

Springer  
**Handbook** *of*  
**Model-Based  
Science**

---

---

*Magnani  
Bertolotti  
Editors*

---

---

---

---

---

---

---

**Springer Handbook  
of Model-Based Science**

---

**Springer Handbooks** provide a concise compilation of approved key information on methods of research, general principles, and functional relationships in physical and applied sciences. The world's leading experts in the fields of physics and engineering will be assigned by one or several renowned editors to write the chapters comprising each volume. The content is selected by these experts from Springer sources (books, journals, online content) and other systematic and approved recent publications of scientific and technical information.

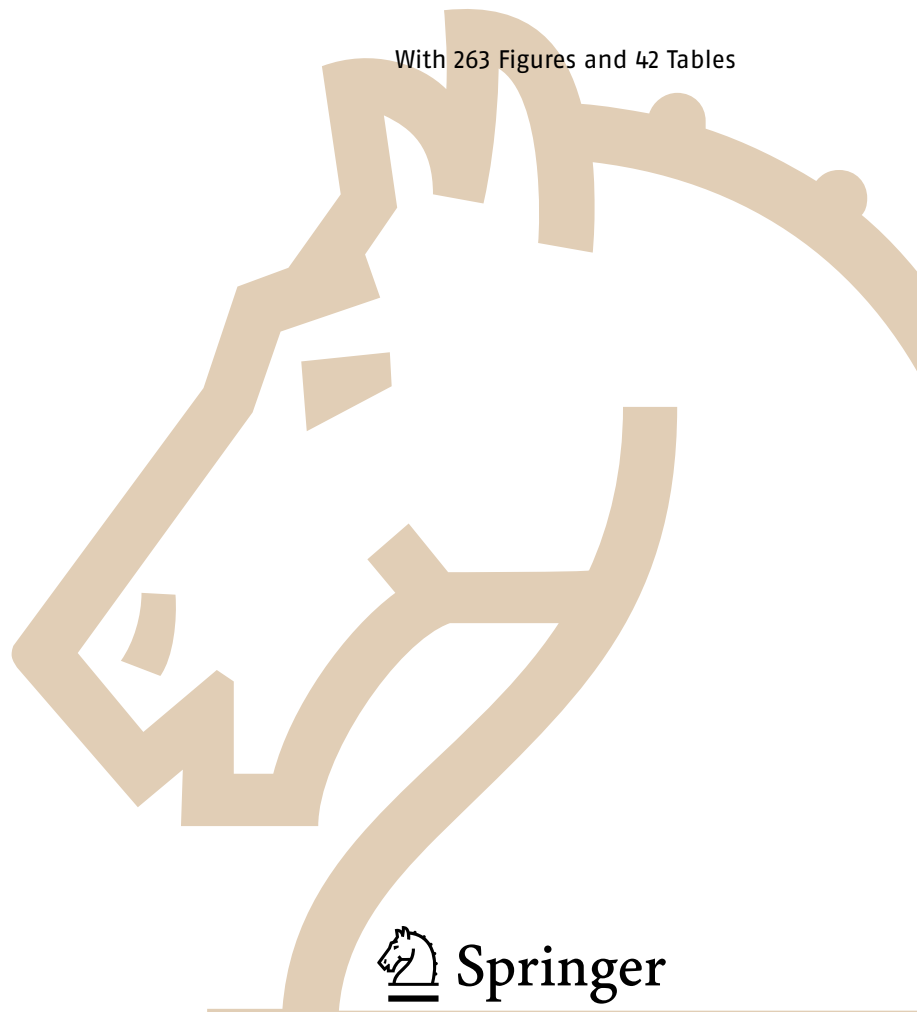
The volumes are designed to be useful as readable desk reference book to give a fast and comprehensive overview and easy retrieval of essential reliable key information, including tables, graphs, and bibliographies. References to extensive sources are provided.

---

# Springer Handbook of Model-Based Science

Lorenzo Magnani, Tommaso Bertolotti (Eds.)

With 263 Figures and 42 Tables



 Springer

---

*Editors*

Lorenzo Magnani  
University of Pavia  
Department of Humanities  
Piazza Botta 6  
Pavia 27100, Italy

Tommaso Bertolotti  
University of Pavia  
Department of Humanities  
Piazza Botta 6  
Pavia 27100, Italy

ISBN: 978-3-319-30525-7 e-ISBN: 978-3-319-30526-4

DOI 10.1007/978-3-319-30526-4

Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2017935722

© Springer International Publishing AG 2017

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Production and typesetting: le-tex publishing services GmbH, Leipzig

Typography and layout: schreiberVIS, Seeheim

Illustrations: Hippmann GbR, Schwarzenbruck

Cover design: eStudio Calamar Steinen, Barcelona

Cover production: WMXDesign GmbH, Heidelberg

Printing and binding: Printer Trento s.r.l., Trento

Printed on acid free paper

This Springer imprint is published by Springer Nature

The registered company is Springer International Publishing AG

The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

## Foreword: Thinking Inside and Outside

Hardly a minute of our lives goes by without reasoning, without “going beyond the information given” (Bruner, 1966). Reasoning happens without our awareness and without our intention. Our visual system rapidly parses light arriving to the retina into figures and ground, into the faces, bodies, objects, scenes, and actions that we recognize in fractions of a second and that enter awareness. The gist stays with us, but the details are not recorded, a process revealed in *change blindness*. Perception is deeply and inherently tied to action, epitomized in mirror neurons, single cells in a monkey cortex that respond both to perceiving an action and to performing it, linking perception and action in a single neuron. The world we see is constantly changing; in order to act, perception is used to predict what happens next, allowing us to catch balls and avoid collisions. Experienced chess and basketball players parse and predict at a more sophisticated level, but the underlying processes seem to be similar, extensive practice in seeing patterns, interpreting them, and selecting appropriate actions. This kind of reasoning is fast thinking (Kahneman, 2011) and the kind of reasoning thoughtfully and thoroughly analyzed in this impressive volume is slow thinking, thinking that is deliberate and reflective and that can unfold over weeks and years and even centuries.

How does deliberative reasoning happen? It happens in a multitude of ways, as is shown insightfully in the many domains analyzed in the chapters that follow. Might there be a way to encompass all of them? Here is one view of thinking, let us call it the *inside-outside* view, a view inspired by theories of Bruner (1966), Piaget (1954), Norman and Rumelhart (1975), Shepard (1984), and Vygotsky (1962), among others, who, however, cannot be held responsible for the present formulation. The inside part is the thinking that happens between the ears; the outside part is the thinking that gets put into the body and the world. This view is iterative, as what gets put in the world can get internalized and then get worked on inside. Inside thinking can be further separated into representations and operations. Representations and operations are useful fictions, ways to think and talk about thinking. Do not expect philosophical precision from them (there is a temperamental difference between and psychologists and philosophers: psychologists live on generalities that ignore sloppy variability, philosophers live

on elegant distinctions) or look for them in the brain. On this view, representations are internalized perceptions. However, representations cannot be copies, they are highly processed. They are interpretations of the content that is the focus of thought. They may select some information from the world and ignore other information, they may rework the information selected, and they may add information, drawing on information already stored in the brain. In this sense, representations are models. On this view, operations are internalized actions, which are analogous to actions in the world. Operations act on representations, transforming them and thereby creating new representations. Examples are in order. We may form a representation of the arrangement of furniture in a room or a corporate structure. We can then draw inferences from the representations by imagining or carrying out actions on the representations, such as comparing parts or tracing paths, for example that the coffee table is too far from the couch or the route from one division to another is too circuitous. You have probably noted that those inferences also depend on functional information stored between the ears. We can then transform the arrangements to create new configurations and act on those to draw further inferences. Seen this way, representations can be created by operations; these processes are iterative and reductive. However, momentarily, representations are regarded as static and transformations as active, changing representations to generate inferences that go beyond the information given to yield new thoughts and discoveries.

The ways that we talk about thinking suggest generality to this view. When we understand, we say that we see something; we have an image or an idea, or a thought or a concept. These are static and they stay still to be acted on. Then we pull ideas together, compare them, turn them inside out, divide them up, reorganize them, or toss them out.

Forming representations, keeping them in mind, and transforming them can quickly overwhelm the mind. When thought overwhelms the mind, the mind puts thought in the body and the world. Counting is a paradigmatic example: actions of the finger or the



**Barbara Tversky**  
Stanford University and  
Columbia Teachers College

hand (or the head or the eye) on an array of objects in the world, pointing to them or moving them, while keeping track of the count with number words. Counting is a sequence of actions on objects linked one-to-one to a sequence of concepts. If representations are internalized perceptions and transformations of thoughts are internalized actions, re-externalizing representations and transformations should promote thought. Moreover, they do, as counting and countless other examples demonstrate. The actions of the body on the objects in the world exemplify the outside kind of thinking and, importantly, they are linked to internal actions, in this case, keeping track of the count.

Putting thought into the world expands the mind. Putting thought into the world allows physical actions to substitute for mental ones. Putting thought in the world makes thought evident to others and to ourselves at other times. Putting thought into the world separates the representing and the transforming and makes them apparent. To be effective, both inside and outside, the representations and the transformations, should be congruent with thought (e.g., Tversky, 2011, 2015). Maps preserve the spatial relations of landmarks and paths. Discrete one-to-one actions help children add and continuous actions help children estimate (Segal, Tversky, and Black, 2014). Gesturing the layout of an environment helps adults remember the spatial relations (Jamalian, Giardino, and Tversky, 2013).

This analysis, representations and operations, inside and outside, is simple, even simplistic. The complexity comes from the interactions between and within inside and outside, in the constructions of the representations, and in the subtleties of the inferences. Representations and operations are intimately interlinked. Representations, internal or external, carry content, but they also have a format, i.e., the way that information is captured and arrayed. The format encourages certain inferences and discourages others. The Arabic number system is friendlier to arithmetic computations than the Roman number system. Maps are friendlier to inferences about routes, distances, and directions than tables of GPS coordinates. Finding the *right* representation, i.e., the one that both captures the information at hand and enables productive inferences, can be hard work; the history of science is filled with such struggles from the structure and workings of the universe to those of the smallest particles. In *The Double Helix* (1968), *Watson* describes the intricate back-and-forth between empirical findings, theory, hypotheses, conversation, photographs, and models that led to the discovery of the model of DNA that succeeded in integrating the biological and chemical structures and phenomena. Typically, there is no single right representation exactly because

different representations capture different information, highlight different relationships, and encourage different inferences. A pilot's map will not serve a hiker or a bicyclist, or a surveyor. Chemists have a multitude of representations of molecules, human biologists of the body, statisticians of a set of data, each designed to simplify, highlight, explain, understand, explore, or discover different aspects of multifaceted, often elusive, phenomena.

This very simple view hides enormous complexity. It can be accused of being an oversimplification. If it were all that simple, design, science, and mathematics would be done with, and they are not. Fortunately, the volume at hand corrects that oversimplification. Introducing this volume is a humbling task. So many kinds of reasoning are revealed so perceptively in so many domains. The diverse thoughtful and thought-provoking contributions reveal fascinating intricacies in model-based reasoning, the nuances of finding suitable representations (models), and the complexities of using them to draw inferences. The many insights in each contribution and in the section overviews cannot readily be summarized, they must be savored. They will be a continuing source of inspiration.

Barbara Tversky

## References

- J. S. Bruner: On cognitive growth. In: *Studies in Cognitive Growth*, ed. by J. S. Bruner, R. R. Olver, P. M. Greenfield (Wiley, Oxford 1966) pp. 1–29
- A. Jamalian, V. Giardino, B. Tversky: Gestures for thinking, Proc. 35th Annu. Conf. Cogn. Sci. Soc., ed. by M. Knauff, M. Pauen, N. Sabaenz, I. Wachsmuth (Cognitive Science Society, Austin 2013)
- J. D. Watson: *The Double Helix: A Personal Account of the Discovery of the Structure of DNA* (Athenum, New York 1968)
- D. Kahneman: *Thinking, Fast and Slow* (Macmillan, New York 2011)
- D. A. Norman, D. E. Rumelhart: *Explorations in cognition* (WH Freeman, San Francisco 1975)
- J. Piaget: *The Construction of Reality in the Child* (Basic Books, New York 1954)
- A. Segal, B. Tversky, J. B. Black: Conceptually congruent actions can promote thought, *J. Res. Mem. Appl. Cogn.* **3**, 124–130 (2014)
- R. N. Shepard: Ecological constraints on internal representation: Resonant kinematics of perceiving, imagining, thinking, and dreaming, *Psychol. Rev.* **91**, 417–447 (1984)

- B. Tversky: Visualizations of thought, *Top. Cogn. Sci.* **3**, 499–535 (2011)
- B. Tversky: The cognitive design of tools of thought, *Rev. Philos. Psychol.* **6**, 99–116 (2015),
- L. Vygotsky: *Thought and Language* (MIT Press, Cambridge 1962)



## Foreword

Thomas Kuhn considered that a sign of maturity in the development of a scientific field was the acquisition of a paradigm; researchers will no longer need to constantly battle over the foundations of their discipline, could agree on the proper questions to ask and the theories, techniques, and procedures to answer them. However, paradigms, Kuhn thought, can also lead to quite esoteric research and, after periods of extended anomalies and crises, are eventually displaced by sufficiently unprecedented, sufficiently open-ended achievements, which, leaving a number of problems still to solve, open new avenues for research.

Whether or not there are Kuhnian paradigms in model-based science is perhaps less important than the undeniable fact that, as a domain of inquiry, we have here all the significant signs of a mature field. Crucial questions are raised and addressed with innovative approaches, earlier and more traditional issues are revisited with new insights, and a cluster of apparently unrelated problems are addressed from a perspective that has noteworthy unifying traits.

The crucial concept that brings all of this together is one that is perhaps as rich and suggestive as that of a paradigm: the concept of a *model*. Some models are concrete, others are abstract. Certain models are fairly rigid; others are left somewhat unspecified. Some models are fully integrated into larger theories; others, or so the story goes, have a life of their own. Models of experiment, models of data, models in simulations, archeological modeling, diagrammatic reasoning, abductive inferences; it is difficult to imagine an area of scientific investigation, or established strategies of research, in which models are not present in some form or another. However, models are ultimately understood, there is no doubt that they play key roles in multiple areas of the sciences, engineering, and mathematics, just as models are central to our understanding of the practices of these fields, their history and the plethora of philosophical, conceptual, logical, and cognitive issues they raise.

What Lorenzo Magnani and Tommaso Bertolotti, together with their team of subject-area editors, have created is more than a snapshot of a growing, sophisticated, multifaceted field. They have provided a rich map (a model!) of a sprawling, genuinely interdisciplinary, eclectic, and endlessly fascinating nested series of domains of research.

With this Springer *Handbook*, model-based science acquires more than a clear sign of maturity: it has the excitement, energy, and innovation of those sufficiently groundbreaking, sufficiently boundless, achievements that will nurture and inspire generation upon generation to come.

Otávio Bueno



**Otávio Bueno**  
University of Miami  
Editor-in-Chief, *Synthese*

## Preface

The debate about models has crossed philosophy along the centuries, ranging from the most speculative to the most pragmatic and cognitive outlooks. Recently, epistemological perspectives and both scientific and cognitive insights sparked new interdisciplinary studies about how models are created and used. The relevance of the discourse about models transcended the boundaries of philosophy of science, as it was immensely boosted by the progress being made in computation since the 1950s, making the discourse on models not only relevant to scientists and philosophers, but also to computer scientists, programmers, and logicians. Another field of study, strictly connected to modeling, was the study of inferential processes that would go beyond traditional logic and yet play a crucial role in the creation and use of models. The most relevant field, in this respect, concerns abduction and studies on hypothetical cognition.

To provide an initial definition, we can agree that a model is something we use in order to gain some benefit in the understanding or explanation of something else, which can be called the target. “A model allows us to infer something about the thing modeled,” as summarized by the late *John Holland* in his 1995 book *Hidden Order*. A model lets us understand the target, and consequently behave in a way that would not be possible without it; different models usually optimize the understanding of different aspects of the target.

This definition of a model should make it easy to appreciate how many situations that we face every day are tackled by making use of models; to deal with other people we make models of their minds and their intentions, to operate machinery we make models of their functioning, in the remote case of trying to escape from wild animals we make models of their hunting strategies and perceptual systems, to explore novel environments we make models of their spatial configurations, to mention only a few. We make use of models in a wide array of circumstances, but what all models actually share is a dimension of non-abstractness; we create them, or make use of models that have already been constructed by other people, and models usually display a distributed nature, since they are either built on external, material supports (i. e., by means of artifacts, paper sheets, sound waves, body gestures) or, in the case of mental models, are encoded in brain wirings by synapses and chemicals (a mental map, for instance, is the mental simulation of the action of draw-

ing a map – a powerful model construction activity – whose embodiment in the brain was made possible by the enhancement of human cognitive capabilities).

In order to grasp to the fullest the rich universe of models, their relevance in model-based science, but also as cognitive architectures, we divided the handbook into nine parts. The first three parts can be seen as the *ABC* of the discourse, providing a cognitive and theoretical alphabet for the understanding of model-based science, while the remaining six parts each deal with precise, and applied, fields of model-based science.

Part A – *Theoretical Issues in Models*, edited by Demetris Portides, sets the foundation for all of the subsequent debates, exploring the relationships between models and core notions such as those of theory, representation, explanation, and simulation; furthermore, the part extensively lays out the contemporary and complex debate about the ontology of models, that is, the different stances concerning their existence and reality, answering questions such as *How real are models?*, *Are they fictitious?*, *Do they exist like an object exists or like an idea exists?*.

In Part B – *Theoretical and Cognitive Issues in Abduction and Scientific Inference*, Editor Woosuk Park selected contributions exploring the fundamental aspects of a key inference in the production of models, both at the cognitive and scientific levels: abduction. This can be defined as the most basic building block of hypothetical reasoning, or the process of inferring certain facts and/or laws and hypotheses that render some sentences plausible, that explain (and also sometimes discover) a certain (eventually new) phenomenon or observation.

Atocha Aliseda edited Part C – *The Logic of Hypothetical Reasoning, Abduction, and Models*, offering a broad perspective on how different kinds of logic can be employed to model modeling itself, and how this sheds light on model-building processes. As a bridge between the more theoretical and the more specific parts, Part D – *Model-Based Reasoning in Science and History of Science*, edited by Nora Alejandrina Schwartz, frames some issues of exemplar theory and cases concerning the use and the understanding of models in the history and philosophy of physics, biology, and social sciences, but is also about the relevant subject of thought experiments.

Albrecht Heeffer edited Part E – *Models in Mathematics*, which illuminates crucial issues such as the

role of diagrams in mathematical reasoning, the importance of models in actual mathematical practice, and the role played by abductive inferences in the emergence of mathematical knowledge.

In Part F – *Model-Based Reasoning in Cognitive Science*, Editor Athanasios Raftopoulos has selected a number of contributions highlighting the strict relationship between model-based science and model-based reasoning (cognitive science being both a model-based science and the science of modeling), namely the model-based processes underpinning vision and diagrammatic reasoning, but also the relevance of deeper cognitive mechanisms such as embodiment and the neural correlates to model-based reasoning.

Francesco Amigoni and Viola Schiaffonati edited the contributions composing Part G – *Modeling and Computational Issues*, concerning the main intersections between computation, engineering, and model-based science, especially with respect to computational rendering and the simulation of model-based reasoning in artificial cognitive processes, up to robotics.

Part H – *Current and Historical Perspectives on the Use of Models in Physics, Chemistry, and Life Sciences*, edited by Mauro Dorato and Matteo Morganti, offers an exemplary outlook on the fundamental aspects concerning models in hard sciences and life sciences, from a perspective that is not chiefly historical (absorbed by Part D), but rather focuses on practical and theoretical issues as they happen in actual scientific practice.

Cameron Shelley edited the final Part I – *Models in Engineering, Architecture, and Economical and Human Sciences*, providing a series of stimulating and innovative contributions focusing on less represented examples of model-based reasoning and science, for instance in archaeology, economics, architecture, design and innovation, but also social policing and moral reasoning. The focus of this closing part also resides in its ability to show that model-based sciences go beyond the tradition of exact and life sciences, as indeed the reliance on models affects nearly all human endeavors.

The brief excursus on the contents does little justice to the richness and the extensive variety of topics reviewed by the Handbook, but it should be enough to convey one of the main ideas of the Handbook: *Models are everywhere*, and the study thereof is crucial in any human science, discipline, or practice. This is why we conceived this book, hoping to make it highly relevant not only for the philosophy, epistemology, cognitive science, logic, and computational science communities, but also for theoretical biologists, physicists, engineers, and other human scientists dealing with models in their daily work.

We like to see the ample theoretical breadth of this Handbook as having a counterpart in its editorial gen-

esis. Indeed, when we were offered the opportunity to be general editors of the Springer Handbook of Model-Based Science, an intense activity of decision-making followed. To be able to make a decision, we had to think about what editing a handbook was like. Otherwise said, in order to decide we had to know better, and in order to know better we had to make ourselves a *model* of handbook editing. This complex model was partly in our heads, partly in sketches and emails. Part of it was deduced from evidence (other handbooks), part of it came out as hypotheses. Once the model was sufficiently stabilized, giving us a good projection of the major criticalities and some (wishful) scheduling, we accepted the challenge, and the model – continuously updating the progression of the work – would guide our behavior step by step.

We undertook the editing of this Handbook because, so far, there is a vast amount of literature on models, on the inferential and logical processes underdetermining them, and on the philosophy of model-based science, but it is dispersed in more or less recent (and variably authoritative) collections and monographs, journal articles, and conference proceedings. The aim of this Handbook is to offer the possibility to access the core and the state-of-the-art of these studies in a unique, reliable source on the topic, authored by a team of renowned experts.

The present Handbook is the exemplary fruit of research and collaboration. As general editors, we were able to rely on the formidable team of editors we mentioned above, who took the reigns of their parts: Atocha Aliseda, Francesco Amigoni, Mauro Dorato, Albrecht Heeffer, Matteo Morganti, Woosuk Park, Demetris Portides, Athanasios Raftopoulos, Viola Schiaffonati, Nora Alejandrina Schwartz, and Cameron Shelley. They are all remarkable and hard-working academics, and we are most grateful to them for taking the time and shouldering the burden to contact authors, inspire and review contributions, whilst keeping in touch with us. In turn, the editors could count on some of the most renowned and promising scholars in each discipline and field; they are too many to mention here, but our undying recognition and gratitude go to them as well. In addition to our recognition, all of the editors and authors certainly have our congratulations and admiration upon the completion of this work.

Many of the editors and contributors were already part of the ever-growing *MBR (model-based reasoning) community*, an enthusiastic group of philosophers, epistemologists, logicians, cognitive scientists, computer scientists, engineers, and other academics working in the different and multidisciplinary facets of what is known as *model-based reasoning*, especially focusing on hypothetical-abductive reasoning and its role in sci-

entific rationality. The outreach of this handbook goes far beyond the theoretical and personal borders of the MBR community, but it can nevertheless be saluted as a celebration of the 17 years of work and exchange since the first MBR conference was held in Pavia, Italy, in 1998. For us, this Handbook is also a recognition of the work and lives of the many beautiful minds who came to join us, or interacted with the MBR community, who are no longer among us but who will be forever remembered and appreciated.

Last but clearly not least, we are most grateful to Springer's editorial and production team for their constant trust, encouragement, and support. In particular, we wish to thank Leontina Di Cecco, Judith Hinterberg, and Holger Schaepe, as their resilient help and collaboration made a difference in achieving this Handbook.

Finally, beyond its tutorial value for our community, it is our hope that the Handbook will serve as a useful source to attract new researchers to model-based science (and model-based reasoning), and inspire decades of vibrant progress in this fascinating interdisciplinary area. The contents of this Handbook admirably present a very useful bringing together of the vast accomplishments that have taken place in the last 50 years. Certainly, the contents of this Handbook will serve as a valuable tool and guide to those who will produce the even more capable and diverse next generations of research on models.

Pavia, Italy  
Lorenzo Magnani  
Tommaso Bertolotti

## About the Editors

**Lorenzo Magnani**, philosopher, epistemologist, and cognitive scientist, is a Professor of Philosophy of Science at the University of Pavia, Italy, and the Director of its Computational Philosophy Laboratory.

He has been a visiting researcher at Carnegie Mellon University, McGill University, the University of Waterloo and Georgia Institute of Technology and a Visiting Professor at Georgia Institute of Technology, City University of New York, and at Sun Yat-sen University, China.

In the event of the 50th anniversary of the re-building of the Philosophy Department of Sun Yat-sen University in 2010, an award was given to him to acknowledge his contributions to the areas of philosophy, philosophy of science, logic, and cognitive science. He was appointed Member of the International Academy for the Philosophy of the Sciences (AIPS) in 2015.

Amongst his various publications, his book *Abduction, Reason, and Science* has become a well-respected work in the field of human cognition. The book *Morality in a Technological World* develops a philosophical and cognitive theory of the relationships between ethics and technology in a naturalistic perspective. The book *Abductive Cognition. The Epistemological and Eco-Cognitive Dimensions of Hypothetical Reasoning* and the last monograph *Understanding Violence. The Intertwining of Morality, Religion, and Violence: A Philosophical Stance* have been more recently published by Springer, in 2009 and 2011.

Since 1998, initially in collaboration with Nancy J. Nersessian and Paul Thagard, he has created and promoted the MBR Conferences on Model-Based Reasoning. Since 2011 he is the editor of the book series *Studies in Applied Philosophy, Epistemology and Rational Ethics (SAPERE)* by Springer.



**Tommaso Bertolotti** is a Postdoctoral Fellow in Philosophy of Science and Adjunct Professor of Cognitive Philosophy in the Department of Humanities – Philosophy Section, University of Pavia. His research interests include philosophy of science, niche construction theories, cognitive science of religion, social epistemology, and philosophy of technology. He was recently invited to several expert workshops and conferences concerning cyberbullying and internet safety in general, organized by the European Commission together with other institutions and major IT companies. Among his latest publications are *Contemporary finance as a critical cognitive niche* (2015), *An epistemological analysis of gossip and gossip-based knowledge* (2014), *Camouflaging truth: A biological, argumentative and epistemological outlook* (2014), *Generative and Demonstrative Experiments* (2013), *From mindless modeling to scientific models: The case of emerging models* (2012), and the book *Patterns of Rationality. Recurring Inferences in Science, Social Cognition and Religious Thinking* (Springer 2015).



---

## About the Part Editors



### Demetris Portides

University of Cyprus  
Dept. of Classics and Philosophy  
1678 Nicosia, Cyprus  
[portides@ucy.ac.cy](mailto:portides@ucy.ac.cy)

### Part A

Demetris Portides received his PhD from the University of London (LSE) in 2000. He teaches in the Department of Classics and Philosophy at the University of Cyprus. His research interests include topics in philosophy of science, particularly idealization and abstraction, scientific models, and scientific representation. He is currently working on the ways by which idealization and abstraction are employed in the construction of scientific models.

---

### Woosuk Park

Korea Advanced Institute of Science and Technology (KAIST)  
School of Humanities and Social Science  
Daejeon, 34141, Korea  
[woosukpark@kaist.ac.kr](mailto:woosukpark@kaist.ac.kr)



### Part B

Woosuk Park is currently Full Professor of Philosophy at Korea Advanced Institute of Science and Technology (KAIST). He received his PhD degree from the State University of New York at Buffalo with his dissertation on Duns Scotus' haecceity theory of individuation. He was interested in the issues on the border between logic and ontology. He has expanded the scope of research primarily by his historical approach. His research area now includes the history of logic and axiomatic methods and the history of medieval science and philosophy. Recently, he has become the most interested in abductive cognition. In a series of papers, he has dealt with Lorenzo Magnani's innovative ideas such as animal abduction or visual abduction, thereby examining to what extent Magnani goes with and beyond Peirce. He has published articles in international journals, including *Journal of Applied Logic*, *Erkenntnis*, *Foundations of Science*, *Review of Metaphysics*, and *The Modern Schoolman*.

---

### Atocha Aliseda

Universidad Nacional Autónoma de México (UNAM)  
Instituto de Investigaciones Filosóficas  
04510 Ciudad de México, Mexico  
[atocha@filosoficas.unam.mx](mailto:atocha@filosoficas.unam.mx)



### Part C

Atocha Aliseda received her PhD from Stanford University in Philosophy and Symbolic Systems (1997) and later held a postdoctoral position at Groningen University (2000–2002). She is Full Professor at the Institute for Philosophical Research at the National Autonomous University of México (UNAM). She has published and edited a number of books and articles on logic and the philosophy of science and specializes on abduction and the logics of scientific discovery. She is currently working on clinical reasoning.



### Nora Alejandrina Schwartz

Universidad de Buenos Aires  
Faculty of Economics  
Buenos Aires, Argentina  
[nora\\_schwartz@yahoo.com.ar](mailto:nora_schwartz@yahoo.com.ar)

### Part D

Nora A. Schwartz received her Professor degree in Philosophy from the Universidad de Buenos Aires, Argentina, where she has been teaching since 1997. Professor Schwartz is a member of Asociación de Filosofía de la República Argentina and Sociedad Argentina de Análisis Filosófico. Her research focuses on physical models implicated in creative scientific reasoning leading to conceptual innovation. She takes the cognitive-historical approach developed by Nancy Nersessian in a critical way, emphasizing the cultural dimension of physical models with which scientists reason. She also analyzes the features that satisfactory scientific models should have in order to draw inferences about the reality modeled by them. Currently, she is investigating Luigi Galvani's "discovery of animal electricity". In particular, she is interested in his selection of different three-dimensional objects as model candidates to reason about human neuromuscular system and in the improvement that this historical research case may make to the understanding of analogy.

**Albrecht Heeffer**

Part E

Ghent University  
Centre for Logic and Philosophy of Science  
9000 Ghent, Belgium  
[albrecht.heeffer@ugent.be](mailto:albrecht.heeffer@ugent.be)

Albrecht Heeffer holds a degree in Engineering and a PhD in Philosophy. He publishes on the history of optics, the history of mathematics, and the philosophy of mathematical practice. He is a Research Fellow at Ghent University, Belgium. Albrecht has been a Visiting Fellow at Kobe University (2008), the Sydney Center for the Foundations of Science (Sydney University, 2011), and The Max Planck Institute for History of Science (Berlin, 2014).

**Athanasios Raftopoulos**

Part F

University of Cyprus  
Department of Psychology  
1678 Nicosia, Cyprus  
[raftop@ucy.ac.cy](mailto:raftop@ucy.ac.cy)



Athanasios Raftopoulos received his PhD from the Johns Hopkins University in 1993. He teaches in the Department of Psychology at the University of Cyprus. His research interests include philosophy of science, philosophy of perception, epistemology, philosophy of the mind, and cognitive science. He is currently working on the relation between cognition and perception.

**Francesco Amigoni**

Part G

Politecnico di Milano  
Dipartimento di Elettronica,  
Informazione e Bioingegneria  
20133 Milano, Italy  
[francesco.amigoni@polimi.it](mailto:francesco.amigoni@polimi.it)



Francesco Amigoni received his PhD in Computer Engineering and Automatica from the Politecnico di Milano (Italy) in 2000. He has been an Associate Professor at the Dipartimento di Elettronica, Informazione e Bioingegneria of the Politecnico di Milano since 2007. His main research interests include agents and multiagent systems, autonomous mobile robotics, and the philosophical aspects of artificial intelligence.

**Viola Schiaffonati**

Part G



Politecnico di Milano  
Dipartimento di Elettronica, Informazione  
e Bioingegneria  
20133 Milano, Italy  
[viola.schiaffonati@polimi.it](mailto:viola.schiaffonati@polimi.it)

Viola Schiaffonati has Laurea (Milano, 1999) and PhD (Genova, 2004) degrees. She is Associate Professor of Logic and Philosophy of Science at the Dipartimento di Elettronica, Informazione e Bioingegneria of Politecnico di Milano. Her main research interests include the philosophical foundations of artificial intelligence and robotics and the philosophy of computing sciences and information, with particular attention on the philosophical issues of computational science and the epistemology of experiments.

**Mauro Dorato**

Part H

University of Rome  
Dipartimento di Filosofia, Comunicazione  
e Spettacolo  
00144 Rome, Italy  
[mauro.dorato@uniroma3.it](mailto:mauro.dorato@uniroma3.it)

Mauro Dorato earned his PhD in Philosophy from the Johns Hopkins University (1992). He is Full Professor for Philosophy of Science at the University of Rome 'Tre'. He has been Director of the PhD Program in Philosophy since 2013. His research focuses on the philosophy of physics, philosophy of time and the nature of scientific laws. Currently he is Co-Editor of the *European Journal for Philosophy of Science*.

**Matteo Morganti**

Part H

University of Rome  
Dipartimento di Filosofia, Comunicazione  
e Spettacolo  
00144 Rome, Italy  
[matteo.morganti@uniroma3.it](mailto:matteo.morganti@uniroma3.it)



Matteo Morganti is Associate Professor at the University of Rome 'Tre'. He works in the field of philosophy of science and is particularly interested in the interplay between science and analytic metaphysics and in the issue of scientific realism versus antirealism and the interpretation of contemporary physical theories. He earned his PhD at the London School of Economics and has worked at IHPST in Paris and the University of Konstanz.

**Cameron Shelley**

Part I

University of Waterloo  
Centre for Society, Technology & Values  
Waterloo, N2L 3G1, Canada  
[cam\\_shelley@yahoo.ca](mailto:cam_shelley@yahoo.ca)



Cameron Shelley received his PhD in Philosophy from the University of Waterloo in 1999. He is currently a Lecturer at the Centre for Society, Technology & Values in the Department of Systems Design Engineering at the University of Waterloo. His research focuses on model-based reasoning and philosophical issues in technology, such as the involvement of ideology and fairness in technological design.

## List of Authors

### Mark Addis

St Mary's University  
School of Arts and Humanities  
Waldegrave Road  
Twickenham, TW1 4SX, UK  
*mark.addis@stmarys.ac.uk*

### Atocha Aliseda

Universidad Nacional Autónoma de México  
(UNAM)  
Instituto de Investigaciones Filosóficas  
Circuito Mario de la Cueva S/N, Ciudad  
Universitaria, Coyoacán  
04510 Ciudad de México, Mexico  
*atocha@filosoficas.unam.mx*

### Francesco Amigoni

Politecnico di Milano  
Dipartimento di Elettronica, Informazione e  
Bioingegneria  
Piazza Leonardo da Vinci 32  
20133 Milano, Italy  
*francesco.amigoni@polimi.it*

### Margherita Arcangeli

Humboldt-Universität of Berlin  
Department of Philosophy  
Unter den Linden 6  
10099 Berlin, Germany  
*margheritarcangeli@gmail.com*

### Cristina Barés Gómez

Universidad de Sevilla  
Grupo de Investigación en Lógica, Lenguaje e  
Información  
C/ Camilo José Cela S/N  
41018 Sevilla, Spain  
*crisbares@gmail.com*

### Alessandra Basso

University of Helsinki  
Department of Political and Economic Studies  
Unioninkatu 40A  
Helsinki, 00014, Finland  
*alessandra.basso@helsinki.fi*

### William Bechtel

University of California San Diego  
Department of Philosophy and Center for  
Circadian Biology  
9500 Gilman Drive  
La Jolla, CA 92093-0119, USA  
*wbechtel@ucsd.edu*

### Mathieu Beirlaen

Ruhr University Bochum  
Institute for Philosophy II  
Universitätsstraße 150  
44801 Bochum, Germany  
*mathieubeirlaen@gmail.com*

### Alisa Bokulich

Boston University  
Center for Philosophy and History of Science  
745 Commonwealth Ave.  
Boston, MA 02215, USA  
*abokulich@bu.edu*

### Tibor Bosse

VU University Amsterdam  
Department of Computer Science  
De Boelelaan 1081  
1081 HV, Amsterdam, The Netherlands  
*t.bosse@vu.nl*

### Juliana Bueno-Soler

University of Campinas  
School of Technology  
Rua Paschoal Marmo 1888  
SP 13484-332, Limeira, Brazil  
*juliana@ft.unicamp.br*

### Angelo Cangelosi

Plymouth University  
Centre for Robotics and Neural Systems  
Drake Circus  
Plymouth, PL4 8AA, UK  
*acangelosi@plymouth.ac.uk*

### Walter Carnielli

University of Campinas  
Centre for Logic, Epistemology and the History of  
Science  
Rua Sérgio Buarque de Holanda, 251 Barão  
Geraldo, DEP  
SP 13083-85, Campinas, Brazil  
*walter.carnielli@cle.unicamp.br*



**Antonio Cicchetti**

Mälardalen University  
Department of Innovation, Design, and  
Engineering  
Mälardalens högskola  
72123 Västerås, Sweden  
*antonio.cicchetti@mdh.se*

**Marcelo E. Coniglio**

University of Campinas  
Centre for Logic, Epistemology and the History of  
Science  
Rua Sérgio Buarque de Holanda, 251 Barão  
Geraldo, DEP  
SP 13083-85, Campinas, Brazil  
*coniglio@cle.unicamp.br*

**Ralf F.A. Cox**

University of Groningen  
Department of Psychology  
Grote Kruisstraat 2/1  
9712 TS, Groningen, The Netherlands  
*r.f.a.cox@rug.nl*

**Edoardo Datteri**

Università degli Studi di Milano-Bicocca  
Dipartimento di Scienze Umane per la  
Formazione "R.Massa"  
Piazza dell'Ateneo Nuovo 1  
20126 Milano, Italy  
*edoardo.datteri@unimib.it*

**Paul Davidsson**

Malmö University  
Department of Computer Science  
Östra Varvgata 11A  
20506 Malmö, Sweden  
*paul.davidsson@mah.se*

**Ruud J.R. Den Hartigh**

University of Groningen  
Department of Psychology  
Grote Kruisstraat 2/1  
9712 TS, Groningen, The Netherlands  
*j.r.den.hartigh@rug.nl*

**Alessandro Di Nuovo**

University of Enna "Kore"  
Faculty of Engineering and Architecture  
Cittadella Universitaria  
94100 Enna, Italy  
*alessandro.dinuovo@unikore.it*

**Santo Di Nuovo**

University of Catania  
Department of Education  
4 via Biblioteca  
95124 Catania, Italy  
*s.dinuovo@unict.it*

**Gordana Dodig-Crnkovic**

Chalmers University of Technology  
Department of Applied Information Technology  
Forskningsgången 6  
41296 Göteborg, Sweden  
*dodig@chalmers.se*

**Matthieu Fontaine**

Universidad Nacional Autónoma de México  
(UNAM)  
Instituto de Investigaciones Filosóficas  
Circuito Mario de la Cueva S/N, Ciudad  
Universitaria  
04510 Ciudad de México, Mexico  
*fontaine.matthieu@gmail.com*

**Joachim Frans**

Vrije Universiteit Brussel  
Centre for Logic and Philosophy of Science  
Pleinlaan 2  
1050 Brussels, Belgium  
*joachim.frans@vub.ac.be*

**Roman Frigg**

London School of Economics and Political Science  
Houghton Street  
London, WC2A 2AE, UK  
*r.p.frigg@lse.ac.uk*

**Tjerk Gauderis**

Ghent University  
Centre for Logic and Philosophy of Science  
Blandijnberg 2  
9000 Gent, Belgium  
*tjerk.gauderis@ugent.be*

**Axel Gelfert**

National University of Singapore  
Dept. of Philosophy  
3 Arts Link  
117570 Singapore, Singapore  
*axel@gelfert.net*

**Charlotte Gerritsen**

Netherlands Institute for the Study of Crime and  
Law Enforcement  
De Boelelaan 1077a  
1081 HV, Amsterdam, The Netherlands  
*cgerritsen@nscr.nl*

**Valeria Giardino**

UMR 7117 CNRS – Université de Lorraine  
 Laboratoire d'Histoire des Sciences et de  
 Philosophie – Archives Henri-Poincaré  
 91 avenue de la Libération  
 54001 Nancy Cedex, France  
*valeria.giardino@univ-lorraine.fr*

**Fernand Gobet**

University of Liverpool  
 Department of Psychological Sciences  
 Bedford Street South  
 Liverpool, L69 7ZW, UK  
*fernand.gobet@liv.ac.uk*

**William Goodwin**

University of South Florida  
 Department of Philosophy  
 4202 E. Fowler Avenue  
 Tampa, FL 33620, USA  
*wgoodwin@usf.edu*

**Bartłomiej Górný**

Comarch S.A.  
 al. Jana Pawła II 39A  
 31-864 Krakow, Poland  
*bartlomiej.gorny@comarch.com*

**Isar Goyvaerts**

Università degli Studi di Torino  
 Dipartimento di Matematica "Giuseppe Peano"  
 Via Carlo Alberto 10  
 10123 Torino, Italy  
*igoyvaer@unito.it*

**Teruaki Hayashi**

University of Tokyo  
 Department of Systems Innovation  
 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan  
*teru-h.884@nifty.com*

**Cyrille Imbert**

Université de Lorraine  
 Archives Poincaré  
 91 avenue de la Libération  
 54000 Nancy, France  
*cyrille.imbert@univ-lorraine.fr*

**Hiroyuki Kido**

University of Tokyo  
 Department of Systems Innovation  
 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan  
*kido.hiroyuki@gmail.com*

**Franziska Klügl**

Örebro University  
 School of Science and Technology  
 70182 Örebro, Sweden  
*franziska.klugl@oru.se*

**Peter C.R. Lane**

University of Hertfordshire  
 School of Computer Science  
 College Lane  
 Hatfield, AL109 AB, UK  
*p.c.lane@herts.ac.uk*

**Antoni Ligeza**

AGH University of Science and Technology  
 Applied Computer Science  
 30-350 Krakow, Poland  
*ligeza@agh.edu.pl*

**Chiara Lisciandra**

University of Groningen  
 Faculty of Economics and Business, Department  
 of Economics, Econometrics and Finance  
 9700AV, Groningen, The Netherlands  
*c.lisciandra@rug.nl*

**Elisabeth A. Lloyd**

Indiana University  
 Department of History and Philosophy of Science  
 and Medicine  
 1011 E. Third Street  
 Bloomington, IN 47405, USA  
*ealloyd@indiana.edu*

**Giuseppe Longo**

Centre Cavailles  
 Ecole Normale Sup.  
 29 Rue d'Ulm  
 75005 Paris, France  
*giuseppe.longo@ens.fr*

**Miles MacLeod**

University of Twente  
 Department of Philosophy  
 Cubicus C313  
 7500 AE, Enschede, The Netherlands  
*m.a.j.macleod@utwente.nl*

**Giovanna Magnani**

University of Pavia  
 Department of Economics and Management  
 via San Felice, 5  
 27100 Pavia, Italy  
*g.magnani@unipv.it*

**Caterina Marchionni**

University of Helsinki  
Department of Political and Economic Studies  
Unioninkatu 40A  
Helsinki, 00014, Finland  
*caterina.marchionni@helsinki.fi*

**Davide Marocco**

Plymouth University, School of Computing  
Electronics and Mathematics  
Drake Circus  
Plymouth, PL4 8AA, UK  
*davide.marocco@plymouth.ac.uk*

**Massimo Marraffa**

University of Rome 'Roma Tre'  
Department of Philosophy, Communication and  
Media Studies  
Via Ostiense 234  
00144 Rome, Italy  
*massimo.marraffa@uniroma3.it*

**Mary Ann Metzger**

University of Maryland UMBC  
Department of Psychology  
1000 Hilltop Circle  
Baltimore, MD 21250, USA  
*metzger@umbc.edu*

**Gerhard Minnameier**

Goethe University Frankfurt am Main  
Faculty of Economics and Business Administration  
Theodor-W.-Adorno-Platz 4  
60629 Frankfurt am Main, Germany  
*minnameier@econ.uni-frankfurt.de*

**Maël Montévil**

Laboratoire MSC  
Université Paris 7 Diderot  
10 rue Alice Domon et Léonie Duquet  
75205 Paris, France  
*mael.montevil@gmail.com*

**Eleonora Montuschi**

Università Ca' Foscari Venezia  
Department of Philosophy and Cultural Heritage  
Palazzo Malcanton Marcorà, Dorsoduro 3484/D  
30123 Venice, Italy  
*eleonora.montuschi@unive.it*

**John Mumma**

California State University San Bernardino  
Department of Philosophy,  
5500 University Parkway  
San Bernadino, CA 92407, USA  
*jmumma@csusb.edu*

**Angel Nepomuceno-Fernández**

Universidad de Sevilla  
Grupo de Investigación en Lógica, Lenguaje e  
Información  
C/ Camilo José Cela S/N  
41018 Sevilla, Spain  
*nepomuce@us.es*

**Nancy J. Nersessian**

Harvard University  
Dept. of Psychology, William James Hall  
33 Kirkland St.  
Cambridge, MA 02138, USA  
*nancynersessian@fas.harvard.edu*

**James Nguyen**

London School of Economics and Political Science  
Houghton Street  
London, WC2A 2AE, UK  
*j.nguyen1@lse.ac.uk*

**Yukio Ohsawa**

University of Tokyo  
Department of Systems Innovation  
7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan  
*ohsawa@sys.t.u-tokyo.ac.jp*

**Naomi Oreskes**

Harvard University  
Department of the History of Science  
1 Oxford Street  
Cambridge, MA 02138, USA  
*oreskes@fas.harvard.edu*

**Woosuk Park**

Korea Advanced Institute of Science and  
Technology (KAIST)  
School of Humanities and Social Science  
Guseong Dong, Yuseong Gu  
Daejeon, 34141, Korea  
*woosukpark@kaist.ac.kr*

**Alfredo Paternoster**

Università di Bergamo  
Department of Letters, Philosophy,  
Communication  
Via Pignolo 123  
24121 Bergamo, Italy  
*alfredo.paternoster@unibg.it*

**Pieter Pauwels**

Ghent University  
Department of Architecture and Urban Planning  
J. Plateaustraat 22  
9000 Ghent, Belgium  
*pipauwel.pauwels@ugent.be*

**Demetris Portides**

University of Cyprus  
Dept. of Classics and Philosophy  
1 Eressou Street  
1678 Nicosia, Cyprus  
*portides@ucy.ac.cy*

**Athanassios Raftopoulos**

University of Cyprus  
Department of Psychology  
65 Kallipoleos Ave  
1678 Nicosia, Cyprus  
*raftop@ucy.ac.cy*

**Ferdinand Rivera**

San José University  
Department of Mathematics & Statistics  
1 Washington Square  
San Jose, CA 95192, USA  
*ferdinand.rivera@sjsu.edu*

**Abilio Rodrigues Filho**

Federal University of Minas Gerais  
Department of Philosophy  
Av. Antonio Carlos 6627, FAFI CH  
31270-901 Pampulha, Belo Horizonte, Brazil  
*abilio@ufmg.br*

**Federica Russo**

University of Amsterdam  
Department of Philosophy  
Oude Turtmarkt 141-147  
1012GC, Amsterdam, The Netherlands  
*f.russo@uva.nl*

**Flavia Santoianni**

University of Naples Federico II  
Department of Humanities Section of Philosophy  
Via Porta di Massa 1  
80133 Naples, Italy  
*flavia.santoianni@unina.it*

**Viola Schiaffonati**

Politecnico di Milano  
Dipartimento di Elettronica, Informazione e  
Bioingegneria  
Piazza Leonardo da Vinci 32  
20133 Milano, Italy  
*viola.schiaffonati@polimi.it*

**Gerhard Schurz**

Heinrich Heine University  
Department of Philosophy, DCLPS  
Universitätsstrasse 1, Geb. 24.52  
40225 Dusseldorf, Germany  
*schurz@phil.uni-duesseldorf.de*

**Nora Alejandrina Schwartz**

Universidad de Buenos Aires  
Faculty of Economics  
Av. Córdoba 2122  
Buenos Aires, Argentina  
*nora\_schwartz@yahoo.com.ar*

**Cameron Shelley**

University of Waterloo  
Centre for Society, Technology & Values  
200 University Avenue West  
Waterloo, N2L 3G1, Canada  
*cam\_shelley@yahoo.ca*

**Fernando Soler-Toscano**

Universidad de Sevilla  
Grupo de Investigación en Lógica, Lenguaje e  
Información  
C/ Camilo José Cela S/N  
41018 Sevilla, Spain  
*fsoler@us.es*

**Peter D. Sozou**

London School of Economics and Political Science  
Centre for Philosophy of Natural and Social  
Science  
Houghton Street  
London, WC2A 2AE, UK  
*p.sozou@lse.ac.uk*

**Susan G. Sterrett**

Wichita State University  
Department of Philosophy  
1845 North Fairmount  
Wichita, KS 67260-0074, USA  
*susangsterrett@gmail.com*

**Ryan D. Tweney**

Bowling Green State University  
Department of Psychology  
Bowling Green, OH 43403, USA  
*tweney@bgsu.edu*

**Paul L.C. Van Geert**

University of Groningen  
Department of Psychology  
Grote Kruisstraat 2/1  
9712 TS, Groningen, The Netherlands  
*p.l.c.van.geert@rug.nl*

**Bart Van Kerkhove**

Vrije Universiteit Brussel  
Centre for Logic and Philosophy of Science  
Pleinlaan 2  
1050 Brussels, Belgium  
*bart.van.kerkhove@vub.ac.be*

**Fernando R. Velázquez-Quesada**

Universidad de Sevilla  
Grupo de Investigación en Lógica, Lenguaje e  
Información  
C/ Camilo José Cela S/N  
41018 Sevilla, Spain  
*frvelazquezquesada@us.es*

**Harko Verhagen**

Stockholm University  
Department of Computer and Systems Sciences  
Borgarfjordsgatan 8  
164 07 Kista, Sweden  
*verhagen@dsv.su.se*

**Jonathan Waskan**

University of Illinois at Urbana-Champaign  
Department of Philosophy  
810 South Wright Street  
Urbana, IL 61801, USA  
*waskan@illinois.edu*

**John Woods**

University of British Columbia  
Dept. of Philosophy, Vancouver Campus  
1866 Main Mall, Buchanan E370  
Vancouver, BC V6T 1Z1, Canada  
*john.woods@ubc.ca*

**Alison Wylie**

University of Washington  
Department of Philosophy and Anthropology,  
Savery Hall  
Seattle, WA 98195-3350, USA  
*aw26@uw.edu*

## Contents

<b>List of Abbreviations</b> .....	XXXVII
------------------------------------	--------

### Part A Theoretical Issues in Models

<b>1 The Ontology of Models</b>	
<i>Axel Gelfert</i> .....	5
1.1 Kinds of Models: Examples from Scientific Practice .....	6
1.2 The Nature and Function of Models .....	8
1.3 Models as Analogies and Metaphors .....	10
1.4 Models Versus the Received View: Sentences and Structures .....	12
1.5 The Folk Ontology of Models .....	16
1.6 Models and Fiction .....	18
1.7 Mixed Ontologies: Models as Mediators and Epistemic Artifacts ....	20
1.8 Summary .....	21
<b>References</b> .....	22
<b>2 Models and Theories</b>	
<i>Demetris Portides</i> .....	25
2.1 The Received View of Scientific Theories .....	26
2.2 The Semantic View of Scientific Theories .....	36
<b>References</b> .....	47
<b>3 Models and Representation</b>	
<i>Roman Frigg, James Nguyen</i> .....	49
3.1 Problems Concerning Model-Representation .....	51
3.2 General Griceanism and Stipulative Fiat .....	55
3.3 The Similarity Conception .....	57
3.4 The Structuralist Conception .....	66
3.5 The Inferential Conception .....	76
3.6 The Fiction View of Models .....	83
3.7 Representation-as .....	91
3.8 Envoi .....	96
<b>References</b> .....	96
<b>4 Models and Explanation</b>	
<i>Alisa Bokulich</i> .....	103
4.1 The Explanatory Function of Models .....	104
4.2 Explanatory Fictions: Can Falsehoods Explain? .....	108
4.3 Explanatory Models and Noncausal Explanations .....	112
4.4 How-Possibly versus How-Actually Model Explanations .....	114
4.5 Tradeoffs in Modeling: Explanation versus Other Functions for Models .....	115
4.6 Conclusion .....	116
<b>References</b> .....	117

<b>5 Models and Simulations</b>	
<i>Nancy J. Nersessian, Miles MacLeod</i> .....	119
5.1 Theory-Based Simulation .....	119
5.2 Simulation not Driven by Theory .....	121
5.3 What is Philosophically Novel About Simulation? .....	124
5.4 Computational Simulation and Human Cognition .....	127
<b>References</b> .....	130
<b>Part B Theoretical and Cognitive Issues on Abduction and Scientific Inference</b>	
<b>6 Reorienting the Logic of Abduction</b>	
<i>John Woods</i> .....	137
6.1 Abduction .....	138
6.2 Knowledge .....	141
6.3 Logic .....	148
<b>References</b> .....	149
<b>7 Patterns of Abductive Inference</b>	
<i>Gerhard Schurz</i> .....	151
7.1 General Characterization of Abductive Reasoning and Ibe .....	152
7.2 Three Dimensions for Classifying Patterns of Abduction .....	154
7.3 Factual Abduction .....	155
7.4 Law Abduction .....	158
7.5 Theoretical-Model Abduction .....	159
7.6 Second-Order Existential Abduction .....	161
7.7 Hypothetical (Common) Cause Abduction Continued .....	162
7.8 Further Applications of Abductive Inference .....	169
<b>References</b> .....	171
<b>8 Forms of Abduction and an Inferential Taxonomy</b>	
<i>Gerhard Minnameier</i> .....	175
8.1 Abduction in the Overall Inferential Context .....	177
8.2 The Logicity of Abduction, Deduction, and Induction .....	183
8.3 Inverse Inferences .....	185
8.4 Discussion of Two Important Distinctions Between Types of Abduction .....	189
8.5 Conclusion .....	193
<b>References</b> .....	193
<b>9 Magnani's Manipulative Abduction</b>	
<i>Woosuk Park</i> .....	197
9.1 Magnani's Distinction Between Theoretical and Manipulative Abduction .....	197
9.2 Manipulative Abduction in Diagrammatic Reasoning .....	198
9.3 When Does Manipulative Abduction Take Place? .....	203
9.4 Manipulative Abduction as a Form of Practical Reasoning .....	204
9.5 The Ubiquity of Manipulative Abduction .....	206
9.6 Concluding Remarks .....	212
<b>References</b> .....	212

## Part C The Logic of Hypothetical Reasoning, Abduction, and Models

<b>10 The Logic of Abduction: An Introduction</b>	
<i>Atocha Aliseda</i> .....	219
10.1 Some History .....	219
10.2 Logical Abduction .....	222
10.3 Three Characterizations .....	225
10.4 Conclusions .....	228
<b>References</b> .....	229
<b>11 Qualitative Inductive Generalization and Confirmation</b>	
<i>Mathieu Beirlaen</i> .....	231
11.1 Adaptive Logics for Inductive Generalization .....	231
11.2 A First Logic for Inductive Generalization .....	232
11.3 More Adaptive Logics for Inductive Generalization .....	237
11.4 Qualitative Inductive Generalization and Confirmation .....	240
11.5 Conclusions .....	245
11.A Appendix: Blocking the Raven Paradox? .....	246
<b>References</b> .....	247
<b>12 Modeling Hypothetical Reasoning by Formal Logics</b>	
<i>Tjerk Gauderis</i> .....	249
12.1 The Feasibility of the Project .....	249
12.2 Advantages and Drawbacks .....	251
12.3 Four Patterns of Hypothetical Reasoning .....	252
12.4 Abductive Reasoning and Adaptive Logics .....	255
12.5 The Problem of Multiple Explanatory Hypotheses .....	256
12.6 The Standard Format of Adaptive Logics .....	256
12.7 $LA_s^r$ : A Logic for Practical Singular Fact Abduction .....	258
12.8 $MLA_s^s$ : A Logic for Theoretical Singular Fact Abduction .....	261
12.9 Conclusions .....	265
12.A Appendix: Formal Presentations of the Logics $LA_s^r$ and $MLA_s^s$ .....	265
<b>References</b> .....	267
<b>13 Abductive Reasoning in Dynamic Epistemic Logic</b>	
<i>Angel Nepomuceno-Fernández, Fernando Soler-Toscano,</i> <i>Fernando R. Velázquez-Quesada</i> .....	269
13.1 Classical Abduction .....	270
13.2 A Dynamic Epistemic Perspective .....	272
13.3 Representing Knowledge and Beliefs .....	275
13.4 Abductive Problem and Solution .....	278
13.5 Selecting the Best Explanation .....	281
13.6 Integrating the Best Solution .....	284
13.7 Working with the Explanations .....	287
13.8 A Brief Exploration to Nonideal Agents .....	289
13.9 Conclusions .....	290
<b>References</b> .....	292
<b>14 Argumentation and Abduction in Dialogical Logic</b>	
<i>Cristina Barés Gómez, Matthieu Fontaine</i> .....	295
14.1 Reasoning as a Human Activity .....	295



14.2	Logic and Argumentation: The Divorce .....	297
14.3	Logic and Argumentation: A Reconciliation .....	299
14.4	Beyond Deductive Inference: Abduction .....	303
14.5	Abduction in Dialogical Logic.....	306
14.6	Hypothesis: What Kind of Speech Act?.....	310
14.7	Conclusions .....	312
	<b>References</b> .....	312
<b>15</b>	<b>Formal (In)consistency, Abduction and Modalities</b>	
	<i>Juliana Bueno-Soler, Walter Carnielli, Marcelo E. Coniglio,</i>	
	<i>Abilio Rodrigues Filho</i> .....	315
15.1	Paraconsistency.....	315
15.2	Logics of Formal Inconsistency .....	316
15.3	Abduction.....	322
15.4	Modality.....	327
15.5	On Alternative Semantics for mbC.....	331
15.6	Conclusions .....	333
	<b>References</b> .....	334
<b>Part D Model-Based Reasoning in Science</b>		
<b>and the History of Science</b>		
<b>16</b>	<b>Metaphor and Model-Based Reasoning</b>	
	<b>in Mathematical Physics</b>	
	<i>Ryan D. Tweney</i> .....	341
16.1	Cognitive Tools for Interpretive Understanding .....	343
16.2	Maxwell's Use of Mathematical Representation .....	345
16.3	Unpacking the Model-Based Reasoning .....	348
16.4	Cognition and Metaphor in Mathematical Physics .....	350
16.5	Conclusions .....	351
	<b>References</b> .....	352
<b>17</b>	<b>Nancy Nersessian's Cognitive-Historical Approach</b>	
	<i>Nora Alejandrina Schwartz</i> .....	355
17.1	Questions About the Creation of Scientific Concepts.....	356
17.2	The Epistemic Virtues of Cognitive Historical Analysis .....	359
17.3	Hypothesis About the Creation of Scientific Concepts .....	363
17.4	Conclusions .....	373
	<b>References</b> .....	373
<b>18</b>	<b>Physically Similar Systems – A History of the Concept</b>	
	<i>Susan G. Sterrett</i> .....	377
18.1	Similar Systems, the Twentieth Century Concept .....	379
18.2	Newton and Galileo.....	380
18.3	Late Nineteenth and Early Twentieth Century .....	383
18.4	1914: The Year of <i>Physically Similar Systems</i> .....	397
18.5	Physically Similar Systems: The Path in Retrospect.....	408
	<b>References</b> .....	409
<b>19</b>	<b>Hypothetical Models in Social Science</b>	
	<i>Alessandra Basso, Chiara Lisciandra, Caterina Marchionni</i> .....	413
19.1	Hypothetical Modeling as a Style of Reasoning.....	413

19.2	Models Versus Experiments: Representation, Isolation and Resemblance .....	416
19.3	Models and Simulations: Complexity, Tractability and Transparency .....	420
19.4	Epistemology of Models .....	423
19.5	Conclusions .....	428
19.A	Appendix: J.H. von Thünen's Model of Agricultural Land Use in the Isolated State .....	429
19.B	Appendix: T. Schelling's Agent-Based Model of Segregation in Metropolitan Areas .....	430
	<b>References</b> .....	431
<b>20</b>	<b>Model-Based Diagnosis</b>	
	<i>Antoni Ligęza, Bartłomiej Górný</i> .....	435
20.1	A Basic Model for Diagnosis .....	437
20.2	A Review and Taxonomy of Knowledge Engineering Methods for Diagnosis .....	438
20.3	Model-Based Diagnostic Reasoning .....	440
20.4	A Motivation Example .....	440
20.5	Theory of Model-Based Diagnosis .....	442
20.6	Causal Graphs .....	444
20.7	Potential Conflict Structures .....	446
20.8	Example Revisited. A Complete Diagnostic Procedure .....	448
20.9	Refinement: Qualitative Diagnoses .....	450
20.10	Dynamic Systems Diagnosis: The Three-Tank Case .....	454
20.11	Incremental Diagnosis .....	456
20.12	Practical Example and Tools .....	458
20.13	Concluding Remarks .....	459
	<b>References</b> .....	460
<b>21</b>	<b>Thought Experiments in Model-Based Reasoning</b>	
	<i>Margherita Arcangeli</i> .....	463
21.1	Overview .....	464
21.2	Historical Background .....	467
21.3	What Is a Thought Experiment? .....	469
21.4	What Is the Function of Thought Experiments? .....	475
21.5	How Do Thought Experiments Achieve Their Function? .....	484
	<b>References</b> .....	487
 <b>Part E Models in Mathematics</b>		
<b>22</b>	<b>Diagrammatic Reasoning in Mathematics</b>	
	<i>Valeria Giardino</i> .....	499
22.1	Diagrams as Cognitive Tools .....	499
22.2	Diagrams and (the Philosophy of) Mathematical Practice .....	501
22.3	The Euclidean Diagram .....	503
22.4	The Productive Ambiguity of Diagrams .....	509
22.5	Diagrams in Contemporary Mathematics .....	510
22.6	Computational Approaches .....	515
22.7	Mathematical Thinking: Beyond Binary Classifications .....	518
22.8	Conclusions .....	520
	<b>References</b> .....	521

<b>23 Deduction, Diagrams and Model-Based Reasoning</b>	
<i>John Mumma</i> .....	523
23.1 Euclid's Systematic Use of Geometric Diagrams .....	524
23.2 Formalizing Euclid's Diagrammatic Proof Method .....	525
23.3 Formal Geometric Diagrams as Models .....	532
<b>References</b> .....	534
<b>24 Model-Based Reasoning in Mathematical Practice</b>	
<i>Joachim Frans, Isar Goyvaerts, Bart Van Kerkhove</i> .....	537
24.1 Preliminaries.....	537
24.2 Model-Based Reasoning: Examples .....	538
24.3 The Power of Heuristics and Plausible Reasoning .....	540
24.4 Mathematical Fruits of Model-Based Reasoning .....	542
24.5 Conclusion .....	546
24.A Appendix.....	546
<b>References</b> .....	548
<b>25 Abduction and the Emergence of Necessary Mathematical Knowledge</b>	
<i>Ferdinand Rivera</i> .....	551
25.1 An Example from the Classroom .....	551
25.2 Inference Types .....	555
25.3 Abduction in Math and Science Education .....	561
25.4 Enacting Abductive Action in Mathematical Contexts .....	564
<b>References</b> .....	566
<b>Part F Model-Based Reasoning in Cognitive Science</b>	
<b>26 Vision, Thinking, and Model-Based Inferences</b>	
<i>Athanasios Raftopoulos</i> .....	573
26.1 Inference and Its Modes .....	576
26.2 Theories of Vision .....	577
26.3 Stages of Visual Processing .....	585
26.4 Cognitive Penetrability of Perception and the Relation Between Early Vision and Thinking .....	588
26.5 Late Vision, Inferences, and Thinking .....	591
26.6 Concluding Discussion .....	596
26.A Appendix: Forms of Inferences .....	597
26.B Appendix: Constructivism .....	598
26.C Appendix: Bayes' Theorem and Some of Its Epistemological Aspects .....	600
26.D Appendix: Modal and Amodal Completion or Perception.....	600
26.E Appendix: Operational Constraints in Visual Processing .....	601
<b>References</b> .....	602
<b>27 Diagrammatic Reasoning</b>	
<i>William Bechtel</i> .....	605
27.1 Cognitive Affordances of Diagrams and Visual Images .....	606
27.2 Reasoning with Data Graphs .....	608
27.3 Reasoning with Mechanism Diagrams .....	613
27.4 Conclusions and Future Tasks .....	616
<b>References</b> .....	617

<b>28 Embodied Mental Imagery in Cognitive Robots</b>	
<i>Alessandro Di Nuovo, Davide Marocco, Santo Di Nuovo, Angelo Cangelosi.</i>	619
28.1 Mental Imagery Research Background .....	620
28.2 Models and Approaches Based on Mental Imagery in Cognitive Systems and Robotics .....	622
28.3 Experiments .....	624
28.4 Conclusion .....	635
<b>References</b> .....	635
<b>29 Dynamical Models of Cognition</b>	
<i>Mary Ann Metzger</i> .....	639
29.1 Dynamics .....	639
29.2 Data-Oriented Models .....	641
29.3 Cognition and Action Distinct .....	644
29.4 Cognition and Action Intrinsically Linked .....	648
29.5 Conclusion .....	653
<b>References</b> .....	655
<b>30 Complex versus Complicated Models of Cognition</b>	
<i>Ruud J.R. Den Hartigh, Ralf F.A. Cox, Paul L.C. Van Geert</i> .....	657
30.1 Current Views on Cognition .....	658
30.2 Explaining Cognition .....	660
30.3 Is Cognition Best Explained by a Complicated or Complex Model? ..	662
30.4 Conclusion .....	666
<b>References</b> .....	666
<b>31 From Neural Circuitry to Mechanistic Model-Based Reasoning</b>	
<i>Jonathan Waskan</i> .....	671
31.1 Mechanistic Reasoning in Science .....	672
31.2 The Psychology of Model-Based Reasoning .....	673
31.3 Mental Models in the Brain: Attempts at Psycho-Neural Reduction .....	675
31.4 Realization Story Applied .....	686
31.5 Mechanistic Explanation Revisited .....	687
31.6 Conclusion .....	690
<b>References</b> .....	690

## Part G Modelling and Computational Issues

<b>32 Computational Aspects of Model-Based Reasoning</b>	
<i>Gordana Dodig-Crnkovic, Antonio Cicchetti</i> .....	695
32.1 Computational Turn Seen from Different Perspectives .....	695
32.2 Models of Computation .....	697
32.3 Computation Versus Information .....	700
32.4 The Difference Between Mathematical and Computational (Executable) Models .....	702
32.5 Computation in the Wild .....	703
32.6 Cognition: Knowledge Generation by Computation of New Information .....	706
32.7 Model-Based Reasoning and Computational Automation of Reasoning .....	709

32.8	Model Transformations and Semantics: Separation Between Semantics and Ontology .....	712
	<b>References</b> .....	715
<b>33</b>	<b>Computational Scientific Discovery</b>	
	<i>Peter D. Sozou, Peter C.R. Lane, Mark Addis, Fernand Gobet</i> .....	719
33.1	The Roots of Human Scientific Discovery .....	720
33.2	The Nature of Scientific Discovery .....	721
33.3	The Psychology of Human Scientific Discovery .....	722
33.4	Computational Discovery in Mathematics .....	723
33.5	Methods and Applications in Computational Scientific Discovery ...	725
33.6	Discussion .....	730
	<b>References</b> .....	731
<b>34</b>	<b>Computer Simulations and Computational Models in Science</b>	
	<i>Cyrille Imbert</i> .....	735
34.1	Computer Simulations in Perspective .....	736
34.2	The Variety of Computer Simulations and Computational Models...	739
34.3	Epistemology of Computational Models and Computer Simulations	743
34.4	Computer Simulations, Explanation, and Understanding .....	750
34.5	Comparing: Computer Simulations, Experiments and Thought Experiments .....	758
34.6	The Definition of Computational Models and Simulations .....	767
34.7	Conclusion: Human-Centered, but no Longer Human-Tailored Science .....	773
	<b>References</b> .....	775
<b>35</b>	<b>Simulation of Complex Systems</b>	
	<i>Paul Davidsson, Franziska Klügl, Harko Verhagen</i> .....	783
35.1	Complex Systems .....	783
35.2	Modeling Complex Systems .....	785
35.3	Agent-Based Simulation of Complex Systems .....	789
35.4	Summing Up and Future Trends .....	795
	<b>References</b> .....	796
<b>36</b>	<b>Models and Experiments in Robotics</b>	
	<i>Francesco Amigoni, Viola Schiaffonati</i> .....	799
36.1	A Conceptual Premise .....	799
36.2	Experimental Issues in Robotics .....	801
36.3	From Experimental Computer Science to Good Experimental Methodologies in Autonomous Robotics .....	802
36.4	Simulation .....	804
36.5	Benchmarking and Standards .....	807
36.6	Competitions and Challenges .....	809
36.7	Conclusions .....	812
	<b>References</b> .....	812
<b>37</b>	<b>Biorobotics</b>	
	<i>Edoardo Datteri</i> .....	817
37.1	Robots as Models of Living Systems .....	817
37.2	A Short History of Biorobotics .....	825

37.3	Methodological Issues .....	826
37.4	Conclusions .....	833
	<b>References</b> .....	834

## Part H Models in Physics, Chemistry and Life Sciences

### 38 Comparing Symmetries in Models and Simulations

	<i>Giuseppe Longo, Maël Montévil</i> .....	843
38.1	Approximation .....	844
38.2	What Do Equations and Computations Do? .....	845
38.3	Randomness in Biology .....	848
38.4	Symmetries and Information in Physics and Biology .....	849
38.5	Theoretical Symmetries and Randomness .....	852
	<b>References</b> .....	854

### 39 Experimentation on Analogue Models

	<i>Susan G. Sterrett</i> .....	857
39.1	Analogue Models: Terminology and Role .....	858
39.2	Analogue Models in Physics .....	868
39.3	Comparing Fundamental Bases for Physical Analogue Models .....	873
39.4	Conclusion .....	876
	<b>References</b> .....	877

### 40 Models of Chemical Structure

	<i>William Goodwin</i> .....	879
40.1	Models, Theory, and Explanations in Structural Organic Chemistry .....	881
40.2	Structures in the Applications of Chemistry .....	883
40.3	The Dynamics of Structure .....	885
40.4	Conclusion .....	889
	<b>References</b> .....	889

### 41 Models in Geosciences

	<i>Alisa Bokulich, Naomi Oreskes</i> .....	891
41.1	What Are Geosciences? .....	891
41.2	Conceptual Models in the Geosciences .....	892
41.3	Physical Models in the Geosciences .....	893
41.4	Numerical Models in the Geosciences .....	895
41.5	Bringing the Social Sciences Into Geoscience Modeling .....	897
41.6	Testing Models: From Calibration to Validation .....	898
41.7	Inverse Problem Modeling .....	902
41.8	Uncertainty in Geoscience Modeling .....	903
41.9	Multimodel Approaches in Geosciences .....	907
41.10	Conclusions .....	908
	<b>References</b> .....	908

### 42 Models in the Biological Sciences

	<i>Elisabeth A. Lloyd</i> .....	913
42.1	Evolutionary Theory .....	913
42.2	Confirmation in Evolutionary Biology .....	922
42.3	Models in Behavioral Evolution and Ecology .....	925
	<b>References</b> .....	927

<b>43 Models and Mechanisms in Cognitive Science</b>	
<i>Massimo Marraffa, Alfredo Paternoster</i> .....	929
43.1 What is a Model in Cognitive Science? .....	929
43.2 Open Problems in Computational Modeling .....	940
43.3 Conclusions .....	948
<b>References</b> .....	949
<b>44 Model-Based Reasoning in the Social Sciences</b>	
<i>Federica Russo</i> .....	953
44.1 Modeling Practices in the Social Sciences .....	954
44.2 Concepts of Model .....	958
44.3 Models and Reality .....	962
44.4 Models and Neighboring Concepts .....	963
44.5 Conclusion .....	967
<b>References</b> .....	968
<b>Part I Models in Engineering, Architecture, and Economical and Human Sciences</b>	
<b>45 Models in Architectural Design</b>	
<i>Pieter Pauwels</i> .....	975
45.1 Architectural Design Thinking .....	976
45.2 BIM Models and Parametric Models .....	981
45.3 Implementing and Using ICT for Design and Construction .....	984
<b>References</b> .....	987
<b>46 Representational and Experimental Modeling in Archaeology</b>	
<i>Alison Wylie</i> .....	989
46.1 Philosophical Resources and Archaeological Parallels .....	990
46.2 The Challenges of Archaeological Modeling .....	991
46.3 A Taxonomy of Archaeological Models .....	992
46.4 Conclusions .....	1000
<b>References</b> .....	1000
<b>47 Models and Ideology in Design</b>	
<i>Cameron Shelley</i> .....	1003
47.1 Design and Ideology .....	1003
47.2 Models and Ideology .....	1004
47.3 Revivalism: Looking to the Past .....	1005
47.4 Modernism: Transcending History .....	1006
47.5 Industrial Design: The Shape of Things to Come .....	1009
47.6 Biomimicry .....	1011
47.7 Conclusion .....	1013
<b>References</b> .....	1013
<b>48 Restructuring Incomplete Models in Innovators Marketplace on Data Jackets</b>	
<i>Yukio Ohsawa, Teruaki Hayashi, Hiroyuki Kido</i> .....	1015
48.1 Chance Discovery as a Trigger to Innovation .....	1016
48.2 Chance Discovery from Data and Communication .....	1016

48.3	IM for Externalizing and Connecting Requirements and Solutions .	1020
48.4	Innovators Marketplace on Data Jackets .....	1022
48.5	IMDJ as Place for Reasoning on Incomplete Models .....	1023
48.6	Conclusions .....	1029
	<b>References</b> .....	1029
<b>49</b>	<b>Models in Pedagogy and Education</b>	
	<i>Flavia Santoianni</i> .....	1033
49.1	Pluralism .....	1034
49.2	Dialecticity .....	1039
49.3	Applied Models .....	1042
49.4	Conclusions .....	1048
	<b>References</b> .....	1048
<b>50</b>	<b>Model-Based Reasoning in Crime Prevention</b>	
	<i>Charlotte Gerritsen, Tibor Bosse</i> .....	1051
50.1	Ambient Intelligence .....	1053
50.2	Methodology .....	1054
50.3	Domain Model .....	1055
50.4	Analysis Model .....	1058
50.5	Support Model .....	1060
50.6	Results .....	1060
50.7	Discussion .....	1062
	<b>References</b> .....	1062
<b>51</b>	<b>Modeling in the Macroeconomics of Financial Markets</b>	
	<i>Giovanna Magnani</i> .....	1065
51.1	The Intrinsic Instability of Financial Markets .....	1066
51.2	The Financial Theory of Investment .....	1071
51.3	The Financial Instability Hypothesis Versus the Efficient Markets Hypothesis .....	1074
51.4	Irving Fisher's Debt-Deflation Model .....	1074
51.5	Policy Implications and the Shareholder Maximization Value Model .....	1079
51.6	Integrating the Minskyian Model with New Marxists and Social Structure of Accumulation (SSA) Theories .....	1085
51.7	Risk and Uncertainty .....	1086
	<b>References</b> .....	1098
<b>52</b>	<b>Application of Models from Social Science to Social Policy</b>	
	<i>Eleonora Montuschi</i> .....	1103
52.1	Unrealistic Assumptions .....	1105
52.2	Real Experiments, Not Models Please! .....	1110
52.3	Conclusions .....	1115
	<b>References</b> .....	1116
<b>53</b>	<b>Models and Moral Deliberation</b>	
	<i>Cameron Shelley</i> .....	1117
53.1	Rules .....	1118
53.2	Mental Models .....	1119
53.3	Schemata .....	1121



53.4 Analogy .....	1122
53.5 Empathy.....	1124
53.6 Role Models.....	1125
53.7 Discussion.....	1126
<b>References</b> .....	<b>1127</b>
<b>About the Authors</b> .....	<b>1129</b>
<b>Detailed Contents</b> .....	<b>1141</b>
<b>Subject Index</b> .....	<b>1163</b>

## List of Abbreviations

1-D	one-dimensional
2-D	two-dimensional
2.5-D	two-and-a-half-dimensional
3-D	three-dimensional
3M	mechanism-model-mapping constraint
4-D	four-dimensional
5-D	five-dimensional

### A

AAMAS	autonomous agents and multiagent systems
AB	abnormal behavior
ABC	amoeba-based computing
ABM	agent-based model
ABS	agent-based simulation
ACT-R	Adaptive Control of Thought-Rational
AEC	architecture, engineering and construction
AGM	Alchourrón, Gärdenfors and Makinson
AI	artificial intelligence
AKM	Aliseda–Kakas/Kowalski–Magnani/Meheus
AL	adaptive logic
ALP	abductive logic programming
AM	Automated Mathematician
AmI	ambient intelligence
ANN	artificial neural network
ART	adaptive resonance theory

### B

bd	basic property of determinedness
BDI	belief, desire, intention
BIM	building information model
BMN	braided monoidal categories
BN	Bayesian network
BOID	beliefs, obligations, intentions and desires
BPTT	backpropagation through time

### C

CA	cellular automata
CAD	computer-aided design
CAESAR	cellular automaton evolutionary slope and river
CAPM	capital asset pricing model
CBR	case-based reasoning
CC	causal connection
CCG	clock-controlled genes
CDS	complex dynamic systems
CERN	European Organization for Nuclear Research
CG	causal graph

CHILD	channel-hillslope integrated landscape development
CI	cognitively impenetrable
CIF	cash inflow
CL	classical logic
CMIP	coupled model intercomparison project
CMR	computational matrix representation
COF	cash outflow
COR	chains of reasoning connections
CORMAS	common-pool resources and multi-agent systems
CP	cognitive penetrability
CPL	classical propositional logic
CPM	classical particle mechanic
CPU	central processing unit
CR	causal response
CRTM	computational and representational theory of mind

### D

DAT	derivability adjustment theorem
DC	demarcation criterion
DCF	disjunctive conceptual fault
DDI	denotation, demonstration, interpretation
DEKI	denotation, exemplification, keying-up and imputation
DEL	dynamic epistemic logic
DEVS	Discrete Event System Specification
DIAMOND	diagrammatic reasoning and deduction
DJ	data jacket
DL	description logics
DN	deductive-nomological
DNA	deoxyribonucleic acid
DoF	degree of freedom
DRNN	dual recurrent neural network
DSL	domain specific language
DSM	discrete state machine

### E

E.O.L.	ease of learning
EELD	evidence extraction and link discovery
EL	epistemic logic
ENIAC	electronic numerical integrator and computer
EPR	Einstein–Podolsky–Rosen
ER	epistemic representation
ERH	external reality hypothesis
ERP	event-related potentials
ESM	Earth system model
ESS	evolutionary stable strategy
EU	expected utility
EUT	expected utility theory

<b>F</b>		KJ	Kawakita Jiro
F.O.K.	feeling of knowing	KL	Kullback–Liebler
FBM	feature-based modeling	<b>L</b>	
FD	fractal dimension	LDI	logics of deontic (in)consistency
FEF	front eye fields	LEM	landscape evolution model
FEL	fast enabling link	LFI	logics of formal inconsistency
FEM	finite element model	LLL	lower limit logic
FFNN	feedforward neural network	LOC	lateral occipital complex
FFS	feedforward sweep	LRP	local recurrent processing
fMRI	functional magnetic resonance imaging	LSF	low spatial frequency
FOL	first-order logic	LT	Logic Theorist
FT	fundamental theorem	LTM	long term memory
<b>G</b>		LTP	long term potentiation
GCM	general circulation model	LTWM	long term working memory
GDI	general definition of information	<b>M</b>	
GEM	good experimental methodology	MABS	multi-agent-based simulation
GG	General Griceanism	MASON	multi-agent simulator of neighborhoods
GIS	geographical information system	MAYA	most advanced yet acceptable
GLUE	generalized likelihood uncertainty estimation	MBR	model-based reasoning
GOLEM	geomorphic-oro-genic landscape evolution model	MG	mechanism governing
GPL	general purpose language	MI	mechanism implemented
GRM	Gaussian random matrix	MMH	massive modularity hypothesis
GRP	global recurrent processing	MoralDM	moral decision-making
GTF	geomorphic transport function	MSE	mean squared error
GW	Gabbay–Woods	<b>N</b>	
<b>H</b>		NCF	net cash inflow
HD	hypothetico-deductive model of confirmation	NFC	nonfinancial corporation
HM	Hopf monoids	<b>O</b>	
HS	hypothetico-structural	OBS	observations
HSF	high spatial frequency	OMG	object management group
<b>I</b>		OOP	object oriented programming
IBAE	inference to the best available explanation	<b>P</b>	
IBE	inference to the best explanation	P.T.R.	prediction of total recall
ICB	idle cash balance	PCN	pre-nex conjunctive normal
ICT	information and communication technologies	PCS	potential conflict structure
IPE	intuitive physics engine	pdf	probability density function
IQ	intelligence quotient	PDP	parallel distributed processing
ISB	integrative systems biology	PI	processes of induction
IT	inferotemporal cortex in the brain	POPI	principle of property independence
<b>J</b>		PSI	principles of synthetic intelligence
J.O.L.	judgment of learning	PT	prospect theory
<b>K</b>		PTS	possible-translations semantics
KB	knowledge base	<b>Q</b>	
KE	knowledge engineering	QALY	quality adjusted life year
		QM	quantum mechanics

**R**


---

RBC	reference benchmark controller
RC	conditional rule
RCM	regional climate model
RCT	randomized controlled trial
RE	real experiment
RL	reinforcement learning
RNN	recurrent neural network
RP	recurrent processes
RU	unconditional rule
RV	received view
RVT	rational value theory

**S**


---

SCN	suprachiasmatic nucleus
SD	system description
SES	socioeconomic status
SeSAm	Shell for Simulated Agent Systems
SEU	subjective expected utility
SF	standard format
SMG	specialist modelling group
SSA	social structure of accumulation
SSK	sociology of scientific knowledge

STEM	science, technology, engineering, and mathematics
STM	short term memory
SV	semantic view

**T**


---

TCE	transaction cost economics
TE	thought experiment
ToMM	theory of mind mechanism
TRoPICAL	two route, prefrontal instruction, competition of affordances, language simulation

**U**


---

UML	unified modeling language
-----	---------------------------

**V**


---

VaR	value at risk
-----	---------------

**W**


---

WM	working memory
----	----------------

---

# Theoretical

# Part A

## Part A Theoretical Issues in Models

Ed. by Demetris Portides

**1 The Ontology of Models**

Axel Gelfert, Singapore, Singapore

**2 Models and Theories**

Demetris Portides, Nicosia, Cyprus

**3 Models and Representation**

Roman Frigg, London, UK  
James Nguyen, London, UK

**4 Models and Explanation**

Alisa Bokulich, Boston, USA

**5 Models and Simulations**

Nancy J. Nersessian, Cambridge, USA  
Miles MacLeod, Enschede,  
The Netherlands

It is not hard to notice the lack of attention paid to scientific models in mid-twentieth century philosophy of science. Models were, for instance, absent from philosophical theories of scientific explanation; they were also absent from attempts to understand how theoretical concepts relate to experimental results. In the last few decades, however, this has changed, and philosophers of science are increasingly turning their attention to scientific models. Models and their relation to other parts of the scientific apparatus are now under philosophical scrutiny; at the same time, they are instrumental parts of approaches that aim to address certain philosophical questions.

After recognizing the significance of models in scientific inquiry and in particular the significance of models in linking theoretical concepts to experimental reports, philosophers have begun to explore a number of questions about the nature and function of models. There are several philosophically interesting questions that could fit very well into the theme of this set of chapters. For example, *what is the function of models?* and *what is the role of idealization and abstraction in modeling?* It is, however, not the objective of this set of chapters to address every detail about models that has gained philosophical interest over time. In this part of the book five model-related philosophical questions are isolated from others and are explored in separate chapters:

1. What is a scientific model?
2. How do models and theories relate?
3. How do models represent phenomena?
4. How do models function in scientific explanation?
5. How do models and other modes of scientific theorizing, such as simulations, relate?

Of course, the authors of these chapters are all aware that isolating these questions is only done in order to reach an intelligible exposition of the explored problems concerning models, and not because different questions have been kept systematically apart in the philosophical literature that preceded this work. In fact, the very nature of some of these questions dictates an interrelation with others and attempts to address one leads to overlaps with attempts to address others. For example, how one addresses the question *what sort of entities are models?* or how one conceives the theory–model relation affects the understanding of their scientific representation and scientific explanation, and vice versa. Although this point becomes evident in the subsequent chapters, a conscious attempt was made by each author to focus on the one question of concern of their chapter and to attempt to extrapolate and explicate the different proposed philosophical accounts that

have been offered in the quest to answer that particular question. We hope that the final outcome is helpful and illuminating to the reader.

*Axel Gelfert* in his contribution, **Chap. 1: *The Ontology of Models***, explicates the different ways in which philosophers have addressed the issue of what a scientific model is. For historical reasons, he begins by examining the view that was foremost almost a century ago, which held that models could be understood as analogies. He then quickly turns his attention to a debate that took place in the second half of the twentieth century between advocates of logical positivism, who held that models are interpretations of a formal calculus, and advocates of the semantic view, which maintained that models are directly defined mathematical structures. He continues by examining the more recent view, which identifies models with fictional entities. He closes his chapter with an explication of what he calls the more pragmatic accounts, which hold that models can best be understood with the use of a mixed ontology.

In **Chap. 2: *Models and Theories***, *Demetris Portides* explicates the two main conceptions of the structure of scientific theories (and subsequently the two main conceptions of the theory–model relation) in the history of the philosophy of science, the received and the semantic views. He takes the reader through the main arguments that led to the collapse of the received view and gives the reader a lens by which to distinguish the different versions of the semantic view. He finally presents the main arguments against the semantic view and in doing so he explicates a more recent philosophical trend that conceives the theory–model relation as too complex to sufficiently capture with formal tools.

*Roman Frigg* and *James Nguyen*, in **Chap. 3: *Models and Representation*** begin by analyzing the concept of representation and clarifying its main characteristics and the conditions of adequacy any theory of representation should meet. They then proceed to explain the main theories of representation that have been proposed in the literature and explain with reference to their proposed set of characteristics and conditions of adequacy where each theory is found wanting. The similarity, the structuralist, the inferential, the fictionalist, and the denotational accounts of representation are all thoroughly explained and critically assessed. By doing this the authors expose and explicate many of the weaknesses of the different accounts of representation.

In **Chap. 4: *Models and Explanation***, *Alisa Bokulich* explains that by recognizing the extensive use of models in science and by realizing that models are more often than not highly idealized and incomplete

---

descriptions of phenomena that frequently incorporate fictional elements, philosophers have been led to revise previous philosophical accounts of scientific explanation. By scrutinizing different model-based accounts of scientific explanation offered in the literature and exposing the problems involved, she highlights the difficulties involved in resolving the issue of whether or not the falsehoods present in models are operative in scientific explanation.

Finally, *Nancy Nersessian* and *Miles McLeod*, in **Chap. 5: *Models and Simulations***, explicate a more recent issue that is increasingly gaining the interest of philosophers: how scientific models, i.e., the mathematical entities that scientists traditionally use to represent phenomena, relate to simulations, particularly computational simulations. They give a flavor of the character-

istics of computational simulations both in the context of well-developed overarching theories and in the context where an overarching theory is absent. The authors also highlight the epistemological significance of simulations for all such contexts by elaborating on how simulations introduce novel problems that should concern philosophers. Finally, they elaborate on the relation between simulations and other constructs of human cognition such as thought experiments.

In most cases, in all chapters the technical aspects of the philosophical arguments have been kept to a minimum in order to make them accessible even to readers working outside the sphere of the philosophy of science. Suppressing the technical aspects has not, however, introduced misrepresentation or distortion to philosophical arguments.

# The Ontology of Models

Axel Gelfert

The term *scientific model* picks out a great many things, including scale models, physical models, sets of mathematical equations, theoretical models, toy models, and so forth. This raises the question of whether a general answer to the question *What is a model?* is even possible. This chapter surveys a number of philosophical approaches that bear on the question of what, in general, a scientific model is. While some approaches aim for a unitary account that would apply to models in general, regardless of their specific features, others take as their basic starting point the manifest heterogeneity of models in scientific practice. This chapter first motivates the ontological question of what models are by reflecting on the diversity of different kinds of models and arguing that models are best understood as *functional entities*. It then provides some historical background regarding the use of analogy in science as a precursor to contemporary notions of *scientific model*. This is followed by a contrast between the syntactic and the semantic views of theories and models and their different stances toward the question of what a model is. Scientists, too, typically operate with tacit assumptions about the ontological status of models: this gives rise to what has been called the *folk ontology* of models, according to which models may be thought of as descriptions of missing (i. e., uninstantiated) systems. There is a close affinity between this view and recent philosophical positions (to be discussed in the

1.1	<b>Kinds of Models: Examples from Scientific Practice</b> .....	6
1.2	<b>The Nature and Function of Models</b> .....	8
1.3	<b>Models as Analogies and Metaphors</b> .....	10
1.4	<b>Models Versus the Received View: Sentences and Structures</b> .....	12
1.4.1	Models and the Study of Formal Languages .....	13
1.4.2	The Syntactic View of Theories .....	13
1.4.3	The Semantic View .....	14
1.4.4	Partial Structures .....	15
1.5	<b>The Folk Ontology of Models</b> .....	16
1.6	<b>Models and Fiction</b> .....	18
1.7	<b>Mixed Ontologies: Models as Mediators and Epistemic Artifacts</b> .....	20
1.7.1	Models as Mediators .....	20
1.7.2	Models as Epistemic Artifacts .....	21
1.8	<b>Summary</b> .....	21
	<b>References</b> .....	22

penultimate section) according to which models are fictions. This chapter concludes by considering various pragmatic conceptions of models, which are typically associated with what may be called *mixed ontologies*, that is, with the view that any quest for a unitary account of the nature of models is bound to be fruitless.

The philosophical discussion about models has emerged from a cluster of concerns, which span a range of theoretical, formal, and practical questions across disciplines ranging from logic and mathematics to aesthetics and artistic representations. In what follows, the term *models* will normally be taken as synonymous to *scientific models*, and any departure from this usage – for example, when discussing the use of models in non-scientific settings – will either be indicated explicitly or will be clear from context. Focusing on scientific models helps to clarify matters, but still leaves a wide

range of competing philosophical approaches for discussion. This chapter will summarize and critically discuss a number of such approaches, especially those that shed light on the question *what is a model?*; these will range from views that, by now, are of largely historical interest to recent proposals at the cutting edge of the philosophy of science. While the emphasis throughout will be on the ontology of models, it will often be necessary to also reflect on their function, use, and construction. This is not meant to duplicate the discussion provided in other chapters of this handbook; rather, it is the natu-



ral result of scientific models having traditionally been defined either in terms of their function (e.g., to provide representations of target systems) or via their relation to other (purportedly) better understood entities, such as scientific theories.

The rest of this chapter is organized as follows: Sect. 1.1 will set the scene by introducing a number of examples of scientific models, thereby raising the question of what degree of unity any philosophical account of scientific models can reasonably aspire to. Section 1.2 will characterize models as functional entities and will provide a general taxonomy for how to classify various possible philosophical approaches. A first important class of specific accounts, going back to nineteenth-century scientists and philosophers, will be discussed in Sect. 1.3, which focuses on models as analogies. Section 1.4 is devoted to formal approaches

that dominated much of twentieth-century discussion of scientific models. In particular, it will survey the syntactic view of theories and models and its main competitor, the semantic view, along with recent formal approaches (such as the partial structures approach) which aim to address the shortcomings of their predecessors. Section 1.5 provides a sketch of what has been called the *folk ontology* of models – that is, a commonly shared set of assumptions that inform the views of scientific practitioners. On this view, models are place-holders for *imaginary concrete systems* and as such are not unlike fictions. The implications of fictionalism about models are discussed in Sect. 1.6. Finally, in Sect. 1.7, recent pragmatic accounts are discussed, which give rise to what may be called a *mixed ontology*, according to which models are best conceived of as a heterogeneous mixture of elements.

## 1.1 Kinds of Models: Examples from Scientific Practice

Models can be found across a wide range of scientific contexts and disciplines. Examples include the Bohr model of the atom (still used today in the context of science education), the billiard ball model of gases, the DNA double helix model, scale models in engineering, the Lotka–Volterra model of predator–prey dynamics in population biology, agent-based models in economics, the Mississippi River Basin model (which is a 200 acres hydraulic model of the waterways in the entire Mississippi River Basin), and general circulation models (GCMs), which allow scientists to run simulations of Earth’s climate system. The list could be continued indefinitely, with the number of models across the natural and social sciences growing day by day.

In philosophical discussions of scientific models, the situation is hardly any different. The *Stanford Encyclopedia of Philosophy* gives the following list of model types that have been discussed by philosophers of science [1.1]:

“Probing models, phenomenological models, computational models, developmental models, explanatory models, impoverished models, testing models, idealized models, theoretical models, scale models, heuristic models, caricature models, didactic models, fantasy models, toy models, imaginary models, mathematical models, substitute models, iconic models, formal models, analogue models and instrumental models.”

The proliferation of models and model types, in the sciences as well as in the philosophical literature, led Goodman to lament in his 1968 *Languages of Art* [1.2, p. 171]: “Few terms are used in popular and scientific

discourse more promiscuously than *model*.” If this was true of science and popular discourse in the late 1960s, it is all the more true of the twenty-first century philosophy of science.

As an example of a physics-based model, consider the *Ising model*, proposed in 1925 by the German physicist Ernst Ising as a model of ferromagnetism in certain metals. The model starts from the idea that a macroscopic magnet can be thought of as a collection of elementary magnets, whose orientation determines the overall magnetization. If all the elementary magnets are aligned along the same axis, then the system will be perfectly ordered and will display a maximum value of the magnetization. In the simplest one-dimensional (1-D) case, such a state can be visualized as a chain of *elementary magnets*, all pointing the same way

... ↑↑↑↑↑↑↑↑↑↑↑↑↑↑ ...

The alignment of elementary magnets can be brought about either by a sufficiently strong external magnetic field or it can occur spontaneously, as will happen below a critical temperature, when certain substances (such as iron and nickel) undergo a ferromagnetic phase transition. Whether or not a system will undergo a phase transition, according to thermodynamics, depends on its energy function, which in turn is determined by the interactions between the component parts of the system. For example, if neighboring *elementary magnets* interact in such a way as to favor alignment, there is a good chance that a spontaneous phase transition may occur below a certain temperature. The energy function, then, is crucial to the model and, in the case of the Ising

model, is defined as

$$E = - \sum_{i,j} J_{ij} S_i S_j ,$$

with the variable  $S_i$  representing the orientation (+1 or -1) of an elementary magnet at site  $i$  in the crystal lattice and  $J_{ij}$  representing the strength of interaction between two such elementary magnets at different lattice sites  $i$  and  $j$ .

Contrast this with *model organisms* in biology, the most famous example of which is the fruit fly *Drosophila melanogaster*. Model organisms are real organisms – actual plants and animals that are alive and can reproduce – yet they are used as representations either of another organism (e.g., when rats are used in place of humans in medical research) or of a biological phenomenon that is more universal (e.g., when fruit flies are used to study the effects of crossover between homologous chromosomes). Model organisms are often bred for specific purposes and are subject to artificial selection pressures, so as to purify and *standardize* certain features (e.g., genetic defects or variants) that would not normally occur, or would occur only occasionally, in populations in the wild. As *Ankeny* and *Leonelli* put it, in their ideal form “model organisms are thought to be a relatively simplified form of the class of organism of interest” [1.3, p. 318]; yet it often takes considerable effort to work out the actual relationships between the model organism and its target system (whether it be a certain biological phenomenon or a specific class of target organisms). Tractability and various experimental desiderata – for example, a short life cycle (to allow for quick breeding) and a relatively small and compact genome (to allow for the quick identification of variants) – take precedence over theoretical questions in the choice of model organisms; unlike for the Ising model, there is no simple mathematical formula that one can rely on to study how one’s model behaves, only the messy world of real, living systems.

The Ising model of ferromagnetism and model organisms such as *Drosophila melanogaster* may be at opposite ends of the spectrum of scientific models. Yet the diversity among those models that occupy the middle ground between theoretical description and experimental system is no less bewildering. How, one might wonder, can a philosophical account of scientific models aspire to any degree of unity or generality in the light of such variety? One obvious strategy is to begin by drawing distinctions between different overarching types of models. Thus, *Black* [1.4] distinguishes between four such types:

1. Scale models

2. Analog models
3. Mathematical models
4. Theoretical models.

The basic idea of scale and analog models is straightforward: a scale model increases or decreases certain (e.g., spatial) features of the target system, so as to render them more manageable in the model; an analog model also involves the change of medium (as in once popular hydraulic models of the economy, where the flow of money was represented by the flow of liquids through a system of pumps and valves). Mathematical models are constructed by first identifying a number of relevant variables and then developing empirical hypotheses concerning the relations that may hold between the variables; through (often drastic) simplification, a set of mathematical equations is derived, which may then be evaluated analytically or numerically and tested against novel observations. Theoretical models, finally, begin usually by extrapolating imaginatively from a set of observed facts and regularities, positing new entities and mechanisms, which may be integrated into a possible theoretical account of a phenomenon; comparison with empirical data usually comes only at a later stage, once the model has been formulated in a coherent way.

*Achinstein* [1.5] includes mathematical models in his definition of *theoretical model*, and proposes an analysis in terms of sets of assumptions about a model’s target system. This allows him to include Bohr’s model of the atom, the DNA double-helix model (considered as a set of structural hypotheses rather than as a physical ball-and-stick model), the Ising model, and the Lotka–Volterra model among the class of theoretical systems. Typically, when a scientist constructs a theoretical model, she will help herself to certain established principles of a more fundamental theory to which she is committed. These will then be adapted or modified, notably by introducing various new assumptions specific to the case at hand. Typically, an inner structure or mechanism is posited which is thought to explain the features of the target system. At the same time, there is the (often explicit) acknowledgment that the target system is far more complex than the model is able to capture: in this sense, a theoretical model is believed by the practitioner to be false as a description of the target system. However, this acknowledgment of the limits of applicability of models also allows researchers to simultaneously use different models of the same target system alongside each other. Thus understood, theoretical models usually involve the combination of general theoretical principles and specific auxiliary assumptions, which may only be valid for a narrow range of parameters.

## 1.2 The Nature and Function of Models

The great variety of models employed in scientific practice, as illustrated by the long list given in the preceding section, suggests two things. First, it makes vivid just how central the use of models is to the scientific enterprise and to the self-image of scientists. As *von Neumann* put it, with some hyperbole [1.6, p. 492]: “The sciences do not try to explain, they hardly even try to interpret, they mainly make models.” Whatever shape and form the scientific enterprise might take without the use of models, it seems safe to say that it would not look anything like science as we presently know it. Second, one might wonder whether it is at all reasonable to look for a unitary philosophical account of models. Given the range of things we call *models*, and the diversity of uses to which they are being put, it may simply not be possible to give a one-size-fits-all answer to the question *what is a model?* This has led some commentators to propose quietism as the only viable attitude toward ontological questions concerning models and theories. As *French* puts it [1.7, p. 245],

“whereas positing the reality of quarks or genes may contribute to the explanation of certain features of the physical world, adopting a similar approach toward theories and models – that is, reifying them as entities for which a single unificatory account can be given – does nothing to explain the features of scientific practice.”

While there are good grounds for thinking that quietism should only be a position of last resort in philosophy, the sentiment expressed by *French* may go some way toward explaining why there has been a relative dearth of philosophical work concerning the ontology of models. The neglect of ontological questions concerning models has been remarked upon by a number of contributors, many of whom, like *Connessa*, find it [1.8, p. 194]

“surprising if one considers the amount of interest raised by analogous questions about the ontology and epistemology of mathematical objects in the philosophy of mathematics.”

A partial explanation of this discrepancy lies in the arguably greater heterogeneity in what the term *scientific models* is commonly thought to refer to, namely, anything from physical ball-and-stick models of chemical molecules to mathematical models formulated in terms of differential equations. (If we routinely included dividers, compasses, set squares, and other technical drawing tools among, say, the class of *geometrical entities*, the ontology of mathematical entities, too, would quickly become rather unwieldy!)

In the absence of any widely accepted unified account of models – let alone one that would provide a conclusive answer to ontological questions arising from models – it may be natural to assume, as indeed many contributors to the debate have done, that “if all scientific models have something in common, this is not their *nature* but their *function*” [1.8, p. 194]. One option would be to follow the quietist strategy concerning the ontology of models and “refuse to engage with this issue and ask, instead, how can we best represent these features [and functions of models] in order that we can understand” [1.7, p. 245] the practice of scientific modeling. Alternatively, however, one might simply accept that the function of models in scientific inquiry is our best – and perhaps only – guide when exploring answers to the question *what is a model?*. At the very least, it is not obvious that an exploration of the ontological aspects of models is necessarily fruitless or misguided. *Ducheyne* puts this nicely when he argues that [1.9, p. 120],

“if we accept that models are functional entities, it should come as no surprise that when we deal with scientific models ontologically, we cannot remain silent on how such models function as carriers of scientific knowledge.”

As a working assumption, then, let us treat scientific models as *functional entities* and explore how much ontological unity – over and above their *mere* functional role – we can give to the notion of *scientific model*.

Two broad classes of functional characterizations of models can be distinguished, according to which it is either *instantiation* or *representation* that lie at the heart of how models function. As *Giere* [1.10] sees it, on the *instantial view*, models instantiate the axioms of a theory, where the latter is understood as being comprised of linguistic statements, including mathematical statements and equations. (For an elaboration of how such an account might turn out, see Sect. 1.4.) By contrast, on the *representational view*, “language connects not directly with the world, but rather with a model, whose characteristics may be precisely defined”; the model then connects with the world “by way of similarity between a model and the designated parts of the world” [1.10, p. 156]. Other proponents of the representational view have de-emphasized the role of similarity, while still endorsing representation as one of the key functions of scientific models. Generally speaking, proponents of the representational view consider models to be “tools for *representing the world*,” whereas those who favor the *instantial view* regard them

primarily as “providing a means for interpreting formal systems” [1.10, p. 44].

Within the class of representational views, one can further distinguish between views that emphasize the *informational* aspects of models and those that take their *pragmatic* aspects to be more central. *Chakravartty* nicely characterizes the informational variety of the representational view as follows [1.11, p. 198]:

“The idea here is that a scientific representation is something that bears an objective relation to the thing it represents, on the basis of which it contains information regarding that aspect of the world.”

The term *objective* here simply means that the requisite relation obtains independently of the model user’s beliefs or intentions as well as independently of the specific representational conventions he or she might be employing. *Giere*’s similarity-based view of representation – according to which scientific models represent in virtue of their being similar to their target systems in certain specifiable ways – would be an example of such an informational view similarity, as construed by *Giere*, is a relation that holds between the model and its target, irrespective of a model user’s beliefs or intentions, and regardless of the cognitive uses to which he or she might put the model. Other philosophical positions that are closely aligned with the informational approach might posit that, for a model to represent its target, the two must stand in a relation of isomorphism, partial isomorphism, or homomorphism to one another.

By contrast, the *pragmatic* variety of the representational view of models posits that models function as representations of their targets in virtue of the cognitive uses to which human reasoners put them. The basic idea is that a scientific model facilitates certain cognitive activities – such as the drawing of inferences about a target system, the derivation of predictions, or perhaps a deepening of the scientific understanding – on the part of its user and, therefore, necessarily involves the latter’s cognitive interests, beliefs, or intentions. *Hughes* [1.12], for example, emphasizes the interplay of three cognitive–theoretical processes – denotation, demonstration, and interpretation – which jointly give rise to the representational capacity of (theoretical) models in science. On *Hughes*’ (aptly named) *DDI* account of model-based representation, *denotation* accounts for the fact that theoretical elements of a model

purport to refer to elements in the physical world. The possibility of *demonstration* from within a model – in particular, the successful mathematical derivation of results for models that lend themselves to mathematical derivation techniques – attests both to the models having a nontrivial internal dynamic and to its being a viable object of fruitful theoretical investigation. Through successful *interpretation*, a model user then relates the theoretically derived results back to the physical world, including the model’s target system. Clearly, the *DDI* account depends crucially on there being someone who engages in the activities of interpreting and demonstrating – that is, it depends on the cognitive activities of human agents, who will inevitably draw on their background knowledge, cognitive interests, and derivational skills in establishing the requisite relations for bringing about representation.

The contrast between informational and pragmatic approaches to model-based representation roughly maps onto another contrast, between what *Knuuttila* has dubbed *dyadic* and *triadic* approaches. The former takes “the model–target dyad as a basic unit of analysis concerning models and their epistemic values” [1.13, p. 142]. This coheres well with the informational approach which, as discussed, tends to regard models as (often abstract) structures that stand in a relation of isomorphism, or partial isomorphism, to a target system. By contrast, *triadic* accounts – in line with pragmatic views of model-based representation – based representation shift attention away from models and the abstract relations they stand in, toward modeling as a theoretical activity pursued by human agents with cognitive interests, intentions, and beliefs. On this account, model-based representation cannot simply be a matter of any abstract relationship between the model and a target system since one cannot, as *Suárez* puts it, “reduce the essentially intentional judgments of representation users to facts about the source and target object or systems and their properties” [1.14, p. 768]. Therefore, so the suggestion goes, the model–target dyad needs to be replaced by a three-place relation between the model, its target, and the model user. *Suárez*, for example, proposes an inferentialist account of model-based representation, according to which a successful model must allow “competent and informed agents to draw specific inferences regarding” [1.14, p. 773] the target system – thereby making the representational success of a model dependent on the qualities of a (putative) model user.

### 1.3 Models as Analogies and Metaphors

Some scholars trace the emergence of the concept of a *scientific model* to the second half of the nineteenth century [1.15]. Applying our contemporary concept of *model* to past episodes in the history of science, we can of course identify prior instances of models being employed in science; however, until the nineteenth century scientists were engaged in little systematic self-reflection on the uses and limitations of models. Philosophy of science took even longer to pay attention to models in science, focusing instead on the role and significance of scientific theories. Only from the middle of the twentieth century onward did philosophical interest in models acquire the requisite momentum to carry the debate forward. Yet in both science and philosophy, the term *model* underwent important transformations, so it will be important to identify some of these shifts, in order to avoid unnecessary ambiguity and confusion in our exploration of the question *What is a model?*

Take, for example, Duhem's dismissal, in 1914, of what he takes to be the excessive use of models in Maxwell's theory of electromagnetism, as presented in an English textbook published at the end of the nineteenth century [1.16, p. 7]:

“Here is a book intended to expound the modern theories of electricity and to expound a new theory. In it there are nothing but strings which move round pulleys which roll around drums, which go through pearl beads, which carry weights; and tubes which pump water while others swell and contract; toothed wheels which are geared to one another and engage hooks. We thought we were entering the tranquil and neatly ordered abode of reason, but we find ourselves in a factory.”

What Duhem is mocking in this passage, which is taken from a chapter titled *Abstract Theories and Mechanical Models*, is a style of reasoning that is dominated by the desire to *visualize* physical processes in purely mechanical terms. His hostility is thus directed at *mechanical* models only – as the implied contrast in the chapter title makes clear – and does not extend to the more liberal understanding of the term *scientific model* in philosophy of science today.

Indeed, when it comes to the use of *analogy* in science, Duhem is much more forgiving. The term *analogy*, which derives from the Greek expression for *proportion*, itself has multiple uses, depending on whether one considers its use as a rhetorical device or as a tool for scientific understanding. Its general form is that of “pointing to a resemblance between relations in two different domains, that is, *A* is related to *B* like *C* is related to *D*” [1.17, p. 110]. An analogy may be considered

merely *formal*, when only the relations (but not the relata) resemble another, or it may be *material*, when the relata from the two domains (i. e., *A* and *B* on one side, *C* and *D* on the other) have certain attributes or characteristics in common. Duhem's understanding of *analogy* is more specific, in that he conceives of analogy as being a relation between two sets of statements, such as between one theory and another [1.16, p. 97]:

“Analogies consist in bringing together two abstract systems; either one of them already known serves to help us guess the form of the other not yet known, or both being formulated, they clarify the other. There is nothing here that can astonish the most rigorous logician, but there is nothing either that recalls the procedures dear to ample but shallow minds.”

Consider the following example: When Christiaan Huygens (1629–1695) proposed his theory of light, he did so on the basis of *analogy* with the theory of sound waves: the relations between the various attributes and characteristics of light are similar to those described by acoustic theory for the rather different domain of sound. Thus understood, analogy becomes a legitimate instrument for learning about one domain on the basis of what we know about another. In modern parlance, we might want to say that sound waves provided Huygens with a good *theoretical model* – at least given what was known at the time – for the behavior of light.

There is, however, a risk of ambiguity in that last sentence – an ambiguity which, as Mellor [1.18, p. 283] has argued, it would be wrong to consider harmless. Saying that *sound waves provide a good model for the theory of light* appears to equate the model with the *sound waves* – as though one physical object (sound waves) could be identified with the model. At first sight, this might seem unproblematic, given that, as far as wave-like behavior is concerned, we do take light and sound to be relevantly analogous. However, while it is indeed the case that “some of the constructs called *analogy* in the nineteenth century would today be routinely referred to as *models*” [1.19, p. 46], it is important to distinguish between, on the one hand, *analogy* as the similarity relation that exists between a theory and another set of statements and, on the other hand, the latter set of statements as the *analog of the theory*. Furthermore, we need to distinguish between the analog (e.g., the theory of sound waves, in Huygens's case) and the set of entities *of which the analog is true* (e.g., the sound waves themselves). (On this point, see [1.18, p. 283].) What Duhem resents about the naïve use of what he refers to as *mechanical models* is the hasty conflation of the visualized entities – (imaginary) pulleys, drums,

pearl beads, and toothed wheels – with what is *in fact* scientifically valuable, namely the relation of analogy that exists between, say, the theory of light and the theory of sound.

This interpretation resolves an often mentioned tension – partly perpetuated by Duhem himself, through his identification of different styles of reasoning (the *English* style of physics with its emphasis on mechanical models, and the *Continental* style which prizes mathematical principles above all) – between Duhem’s account of models and that of the English physicist Norman Campbell. Thus, *Hesse*, in her seminal essay *Models and Analogies in Science* [1.20], imagines a dialogue between a *Campbellian* and a *Duhemist*. At the start of the dialogue, the *Campbellian* attributes to the *Duhemist* the following view: “I imagine that along with most contemporary philosophers of science, you would wish to say that the use of models or analogs is not essential to scientific theorizing and that [...] the theory as a whole does not require to be interpreted by means of any model.” To this, the *Duhemist*, who admits that “models may be useful guides in suggesting theories,” replies: “When we have found an acceptable theory, any model that may have led us to it can be thrown away. Kekulé is said to have arrived at the structure of the benzene ring after dreaming of a snake with its tail in its mouth, but no account of the snake appears in the textbooks of organic chemistry.” The *Campbellian*’s rejoinder is as follows: “I, on the other hand, want to argue that models in some sense are essential to the logic of scientific theories” [1.20, pp. 8–9]. The quoted part of *Hesse*’s dialogue has often been interpreted as suggesting that the bone of contention between Duhem and Campbell is the status of *models in general* (in the modern sense that includes theoretical models), with Campbell arguing in favor and Duhem arguing against. But we have already seen that Duhem, using the language of *analogy*, does allow for theoretical models to play an important role in science. This apparent tension can be resolved by being more precise about the target of Duhem’s criticism: “Kekulé’s snake dream might illustrate the use of a visualizable model, but it certainly does not illustrate the use of an analogy, in Duhem and Campbell’s sense” [1.18, p. 285]. In other words, Duhem is not opposed to scientific models in general, but to its mechanical variety in particular. And, on the point of over-reliance on mechanical models, *Campbell*, too, recognizes that dogmatic attachment to such a style of reasoning is *open to criticism*. Such a dogmatic view would hold “that theories are completely satisfactory only if the analogy on which they are based is mechanical, that is to say, if the analogy is with the laws of mechanics” [1.21, p. 154]. Campbell is clearly more sympathetic than Duhem toward our “craving for

mechanical theories,” which he takes to be firmly rooted in our psychology. But he insists that [1.21, p. 156]

“we should notice that the considerations which have been offered justify only the attempt to adopt some form of theory involving ideas closely related to those of force and motion; it does not justify the attempt to force all such theories into the Newtonian mold.”

To be sure, significant differences between Duhem and Campbell remain, notably concerning what *kinds* of uses of analogies in science (or, in today’s terminology, of scientific – including theoretical – models) are appropriate. For Duhem, such uses are limited to a heuristic role in the discovery of scientific theories. By contrast, *Campbell* claims that “in order that a theory may be valuable [...] it must display analogy” [1.21, p. 129] – though it should be emphasized again, not necessarily analogy *of the mechanical sort*. (As *Mellor* argues, Duhem and Campbell differ chiefly in their views of scientific theories and less so in their take on analogy, with Duhem adopting a more *static* perspective regarding theories and Campbell taking a more realist perspective [1.18].)

It should be said, though, that *Hesse*’s *Campbellian* and *Duhemist* are at least partly intended as caricatures and serve as a foil for *Hesse*’s own account of models as analogies. The account hinges on a three-part distinction between *positive*, *negative*, and *neutral* analogies [1.20]. Using the billiard ball model of gases as her primary example, *Hesse* notes that some characteristics are shared between the billiard balls and the gas atoms (or, rather, are ascribed by the billiard ball model to the gas atoms); these include velocity, momentum, and collision. Together, these constitute the *positive* analogy. Those properties we know to belong to billiard balls, but not to gas atoms – such as color – constitute the *negative* analogy of the model. However, there will typically be properties of the model (i. e., the billiard ball system) of which we do not (yet) know whether they also apply to its target (in this case, the gas atoms). These form the *neutral* analogy of the model. Far from being unimportant, the neutral analogy is crucial to the fruitful use of models in scientific inquiry, since it holds out the promise of acquiring new knowledge about the target system by studying the model in its place [1.20, p. 10]:

“If gases are really like collections of billiard balls, except in regard to the known negative analogy, then from our knowledge of the mechanics of billiard balls, we may be able to make new predictions about the expected behavior of gases.”

In dealing with scientific models we may choose to disregard the negative analogy (which results in what Hesse calls *model*<sub>1</sub>) and consider only the known positive and neutral analogies – that is, only those properties that are shared, or for all we know may turn out to be shared, between the target system and its analog. (On the terminology discussed in Sect. 1.1, due to Black and Achinstein, *model*<sub>1</sub> would qualify as a *theoretical model*.) This, Hesse argues, typically describes our use of models for the purpose of explanation: we resolve to treat *model*<sub>1</sub> as taking the place of the phenomena themselves. Alternatively, we may actively include the negative analogy in our considerations, resulting in what Hesse calls *model*<sub>2</sub> or a form of analog model. Given that, let us assume, the model system (e.g., the billiard balls) was chosen because it was observable – or, at any rate, more accessible than the target system (e.g., the gas) – *model*<sub>2</sub> allows us to study the similarities and dissimilarities between the two analogous domains; *model*<sub>2</sub>, qua being a model for its target, thus has a deeper structure than the system of billiard balls considered in isolation – and, like *model*<sub>1</sub>, importantly includes the neutral analogy, which holds out the promise of novel insights and predictions. As Hesse puts it, in the voice of her Campbellian interlocutor [1.20, pp. 12–13]:

“My whole argument is going to depend on these features [of the neutral analogy] and so I want to make it clear that I am not dealing with static and formalized theories, corresponding only to the known positive analogy, but with theories in the process of growth.”

Models have been discussed not only in terms of analogy, but also in terms of metaphor. *Metaphor*, more explicitly than *analogy*, refers to the linguistic realm:

a metaphor is a linguistic expression that involves at least one part that is being transferred from a domain of discourse where it is common to another – the target domain – where it is uncommon. The existence of an analogy may facilitate such a transfer of linguistic expression; at the same time, it is entirely possible that “it is the metaphor that prompts the recognition of analogy” [1.17, p. 114] – both are compatible with one another and neither is obviously prior to the other. Metaphorical language is widespread in science, not just in connection with models: for example, physicists routinely speak of *black holes* and *quantum tunneling* as important predictions of general relativity theory and quantum theory, respectively. Yet, as Soskice and Harré note, there is a special affinity between models and metaphor [1.22, p. 302]:

“The relationship of model and metaphor is this: if we use the image of a fluid to explicate the supposed action of the electrical energy, we say that the fluid is functioning as a model for our conception of the nature of electricity. If, however, we then go on to speak of the *rate of flow* of an *electrical current*, we are using metaphorical language based on the fluid model.”

In spite of this affinity, it would not be fruitful to simply equate the two – let alone jump to the conclusion that, in the notion of *metaphor*, we have found an answer to the question *What is a model?*. Models and metaphors both issue in descriptions, and as such they may draw on analogies we have identified between two otherwise distinct domains; more, however, needs to be said about the nature of the relations that need to be in place for something to be considered a (successful) model of its target system or phenomenon.

## 1.4 Models Versus the Received View: Sentences and Structures

Much of the philosophical debate about models is indebted to model theory as a branch of (first-order) mathematical logic. Two philosophical frameworks for thinking about scientific models and theories – the *syntactic view* of models and theories and its main competitor, the *semantic view* – can be traced back to these origins; they are the topic of this section. (For a more extensive discussion, see also other chapters in this handbook.) The syntactic view (Sect. 1.4.2) is closely aligned with logical positivism, which dominated much anglophone philosophy of science until the mid-1960s, and is sometimes referred to as *the received view*. Given

that less rigid approaches and an overarching movement toward pluralism have reshaped the philosophy of science over the past half-century or so, this expression is somewhat dated; to make matters worse, other contributors to the debate have, over time, come to apply the same label to the syntactic view’s main competitor, the semantic view of models and theories. Instead of adjudicating which position deserves this dubious honor, the present section will discuss how each view conceives of models. Before doing so, however, a few preliminaries are in order concerning the competing views’ joint origins in logical model theory.

### 1.4.1 Models and the Study of Formal Languages

Model theory originated as the study of formal languages and their interpretations, starting from a Tarski-style truth theory based only on notions from syntax and set theory. On a broader understanding, the restriction to formal languages may be dropped, so as to include scientific languages (which are often closer to natural language than to logic), or even natural languages. However, the distinction between the syntax and the semantics of a language, which is sharpest in logic, also provides a useful framework for studying scientific languages and has guided the development of both the syntactic and the semantic views of theories and models. The *syntax* of a language  $L$  is made up of the vocabulary of  $L$ , along with the rules that determine which sequence of symbols counts as a well-formed expression in  $L$ ; in turn, the *semantics* of  $L$  provides interpretations of the symbolic expressions in  $L$ , by mapping them onto another relational structure  $R$ , such that all well-formed expressions in  $L$  are rendered intelligible (e.g., via rules of composition) and can be assessed in terms of their truth or falsity in  $R$ .

The contrast between the syntax and the semantics of a language allows for two different approaches to the notion of a *theory*. A theory  $T$  may either be defined syntactically, as the set of all those sentences that can be derived, through a proper application of the syntactic rules, from a set of axioms (i. e., statements that are taken to be fundamental); or it may be defined semantically, as all those (first-order) sentences that a particular structure,  $M$ , satisfies. An example of the former would be Euclidean geometry, which consists of five axioms and all the theorems derivable from them using geometrical rules; an example of the latter would be group theory, which simply consists of all those first-order sentences that a set of groups – definable in terms of set-theoretic entities – satisfies. (This example, and much of the short summary in this section, is owed to [1.23]; for further discussion, see references therein.) The syntactic and semantic definitions of what a theory is are closely related: starting from the semantic definition, to see whether a particular structure  $M$  is a model of an axiomatizable first-order theory  $T$ , all that one needs to show is that  $M$  satisfies the axioms.

### 1.4.2 The Syntactic View of Theories

The syntactic view of theories originated from the combination of the insights – or, to put it a little more cautiously, fundamental tenets – of two research programs: the philosophical program, aligned with Pierre Duhem (Sect. 1.3) and Henri Poincaré, of treating

(physical) theories as systems of hypotheses designed to *save the phenomena*, and the mathematical program, pioneered by David Hilbert, which sought to formalize (mathematical) theories as axiomatic systems. By combining the two, it seemed possible to identify a theory with the set of logical consequences that could be derived from its fundamental principles (which were to be treated as axioms), using only the rules of the language in which the theory was formulated. In spite of its emphasis on syntax, the syntactic view is not entirely divorced from questions of semantics. When it comes to scientific theories, we are almost always dealing with *interpreted* sets of sentences, some of which – the fundamental principles or axioms – are more basic than others, with the rest derivable using syntactic rules. The question then arises at which level interpretation of the various elements of a theory is to take place. This is where the slogan *to save the phenomena* points us in the right direction: on the syntactic view, interpretation only properly enters at the level of matching singular theoretical predictions, formulated in strictly observational terms, with the observable phenomena. Higher level interpretations – for example, pertaining to purely theoretical terms of a theory (such as posited unobservable entities, causal mechanisms, laws, etc.) – would be addressed through *correspondence rules*, which offered at least a partial interpretation, so that *some* of the meaning of such higher level terms of a theory could be linked up with observational sentences.

As an example, consider the example of classical mechanics. Similar to how Euclidean geometry can be fully derived from a set of five axioms, classical mechanics is fully determined by Newton's laws of mechanics. At a purely formal level, it is possible to provide a fully syntactic axiomatization in terms of the relevant symbols, variables, and rules for their manipulation – that is, in terms of what Rudolf Carnap calls the *calculus of mechanics*. If one takes the latter as one's starting point, it requires interpretation of the results derived from within this formal framework, in order for the calculus to be recognizable as a theory of mechanics, that is, of physical phenomena. In the case of mechanics, we may have no difficulty stating the axioms in the form of the (physically interpreted) *Newtonian laws of mechanics*, but in other cases – perhaps in quantum mechanics – making this connection with observables may not be so straightforward. As Carnap notes [1.24, p. 57]:

“[t]he relation of this theory [= the physically interpreted theory of mechanics] to the calculus of mechanics is entirely analogous to the relation of physical to mathematical geometry.”



As in the Euclidean case, the syntactic view identifies the theory with a formal language or calculus (including, in the case of scientific theories, relevant correspondence rules), “whose interpretation – what the calculus is a theory of – is fixed at the point of application” [1.25, p. 125].

On the syntactic view of theories, models play at best a very marginal role as limiting cases or approximations. This is for two reasons. First, since the nonobservational part of the theory – that is, the *theory proper*, as one might put it – does not admit of direct interpretation, the route to constructing theoretical models on the basis of our directly interpreting the core ingredients of the theory is obstructed. Interpretation at the level of observational statements, while still available to us, is insufficient to imbue models with anything other than a purely *one-off* auxiliary role. Second, as *Cartwright* has pointedly argued in criticism directed at both the syntactic and the semantic views, there is a shared – mistaken – assumption that theories are a bit like vending machines [1.26, p. 247]:

“[Y]ou feed it input in certain prescribed forms for the desired output; it gurgitates for a while; then it drops out the sought-for-representation, plonk, on the tray, fully formed, as Athena from the brain of Zeus.”

This limits what we can do with models, in that there are only two stages [1.26, p. 247]:

“First, eyeballing the phenomenon, measuring it up, trying to see what can be abstracted from it that has the right form and combination that the vending machine can take as input; secondly, [...] we do either tedious deduction or clever approximation to get a facsimile of the output the vending machine would produce.”

Even if this caricature seems a little too extreme, the fact remains that, by modeling theories after first-order formal languages, the syntactic view limits our understanding of what theories and models are and what we can do with them.

### 1.4.3 The Semantic View

One standard criticism of the syntactic view is that it conflates scientific theories with their linguistic formulations. Proponents of the semantic view argue that by adding a layer of (nonlinguistic) structures between the linguistic formulations of theories and our assessment of them, one can side-step many of the problems faced by the syntactic view. According to the semantic view, a theory should be thought of as the set of set-theoretic structures that satisfy the different linguis-

tic formulations of the theory. A structure that provides an interpretation for, and makes true, the set of sentences associated with a specific linguistic formulation of the theory is called a *model of the theory*. Hence, the semantic view is often characterized as conceiving of theories as *collections of models*. This not only puts models – where these are to be understood in the logical sense outlined earlier – center stage in our account of scientific theories, but also renders the latter fundamentally *extra-linguistic* entities.

An apt characterization of the semantic view is given by *Suppe* as follows [1.27, pp. 82–83]:

“This suggests that theories be construed as propounded abstract *structures* serving as models for sets of interpreted sentences that constitute the linguistic formulations. [...] [W]hat the theory does is directly describe the behavior of abstract systems, known as *physical systems*, whose behaviors depend only on the selected parameters. However, physical systems are abstract replicas of actual phenomena, being what the phenomena *would have been* if no other parameters exerted an influence.”

According to a much-quoted remark by one of the main early proponents of the semantic view, *Suppes*, “the meaning of the concept of model is the same in mathematics and in the empirical sciences.” However, as *Suppe’s* quote above makes clear, models in science have additional roles to play, and it is perhaps worth noting that *Suppes* himself immediately continues: “The difference to be found in these disciplines is to be found in their use of the concept” [1.28, p. 289]. Supporters of the semantic view often claim that it is closer to the scientific practices of modeling and theorizing than the syntactic view. On this view, according to *van Fraassen* [1.29, p. 64],

“[t]o present a theory is to specify a family of structures, its *models*; and secondly, to specify certain parts of those models (the *empirical substructures*) as candidates for the direct representation of observable phenomena.”

Unlike what the syntactic view suggests, scientists do not typically formulate abstract theoretical axioms and only interpret them at the point of their application to observable phenomena; rather, “scientists build in their mind’s eye systems of abstract objects whose properties or behavior satisfy certain constraint (including law)” [1.23, p. 154] – that is, they engage in the construction of theoretical models.

Unlike the syntactic view, then, the semantic view appears to give a more definite answer to the question *what is a model?* In line with the account sketched so far, *a model of a theory is simply a (typically extra-*

linguistic) structure that provides an interpretation for, and makes true, the set of axioms associated with the theory (assuming that the theory is axiomatizable). Yet it is not clear that, in applying their view to actual scientific theories, the semanticists always heed their own advice to treat models as both *giving an interpretation*, and *ensuring the truth*, of a set of statements. More importantly, the model-theoretic account demands that, in a manner of speaking, a model should fulfil its truth-making function *in virtue of* providing an interpretation for a set of sentences. Other ways of ensuring truth – for example by limiting the domain of discourse for a set of fully interpreted sentences, thereby ensuring that the latter will happen to be true – should not qualify. Yet, as Thomson-Jones [1.30] has argued, purported applications of the semantic view often stray from the original model-theoretic motivation. As an example, consider Suppes’ *axiomatization* of Newtonian particle physics. (The rest of this subsection follows [1.30, pp. 530–531].) Suppes [1.31] begins with the following definition (in slightly modified form)

**Definition 1.1**

A system  $\beta = \langle P, T, s, m, f, g \rangle$  is a model of particle mechanics if and only if the following seven axioms are satisfied:

*Kinematical axioms:*

- 1 The set  $P$  is finite and nonempty
- 2 The set  $T$  is an interval of real numbers
- 3 For  $p$  in  $P$ ,  $s_p$  is twice differentiable.

*Dynamical axioms:*

- 4 For  $p$  in  $P$ ,  $m(p)$  is a positive real number
- 5 For  $p$  and  $q$  in  $P$  and  $t$  in  $T$ ,

$$f(p, q, t) = -f(q, p, t).$$

- 6 For  $p$  and  $q$  in  $P$  and  $t$  in  $T$ ,

$$s(p, t) \times f(p, q, t) = -s(q, t) \times f(q, p, t).$$

- 7 For  $p$  in  $P$  and  $t$  in  $T$ ,

$$m(p)D^2s_p(t) = \sum_{q \in P} f(p, q, t) + g(p, t).$$

At first sight, this presentation adheres to core ideas that motivate the semantic view. It sets out to define an extra-linguistic entity,  $\beta$ , in terms of a set-theoretical predicate; the entities to which the predicate applies are then to be singled out on the basis of the seven axioms. But as Thomson-Jones points out, a specific model  $S$  defined in this way “is not a serious interpreter of the

predicate or the axioms that compose it” [1.30, p. 531]; it merely fits a structure to the description provided by the fully interpreted axioms (1)–(7), and in this way ensures that they are satisfied, but it does not make them come out true in virtue of providing an interpretation (i. e., by invoking semantic theory). To Thomson-Jones, this suggests that identifying scientific models with truth-making structures in the model-theoretic sense may, at least in the sciences, be an unfulfilled promise of the semantic view; instead, he argues, we should settle for a less ambitious (but still informative) definition of a model as “a mathematical structure used to represent a (type of) system under study” [1.30, p. 525].

**1.4.4 Partial Structures**

Part of the motivation for the semantic view was its perceived greater ability to account for how scientists actually go about developing models and theories. Even so, critics have claimed that the semantic view is unable to accommodate the great diversity of scientific models and faces special challenges from, for example, the use of inconsistency in many models. In response to such criticisms, a philosophical research program has emerged over the past two decades, which seeks to establish a *middle ground* between the classical semantic view of models discussed in the previous section and those who are sceptical about the prospects of formal approaches altogether. This research program is often called the *partial structures approach*, which was pioneered by Newton da Costa and Steven French and whose vocal proponents include Otávio Bueno, James Ladyman, and others; see [1.32] and references therein.

Like many adherents of the semantic view, partial structures theorists hold that models are to be reconstructed in set-theoretic terms, as ordered  $n$ -tuples of sets: a set of objects with (sets of) properties, quantities and relations, and functions defined over the quantities. A *partial structure* may then be defined as  $\mathfrak{A} = \langle D, R_i \rangle_{i \in I}$ , where  $D$  is a nonempty set of  $n$ -tuples of just this kind and each  $R_i$  is a  $n$ -ary relation. Unlike on the traditional semantic view, the relations  $R_i$  need not be complete isomorphisms, but crucially are *partial relations*: that is, they need not be defined for all  $n$ -tuples of elements of  $D$ . More specifically, for each partial relation  $R_i$ , in addition to the set of  $n$ -tuples for which the relation holds and the set of  $n$ -tuples for which it does not hold, there is also a third set of  $n$ -tuples for which it is underdetermined whether or not it holds. (There is a clear parallel here with Hesse’s notion of positive, negative, and neutral analogies which, as da Costa and French put it, “finds a natural home in the context of partial structures” [1.32, p. 48].) A total structure is said to *extend* a partial structure, if it subsumes the first two

sets without change (i. e., includes all those objects and definite relations that exist in the partial structures) and renders each extended relation well defined for every  $n$ -tuple of objects in its domain. This gives rise to a hierarchy of structures and substructures, which together with the notion of partial isomorphism loosens the requirements on representation, since all that is needed for two partial models  $A$  and  $A'$  to be *partially* isomorphic is that a partial substructure of  $A$  be isomorphic to a partial substructure in  $A'$ .

Proponents of the partial structures approach claim that it “widens the framework of the model-theoretic approach and allows various features of models and theories – such as analogies, iconic models, and so on – to be represented,” [1.33, p. 306] that it can successfully contain the difficulties arising from inconsistencies in models, and that it is able to capture “the existence of a hierarchy of models stretching from the data up to the level of theory” [1.33]. Some critics have voiced criticism about such sweeping claims. One frequent criticism concerns the proliferation of partial isomorphisms, many of which will trivially obtain; however,

if partial relations are so easy to come by, how can one tell the interesting from the vast majority of irrelevant ones? (*Pincock* speaks in this connection of the “danger of trivializing our representational relationships” [1.34, p. 1254].) *Suárez* and *Cartwright* add further urgency to this criticism, by noting that the focus on set-theoretical structures obliterates all those uses of models and aspects of scientific practice that do not amount to the making of claims [1.35, p. 72]:

“So all of scientific practice that does not consist in the making of claims gets left out. [...] Again, we maintain that this inevitably leaves out a great deal of the very scientific practice that we are interested in.”

It is perhaps an indication of the limitations of the partial structures approach that, in response to such criticism, its proponents need to again invoke heuristic factors, which cannot themselves be subsumed under the proposed formal framework of models as set-theoretic structures with partial relations.

## 1.5 The Folk Ontology of Models

If we accept that scientific models are best thought of as functional entities (Sect. 1.2), perhaps something can be learnt about the ontology of scientific models from looking at their functional role in scientific inquiry. What one finds across a range of different kinds of models is the practice of taking models as stand-ins for systems that are not, in fact, instantiated. As *Godfrey-Smith* puts it, “modelers often *take* themselves to be describing imaginary biological populations, imaginary neural networks, or imaginary economies” [1.36, p. 735] – that is, they are aware that due to idealization and abstraction, model systems will differ in their descriptions from a full account of the actual world. A model, thus understood, may be thought of as a “description of a missing system,” and the corresponding research practice of describing and characterizing model systems *as though* they were real instantiated systems (even though they are not) may be called, following *Thomson-Jones*, the “face-value practice” of scientific modeling [1.37, pp. 285–286].

On the heels of the face-value practice of scientific modeling, it has been argued, comes a common – though perhaps not universally shared – understanding of *what models are* [1.36, p. 735]:

“[...] to use a phrase suggested by *Deena Skolnick*, the treatment of model systems as comprising

imagined concrete things is the *folk ontology* of at least many scientific modelers. It is the ontology embodied in many scientists’ unreflective habits of talking about the objects of their study-talk about what a certain kind of population will do, about whether a certain kind of market will clear. [...] One kind of understanding of model-based science requires that we take this *folk ontology* seriously, as part of the scientific strategy.”

The ontology of *imagined concrete things* – that is, of entities that, *if real*, would be on a par with concrete objects in the actual world – leads quickly into the thorny territory of fictionalism. *Godfrey-Smith* is explicit about this when he likens models to “something we are all familiar with, the imagined objects of literary fiction” [1.36] – such as *Sherlock Holmes*, *J.R.R. Tolkien’s Middle Earth*, and so on. Implicit in this suggestion is, of course, a partial answer to our question *What is a model?* – namely, that the ontological status of scientific models is *just like* that of literary (or other) fictions. The advantages and disadvantages of such a position will be discussed in detail in Sect. 1.6 of this chapter.

There is, however, another direction into which a closer analysis of the face-value practice can take us. Instead of focusing on the ontological status of the en-

tities we are imagining when we contemplate models as imagined concrete things, we can focus on the conscious processes that attend such imaginings (or, if one prefers a different way of putting it, the *phenomenology* of interacting with models). Foremost among these is the mental imagery that is conjured up by the descriptions of models. (Indeed, as we shall see in the next section, on certain versions of the fictionalist view, a model *prescribes* imaginings about its target system.) How much significance one should attach to the mental pictures that attend our conscious consideration of models has been a matter of much controversy: recall Duhem’s dismissal of mechanical imagery as a way of conceptualizing electromagnetic phenomena (Sect. 1.3).

Focusing on the mental processes that accompany the use of scientific models might lead one to propose an analysis of models in terms of their cognitive foundations. Nancy Nersessian has developed just such an analysis, which ties the notion of models in science closely to the cognitive processes involved in mental modeling. Whereas the traditional approach in psychology had been to think of reasoning as consisting of the mental application of logical rules to propositional representations, mounting empirical evidence of the role of heuristics and biases suggested that much of human reasoning proceeds via *mental models* [1.38], that is, by carrying out thought experiments on internal models. A *mental model*, on this account, is “a structural analog of a real-world or imaginary situation, event, or process” as constructed by the mind in reasoning (and, presumably, realized by certain underlying brain processes) [1.39, pp. 11–12]:

“What it means for a mental model to be a structural analog is that it embodies a representation of the spatial and temporal relations among, and the causal structures connecting the events and entities depicted and whatever other information that is relevant to the problem-solving talks. [...] The essential points are that a mental model can be non-linguistic in form and the mental mechanisms are such that they can satisfy the model-building and simulative constraints necessary for the activity of mental modeling.”

While this characterization of mental models may have an air of circularity, in that it essentially defines mental models as place-holders for *whatever it takes* to support *the activity of mental modeling*, it nonetheless suggests a place to look for the materials from which models are constructed: the mind itself, with its various types of content and mental representation. As *Nersessian* puts it: “Whatever the format of the model

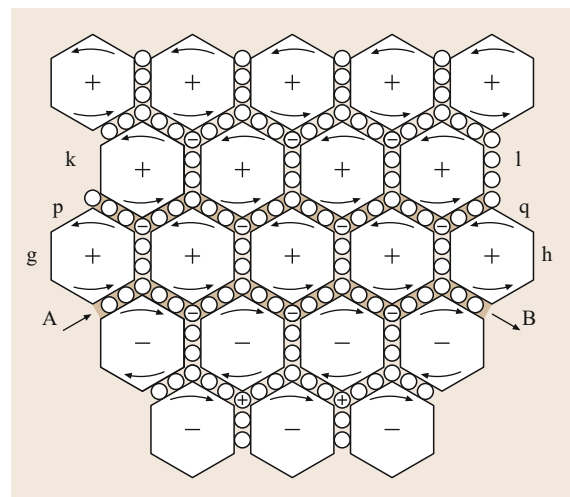
itself, information in various formats, including linguistic, formulaic, visual, auditory, kinesthetic, can be used in its construction” [1.39, p. 12].

How does this apply to the case of *scientific* models? As an example, Nersessian considers James Clerk Maxwell’s famous molecular vortex model, which visualized the lines of magnetic force around a magnet as though they were vortices within a continuous fluid (Fig. 1.1).

As Nersessian sees it, Maxwell’s drawing “is a *visual* representation of an *analogical* model that is accompanied with instructions for *animating* it correctly in thought” [1.39, p. 13]. And indeed *Maxwell* gives detailed instructions regarding how to interpret, and bring to life, the model of which the reader is only given a momentary *snapshot* [1.40, p. 477]:

“Let the current from left to right commence in *AB*. The row of vortices *gh* above *AB* will be set in motion in the opposite direction to a watch [...]. We shall suppose the two of vortices *kl* still at rest, then the layer of particles between these rows will be acted on by the row *gh*,”

and so forth. It does seem plausible to say that such instructions are intended to prescribe certain mental models on the part of the reader. Convincing though this example may be, it still begs the question of what, *in general*, a mental model is. At the same time, it illustrates what is involved in conjuring up a mental model and which materials – in this case, spatial representations, along with intuitions about the mechanical motion of parts in a larger system – are involved in its constitution.



**Fig. 1.1** Maxwell’s drawing of the molecular vortex model (after [1.40])

## 1.6 Models and Fiction

As noted in the previous section, the face-value practice of scientific modeling and its concomitant folk ontology, according to which models are imagined concrete things, have a natural affinity to the way we think about fictions. As one proponent of models as fictions puts it [1.41, p. 253]:

“The view of model systems that I advocate regards them as imagined physical systems, that is, as hypothetical entities that, as a matter of fact, do not exist spatiotemporally but are nevertheless not purely mathematical or structural in that they would be physical things if they were real.”

Plausible though this may sound, the devil is in the details. A first – perhaps trivial – caveat concerns the restriction that model systems *would be physical things if they were real*. In order to allow for the notion of model to be properly applied to the social and cognitive sciences, such as economics and psychology, it is best to drop this restriction to physical systems. (On this point, see [1.30, p. 528].) This leaves as the gist of the folk-ontological view the thought that model systems, *if they were real*, would be *just as we imagine them* (or, more carefully, *just as the model instructs us to imagine them*).

In order to sharpen our intuitions about fictions, let us introduce an example of a literary fiction, such as the following statement from Doyle’s *The Adventure of the Three Garridebs* (1924) [1.42]: “Holmes had lit his pipe, and he sat for some time with a curious smile upon his face.” There is, of course, no actual human being that this statement represents: no one is sitting smilingly at 221B Baker Street, filling up the room with smoke from their pipe. (Indeed, until the 1930s, the address itself had no real-world referent, as the highest number on Baker Street then was No. 85.) And yet there is a sense in which this passage does seem to represent Sherlock Holmes and, within the context of the story, tells us something informative about him. In particular, it seems to lend support to certain statements about Sherlock Holmes as opposed to others. If we say *Holmes is a pipe smoker*, we seem to be asserting something true about him, whereas if we say *Holmes is a nonsmoker*, we appear to be asserting something false. One goal of the ontology of fictions is to make sense of this puzzle.

Broadly speaking, there are two kinds of philosophical approaches – realist and antirealist – regarding fictions. On the realist approach, even though Sherlock Holmes is not an actual human being, we must grant that he *does* exist in some sense. Following

*Meinong* [1.43], we might, for example, distinguish between *being* and *existence* and consider Sherlock Holmes to be an object that has all the requisite properties we normally attribute to him, except for the property of existence. Or we might take fictions to have existence, but only as abstract entities, not as objects in space and time. By contrast, antirealists about fictions deny that they have independent being or existence and instead settle for other ways of making sense of how we interpret fictional discourse. Following Bertrand Russell, we might paraphrase the statement *Sherlock Holmes is a pipe smoker and resides at 221B Baker Street* without the use of a singular term (*Sherlock Holmes*), solely in terms of a suitably quantified existence claim: *There exists one and only one x such that x is a pipe smoker and x resides at 221B Baker Street*. However, while this might allow us to parse the meaning of further statements about Sherlock Holmes more effectively, it does not address the puzzle that certain claims (such as *He is a pipe smoker*) ring true, whereas others do not – since it renders each part of the explicated statement false. This might not seem like a major worry for the case of literary fictions, but it casts doubt on whether we can fruitfully think about scientific models in those terms, given the epistemic role of scientific models as contributors to scientific knowledge.

In recent years, an alternative approach to fictions has garnered the attention of philosophers of science, which takes Walton’s notion of “games of make-believe” as its starting point. Walton introduces this notion in the context of his philosophy of art, where he characterizes (artistic) representations as “things possessing the social function of serving as props in games of make-believe” [1.44, p. 69]. In games of make-believe, participants engage in behavior akin to children’s pretend play: when a child uses a banana as a telephone *to call grandpa*, this action does not amount to actually calling her grandfather (and perhaps not even *attempting* to call him); rather, it is a move within the context of play – where the usual standards of realism are suspended – whereby the child resolves to treat the situation *as if* it were one of speaking to her grandfather on the phone.

The banana is simply a prop in this game of make-believe. The use of the banana as a make-believe telephone may be inspired by some physical similarity between the two objects (e.g., their elongated shape, or the way that each can be conveniently held to one’s ear and mouth at the same time), but it is clear that props can go beyond material objects to include, for example, linguistic representations (as would be the case with

the literary figure of Sherlock Holmes). While the rules governing individual pretend play may be ad hoc, communal games of make-believe are structured by shared normative principles which *authorize* certain moves as legitimate, while excluding other moves as illegitimate. It is in virtue of such principles that fictional truths can be generated: for example, a toy model of a bridge at the scale of 1 : 1000 prescribes that, “if part of the model has a certain length, then, fictionally, the corresponding part of the bridge is a thousand times that length” [1.45, p. 38] – in other words, even though the model itself is only a meter long, it *represents* the bridge as a thousand meters long. Note that the scale model could be a model of a bridge that is yet to be built – in which case it would still be true that, fictionally, the bridge is a thousand meters long: props, via the rules that govern them, *create* fictional truths.

One issue of contention has been what kinds of metaphysical commitments such a view of models entails. Talk of *imagined concrete things* as the material from which models are built has been criticized for amounting to an indirect account of modeling, by which [1.46, pp. 308, fn. 14]

“prepared descriptions and equations of motion ask us to imagine an *imagined concrete system* which then bears some other form of representation relation to the system being modelled.”

A more thoroughgoing direct view of models as fictions is put forward by *Toon*, who considers the following sentence from *Wells’s The War of the Worlds*: “The dome of St. Paul’s was dark against the sunrise, and injured, I saw for the first time, by a huge gaping cavity on its western side” [1.47, p. 229]. As *Toon* argues [1.46, p. 307]:

“There is no pressure on us to postulate a fictional, damaged, St. Paul’s for this passage to represent; the passage simply represents the actual St. Paul’s. Similarly, on my account, our prepared description and equation of motion do not give rise to a fictional, idealised bouncing spring since they represent the actual bouncing spring.”

By treating models as prescribing imaginings about *the actual objects* (where these exist and are the model’s target system), we may resolve to imagine all sorts of

things that are, as a matter of fact, false; however, so the direct view holds, this is nonetheless preferable to the alternative option of positing *independently existing* fictional entities [1.45, p. 42]. Why might one be tempted to posit, as the indirect view does, that fictional objects fitting the model descriptions must exist? An important motivation has to do with the assertoric force of our model-based claims. As *Giere* puts it: “If we insist on regarding principles as genuine statements, we have to find something that they describe, something to which they refer” [1.48, p. 745]. In response, proponents of the direct view have disputed the need “to regard theoretical principles formulated in modeling as genuine statements”; instead, as *Toon* puts it, “they are prescriptions to imagine” [1.45, p. 44].

One potential criticism the models as fictions view needs to address is the worry that, by focusing on the user’s imaginings, what a model is becomes an entirely subjective matter. A similar worry may be raised with respect to the mental models view discussed in Sect. 1.5: if a model is merely a place-holder for whatever is needed to sustain the activity of mental modeling (or imagining) on the part of an agent, how can one be certain that the same kinds of models (or props) reliably give rise to the same kinds of mental modeling (or imaginings)? In this respect, at least, the models as fictions view appears to be in a stronger position. Recall that, unlike in individual pretend play (or unconstrained imagining), in games of make-believe certain imaginations are sanctioned by the prop itself and the – public, shared – rules of the game. As a result, “someone’s imaginings are governed by intersubjective rules, which guarantee that, as long as the rules are respected, everybody involved in the game has the same imaginings” [1.41, p. 264] – though it should be added, not necessarily the same *mental images*.

In his 1963 book, *Models and Metaphors*, *Black* expressed his hope that an “exercise of the imagination, with all its promise and its dangers” may help pave the way for an “understanding of scientific models and archetypes” as “a reputable part of scientific culture” [1.4, p. 243]. Even though *Black* was writing in general terms (and perhaps for rhetorical effect), his characterization would surely be considered apt by the proponents of the models as fictions view, who believe that models allow us to imagine their targets to be a certain way, and that, by engaging in such imaginings, we can gain new scientific insights.

## 1.7 Mixed Ontologies: Models as Mediators and Epistemic Artifacts

In Sect. 1.1, a distinction was drawn between *informational* views of models, which emphasize the objective, two-place relation between the model and what it represents, and *pragmatic* views, according to which a model depends at least in part on the user's beliefs or intentions, thereby rendering model-based representation a three-place relation between model, target, and user. Unsurprisingly, which side one comes down on in this debate will also have an effect on one's take on the ontology of scientific models. Hence, structuralist approaches (e.g., the partial structures approach discussed in Sect. 1.4.4) are a direct manifestation of the informational view, whereas the models as fictions approach – especially insofar as it considers models to be props for the user's imagination – would be a good example of the pragmatic view. The pragmatic dimension of scientific representation has received growing attention in the philosophical literature, and while this is not the place for a detailed survey of pragmatic accounts of model-based representation in particular, the remainder of this section will be devoted to a discussion of the ontological consequences of several alternative pragmatic accounts of models. Particular emphasis will be placed on what I shall call *mixed ontologies*, that is, accounts of models that emphasize the heterogeneity and diversity of their components.

### 1.7.1 Models as Mediators

Proponents of pragmatic accounts of models usually take scientific practice as the starting point of their analysis. This often directly informs how they think about models; in particular, it predisposes them to treat models as the outcome of a process of model construction. On this view, it is not only the *function* of models – for example, their capacity to represent target systems – which depends on the beliefs, intentions, and cognitive interests of a model user, but also the very *nature* of models which is dependent on human agents in this way. In other words, what models are is crucially determined by their being the result of a deliberate process of model construction. Model construction, most pragmatic theorists of models insist, is marked by “piecemeal borrowing” [1.35, p. 63] from a range of different domains. Such conjoining of heterogeneous components to form a model cannot easily be accommodated by structuralist accounts, or so it has been claimed; at the very least, there is considerable tension between, say, the way that the partial structures approach allows for a nested *hierarchy* of models (connected with one another via partial isomorphisms) and the much more ad hoc manner in which modelers piece

together models from a variety of ingredients. (On this point, see especially [1.35, p. 76].)

A number of such accounts have coalesced into what has come to be called the *models as mediators* view (see [1.49] for a collection of case studies). According to this view, models are to be regarded neither as a merely auxiliary intermediate step in applying or interpreting scientific theories, nor as constructed purely from data. Rather, they are thought of as mediating between our theories and the world in a partly autonomous manner. As *Morrison* and *Morgan* put it, models “are *not* situated in the middle of an hierarchical structure between theory and the world,” but operate outside the hierarchical “theory-world axis” [1.50, pp. 17–18]. A central tenet of the models as mediators view is the thesis that models “are made up from a *mixture* of elements, including those from outside the domain of investigation”; indeed, it is thought to be precisely in virtue of this heterogeneity that they are able to retain “an element of independence from both theory and data (or phenomena)” [1.50, p. 23].

At one level, the models as mediators view appears to be making a descriptive point about scientific practice. As *Morrison* and *Morgan* [1.50] point out, there is “no *logical* reason why models should be constructed to have these qualities of partial independence” [1.50, p. 17], though in practice they do exhibit them, and examples that involve the integration of heterogeneous elements beyond theory and data “are not the exception but the rule” [1.50, p. 15]. Yet, there is also the further claim that models could not fulfil their epistemic function *unless* they are partially autonomous entities: “we can only expect to use models to learn about our theories or our world if there is at least partial independence of the model from both” [1.50, p. 17]. Given that models are functional entities (in the sense discussed in Sect. 1.2), this has repercussions for the ontological question of what kind of entities models are. More often than not, models will integrate – perhaps imperfectly, but in irreducible ways – heterogeneous components from disparate sources, including (but not limited to) “elements of theories and empirical evidence, as well as stories and objects which could form the basis for modeling decisions” [1.50, p. 15]. As proponents of the models as mediators view are at pains to show, even in cases where models initially seem to derive straightforwardly from fundamental theory or empirical data, closer inspection reveals the presence of other elements – such as “simplifications and approximations which have to be decided independently of the theoretical requirements or of data conditions” [1.50, p. 16].

For the models as mediators approach, any answer to the question *what is a model?* must be tailored to the specific case at hand: models in high-energy physics will have a very different composition, and will consist of an admixture of different elements, than, say, models in psychology. However, as a general rule, no model – or, at any rate, no *interesting* model – will ever be fully reducible to theory and data; attempts to *clean up* the ontology of scientific models so as to render them either purely theoretical or entirely empirical, according to the models as mediators view, misconstrue the very nature and function of models in science.

### 1.7.2 Models as Epistemic Artifacts

A number of recent pragmatic approaches take the models as mediators view as their starting point, but suggest that it should be extended in various ways. Thus, *Knuuttila* acknowledges the importance of mediation between theory and data, but a richer account of models is needed to account for how this partial independence comes about. For *Knuuttila*, *materiality* is the key enabling factor that imbues models with such autonomy: it is “the material dimension, and not just *additional elements*, that makes models able to mediate” [1.51, p. 48]. Materiality is also seen as explaining the various epistemic functions that models have in inquiry, not least by way of analogy with scientific experiments. For example, just as in experimentation much effort is devoted to minimizing unwanted external factors (such as noise), in scientific models certain methods of approximation and idealization serve the purpose of neutralizing undesirable influences. Models typically draw on variety of formats and representations, in a way that *enables* certain specific uses, but at the same time *constrains* them; this breaks with the traditional assumption that we can “clearly tell apart those features of our scientific representations that are attributable to the phenomena described from the conventions used to describe them” [1.52, p. 268].

On the account sketched thus far, attempting to characterize the nature and function of models in the

language of theories and data would, in the vast majority of cases, give a misleading impression; instead, models are seen as *epistemic tools* [1.52, p. 267]:

“Concrete artifacts, which are built by various representational means, and are constrained by their design in such a way that they enable the study of certain scientific questions and learning through constructing and manipulating them.”

This links the philosophical debate about models to questions in the philosophy of technology, for example concerning the ontology of artifacts, which are likewise construed as both material bodies and functional objects. It also highlights the constitutive role of design and construction, which applies equally to models with a salient material dimension – such as scale models in engineering or ball-and-stick models in chemistry – and to largely theoretical models. For example, it has been argued that mathematical models (e.g., in many-body physics) may be fruitfully characterized not only in theoretical terms (say, as a Hamiltonian) or as mathematical entities (as an operator equation), but also as the output of a *mature mathematical formalism* (in this case, the formalism of second quantization) – that is, a physically interpreted set of notational rules that, while embodying various theoretical assumptions, is not usually reducible to fundamental theory [1.53].

As in the case of the models as mediators approach, the ontological picture that emerges from the artifactual approach to models is decidedly mixed: models will typically consist of a combination of different materials, media and formats, and deploy different representational means (such as pictorial, symbolic, and diagrammatic notations) as well as empirical data and theoretical assumptions. Beyond merely acknowledging the heterogeneity of such a *mixture of elements*, however, the artifactual approach insists that it is *in virtue of their material dimension* that the various elements of a model, taken together, enable and constrain its representational and other epistemic functions.

## 1.8 Summary

As the survey in this chapter demonstrates, the term *model* in science refers to a great variety of things: physical objects such as scale models in engineering, descriptions and sets of sentences, set-theoretic structures, fictional objects, or an assortment of all of the above. This makes it difficult to arrive at a uniform characterization of models *in general*. However, by paying

close attention to philosophical accounts of model-based representation, it is possible to discern certain clusters of positions. At a general level, it is useful to think of models as functional entities, as this allows one to explore how different functional perspectives lead to different conceptions of the ontology of models. Hence, with respect to the representational function of mod-



els, it is possible to distinguish between *informational* views, which we found to be closely associated with structuralist accounts of models, and *pragmatic* views, which tend to give rise to more heterogeneous accounts, according to which models may be thought of as *props for the imagination*, as partly autonomous mediators between theory and data, or as epistemic artifacts consisting of an admixture of heterogeneous elements.

When nineteenth century physicists began to reflect systematically on the role of *analogy* in science,

they did so out of a realization that it may not always be possible to apply fundamental theory directly to reality, either because any attempt to do so faces insurmountable complexities, or because no such fundamental theory is as yet available. At the beginning of the twenty-first century, these challenges have not diminished, and scientists find themselves turning to an ever greater diversity of scientific models, a unified philosophical theory of which is still outstanding.

## References

- 1.1 R. Frigg: Models in science. In: *Stanford Encyclopedia of Philosophy*, ed. by E.N. Zalta <http://plato.stanford.edu/entries/models-science/> (Spring 2012 Edition)
- 1.2 N. Goodman: *Languages of Art* (Bobbs-Merrill, Indianapolis 1968)
- 1.3 R. Ankeny, S. Leonelli: What's so special about model organisms?, *Stud. Hist. Philos. Sci.* **42**(2), 313–323 (2011)
- 1.4 M. Black: *Models and Metaphors: Studies in Language and Philosophy* (Cornell Univ. Press, Ithaca 1962)
- 1.5 P. Achinstein: *Concepts of Science: A Philosophical Analysis* (Johns Hopkins, Baltimore 1968)
- 1.6 J. von Neumann: Method in the physical sciences. In: *Collected Works Vol. VI. Theory of Games, Astrophysics, Hydrodynamics and Meteorology*, ed. by A.H. Taub (Pergamon, Oxford 1961) pp. 491–498
- 1.7 S. French: Keeping quiet on the ontology of models, *Synthese* **172**(2), 231–249 (2010)
- 1.8 G. Contessa: Editorial introduction to special issue, *Synthese* **2010**(2), 193–195 (2010)
- 1.9 S. Ducheyne: Towards an ontology of scientific models, *Metaphysica* **9**(1), 119–127 (2008)
- 1.10 R. Giere: Using models to represent reality. In: *Model-Based Reasoning in Scientific Discovery*, ed. by L. Magnani, N. Nersessian, P. Thagard (Plenum, New York 1999) pp. 41–57
- 1.11 A. Chakravartty: Informational versus functional theories of scientific representation, *Synthese* **217**(2), 197–213 (2010)
- 1.12 R.I.G. Hughes: Models and representation, *Proc. Philos. Sci.*, Vol. 64 (1997) pp. 5325–226
- 1.13 T. Knuuttila: Some consequences of the pragmatist approach to representation. In: *EPSA Epistemology and Methodology of Science*, ed. by M. Suárez, M. Dorato, M. Rédei (Springer, Dordrecht 2010) pp. 139–148
- 1.14 M. Suárez: An inferential conception of scientific representation, *Proc. Philosophy of Science*, Vol. 71 (2004) pp. 67–779
- 1.15 M. Jammer: Die Entwicklung des Modellbegriffs in den physikalischen Wissenschaften, *Stud. Gen.* **18**(3), 166–173 (1965)
- 1.16 P. Duhem: *The Aim and Structure of Physical Theory* (Princeton Univ. Press, Princeton 1954), Transl. by P.P. Wiener
- 1.17 D. Bailer-Jones: Models, metaphors and analogies. In: *The Blackwell Guide to the Philosophy of Science*, ed. by P. Machamer, M. Silberstein (Blackwell, Oxford 2002) pp. 108–127
- 1.18 D.H. Mellor: Models and analogies in science: Duhem versus Campbell?, *Isis* **59**(3), 282–290 (1968)
- 1.19 D. Bailer-Jones: *Scientific Models in Philosophy of Science* (Univ. Pittsburgh Press, Pittsburgh 2009)
- 1.20 M. Hesse: *Models and Analogies in Science* (Sheed Ward, London 1963)
- 1.21 N.R. Campbell: *Physics: The Elements* (Cambridge Univ. Press, Cambridge 1920)
- 1.22 J.M. Soskice, R. Harré: Metaphor in science. In: *From a Metaphorical Point of View: A Multidisciplinary Approach to the Cognitive Content of Metaphor*, ed. by Z. Radman (de Gruyter, Berlin 1995) pp. 289–308
- 1.23 C. Liu: Models and theories I: The semantic view revisited, *Int. Stud. Philos. Sci.* **11**(2), 147–164 (1997)
- 1.24 R. Carnap: *Foundations of Logic and Mathematics* (Univ. Chicago Press, Chicago 1939)
- 1.25 R. Hendry, S. Psillos: How to do things with theories: An interactive view of language and models in science. In: *The Courage of Doing Philosophy: Essays Presented to Leszek Nowak*, ed. by J. Brzeziński, A. Klawiter, T.A.F. Kuipers, K. Lastowski, K. Paprzycka, P. Przybyz (Rodopi, Amsterdam 2007) pp. 123–158
- 1.26 N. Cartwright: Models and the limits of theory: Quantum hamiltonians and the BCS model of superconductivity. In: *Models as Mediators*, ed. by M. Morrison, M. Morgan (Cambridge Univ. Press, Cambridge 1999) pp. 241–281
- 1.27 F. Suppe: *The Semantic Conception of Theories and Scientific Realism* (Univ. Illinois Press, Urbana 1989)
- 1.28 P. Suppes: A comparison of the meaning and uses of models in mathematics and the empirical sciences, *Synthese* **12**(2/3), 287–301 (1960)
- 1.29 B. van Fraassen: *The Scientific Image* (Oxford Univ. Press, Oxford 1980)
- 1.30 M. Thomson-Jones: Models and the semantic view, *Philos. Sci.* **73**(4), 524–535 (2006)
- 1.31 P. Suppes: *Introduction to Logic* (Van Nostrand, Princeton 1957)

- 1.32 N. da Costa, S. French: *Science and Partial Truth: A Unitary Approach to Models and Scientific Reasoning* (Oxford Univ. Press, New York 2003)
- 1.33 S. French: The structure of theories. In: *The Routledge Companion to Philosophy of Science*, 2nd edn., ed. by M. Curd, S. Psillos (Routledge, London 2013) pp. 301–312
- 1.34 C. Pincock: Overextending partial structures: Idealization and abstraction, *Philos. Sci.* **72**(4), 1248–1259 (2005)
- 1.35 M. Suárez, N. Cartwright: Theories: Tools versus models, *Stud. Hist. Philos. Mod. Phys.* **39**(1), 62–81 (2008)
- 1.36 P. Godfrey-Smith: The strategy of model-based science, *Biol. Philos.* **21**(5), 725–740 (2006)
- 1.37 M. Thomson-Jones: Missing systems and the face value practice, *Synthese* **172**(2), 283–299 (2010)
- 1.38 P.N. Johnson-Laird: *Mental Models* (Harvard Univ. Press, Cambridge 1983)
- 1.39 N. Nersessian: Model-based reasoning in conceptual change. In: *Model-Based Reasoning in Scientific Discovery*, ed. by L. Magnani, N. Nersessian, P. Thagard (Plenum, New York 1999) pp. 5–22
- 1.40 J.C. Maxwell: *The Scientific Papers of James Clerk Maxwell*, Vol. 1 (Cambridge Univ. Press, Cambridge 1890), ed. by W.D. Niven
- 1.41 R. Frigg: Models and fiction, *Synthese* **172**(2), 251–268 (2010)
- 1.42 A.C. Doyle: The Adventure of the Three Garridebs. In: *The Casebook of Sherlock Holmes*, ed. by J. Miller (Dover, 2005)
- 1.43 A. Meinong: *Untersuchungen zur Gegenstandstheorie und Psychologie* (Barth, Leipzig 1904)
- 1.44 K. Walton: *Mimesis as Make-Believe: On the Foundations of the Representational Arts* (Harvard Univ. Press, Cambridge 1990)
- 1.45 A. Toon: *Models as Make-Believe: Imagination, Fiction, and Scientific Representation* (Palgrave-Macmillan, Basingstoke 2012)
- 1.46 A. Toon: The ontology of theoretical modelling: Models as make-believe, *Synthese* **172**(2), 301–315 (2010)
- 1.47 H.G. Wells: *War of the Worlds* (Penguin, London 1897), 1978
- 1.48 R. Giere: How models are used to represent reality, *Proc. Philosophy of Science*, Vol. 71 (2004) pp. S742–S752
- 1.49 M. Morrison, M. Morgan (Eds.): *Models as Mediators* (Cambridge Univ. Press, Cambridge 1999)
- 1.50 M. Morrison, M. Morgan: Models as mediating instruments. In: *Models as Mediators*, ed. by M. Morrison, M. Morgan (Cambridge Univ. Press, Cambridge 1999) pp. 10–37
- 1.51 T. Knuuttila: *Models as Epistemic Artefacts: Toward a Non-Representationalist Account of Scientific Representation* (Univ. Helsinki, Helsinki 2005)
- 1.52 T. Knuuttila: Modelling and representing: An artefactual approach to model-based representation, *Stud. Hist. Philos. Sci.* **42**(2), 262–271 (2011)
- 1.53 A. Gelfert: *How to Do Science with Models: A Philosophical Primer* (Springer, Cham 2016)

# Models and Theories

## 2. Models and Theories

Demetris Portides

Part A | 2

Both the received view (RV) and the semantic view (SV) of scientific theories are explained. The arguments against the RV are outlined in an effort to highlight how focusing on the syntactic character of theories led to the difficulty in characterizing theoretical terms, and thus to the difficulty in explicating how theories relate to experiment. The absence of the representational function of models in the picture drawn by the RV becomes evident; and one does not fail to see that the SV is in part a reaction to – what its adherents consider to be an – excessive focus on syntax by its predecessor and in part a reaction to the complete absence of models from its predecessor's philosophical attempt to explain the theory–experiment relation. The SV is explained in an effort to clarify its main features but also to elucidate the differences between its different versions. Finally, two kinds of criticism are explained that affect all versions of the SV but which do not affect the view that models have a warranted degree of importance in scientific theorizing.

2.1	<b>The Received View of Scientific Theories</b> .....	26
2.1.1	The Observation–Theory Distinction .....	27
2.1.2	The Analytic–Synthetic Distinction .....	29
2.1.3	Correspondence Rules .....	30
2.1.4	The Cosmetic Role of Models According to the RV .....	32
2.1.5	Hempel's Provisos Argument .....	33
2.1.6	Theory Consistency and Meaning Invariance .....	34
2.1.7	General Remark on the Received View ....	35
2.2	<b>The Semantic View of Scientific Theories</b> .....	36
2.2.1	On the Notion of Model in the SV .....	38
2.2.2	The Difference Between Various Versions of the SV .....	40
2.2.3	Scientific Representation Does not Reduce to a Mapping of Structures .....	42
2.2.4	A Unitary Account of Models Does not Illuminate Scientific Modeling Practices .....	44
2.2.5	General Remark on the Semantic View ...	46
	<b>References</b> .....	47

Scientists use the term *model* with reference to iconic or scaled representations, analogies, and mathematical (or abstract) descriptions. Although all kinds of models in science may be philosophically interesting, mathematical models stand out. Representation with iconic or scale models, for instance, mostly applies to a particular state of a system at a particular time, or it requires the mediation of a mathematical (or abstract) model in order to relate to theories. Representation via mathematical models, on the other hand, is of utmost interest because it applies to *types* of target systems and it can be used to draw inferences about the time-evolution of such systems, but more importantly for our purposes because of its obvious link to scientific theories.

In the history of philosophy of science, there have been two systematic attempts to explicate the relation of such models to theory. The first is what had been labeled the *received view* (RV) of scientific theories

that grew out of the logical positivist tradition. According to this view, theories are construed as formal axiomatic calculi whose logical consequences extend to observational sentences. Models are thought to have no representational role; their role is understood metamathematically, as interpretative structures of subsets of sentences of the formal calculus. Ultimately it became clear that such a role ascribed to models does not do justice to how science achieves theoretical representations of phenomena. This conclusion was reached largely due to the advent of the second systematic attempt to explore the relation between theory and models, the *semantic view* (SV) or model-theoretic view of scientific theories. The semantic view regards theories as classes of models that are directly defined without resort to a formal calculus. Thus, models in this view are integral parts of theories, but they are also the devices by which representation of phenomena is achieved.

Although, the SV recognized the representational capacity of models and exposed that which was concealed by the logical positivist tradition, namely that one of the primary functions of scientific models is to apply the abstract theoretical principles in ways that actual physical systems can be represented, it also generated a debate concerning the complexities involved in scientific representation. This recent debate has significantly enhanced our understanding of the representational role of scientific models. At the same time it gave rise, among other things, to questions regarding the relation between models and theory. The adherents of the SV claim that a scientific theory is identified with a class of models, hence that models are constitutive parts of theory and thus they represent by means of the conceptual apparatus of theory. The critics of the SV, however, argue that those models that are successful representations of physical systems utilize a much richer conceptual apparatus than that provided by theory and thus claim that they should be understood as partially autonomous from theory.

A distinguishing characteristic of this debate is the notion of representational model, that is, a scientific entity which possesses the necessary features that render it representational of a physical system. In the SV, theoretical models, that is, mathematical models that are constitutive parts of theory structure, are considered to be representational of physical systems. Its critics, however, argue that in order to provide a model with the capacity to represent actual physical systems, the theoretical principles from which the model arises

are typically supplemented with ingredients that derive from background knowledge, from semiempirical results and from experiment. In order to better understand the character of successful representational models, according to this latter view, we must move away from a purely theory-driven view of model construction and also place our emphasis on the idea that representational models are entities that consist of assortments of the aforementioned sorts of conceptual ingredients.

In order to attain insight into how models could relate to theory and also be able to use that insight in addressing other issues regarding models, in what follows, I focus on the RV and the SV of scientific theories. Each of the two led to a different conception of the nature of theory structure and subsequently to a different suggestion for what scientific models are, what they are used for, and how they function. In the process of explicating these two conceptions of theory structure, I will also review the main arguments that have been proposed against them. The RV has long been abandoned for reasons that I shall explore in Sect. 2.1, but the SV lives on despite certain inadequacies that I shall also explore in Sect. 2.2. Toward the end of Sect. 2.2, in Sect. 2.2.4, I shall very briefly touch upon the more recent view that the relation between theory and models is far more complex than advocates of the RV or the SV have claimed, and that models in science demonstrate a certain degree of partial autonomy from the theory that prompts their construction and because of this a unitary account of models obscures significant features of scientific modeling practices.

## 2.1 The Received View of Scientific Theories

What has come to be called the RV of scientific theories is a conception of the structure of scientific theories that is associated with logical positivism, and which was the predominant view for a large part of the twentieth century. It is nowadays by and large overlooked hence it is anything but received. Despite its inappropriate label, clarifying its major features as well as understanding the major philosophical arguments that revealed its inadequacies would not only facilitate acquaintance with the historical background of the debate about the structure of scientific theories and give the reader a flavor of the difficulties involved in characterizing theory structure, but it would also be helpful in understanding some characteristics of contemporary views and how models came to occupy central stage in current debates on how theories represent and explain phenomena. With this intention in mind, I proceed in this section by briefly explaining the major features of the RV and continue with sketching the

arguments that exposed its weaknesses in Sects. 2.1.1–2.1.6.

The RV construes scientific theories as Hilbert-style formal axiomatic calculi, that is, axiomatized sets of sentences in first-order predicate calculus with identity. A scientific theory is thus identified with a formal language,  $L$ , having the following features. The nonlogical terms of  $L$  are divided into two disjoint classes: (1) the theoretical terms that constitute the theoretical vocabulary,  $V_T$ , of  $L$  and (2) the observation terms that constitute the observation vocabulary,  $V_O$ , of  $L$ . Thus,  $L$  can be thought of as consisting of an observation language,  $L_O$ , that is, a language that consists only of observation terms, a theoretical language,  $L_T$ , that is, a language that consists only of theoretical terms, and a part that consists of mixed sentences that are made up of both observation and theoretical terms. The theoretical postulates or the axioms of the theory,  $T$  (i. e., what we, commonly, refer to as the high-level scientific

laws), consist only of terms from  $V_T$ . This construal of theories is a syntactic system, which naturally requires semantics in order to be useful as a model of scientific theories.

It is further assumed that the terms of  $V_O$  refer to directly observable physical objects and directly observable properties and relations of physical objects. Thus the semantic interpretation of such terms, and the sentences belonging to  $L_O$ , is provided by direct observation. The terms of  $V_T$ , and subsequently all the other sentences of  $L$  not belonging to  $L_O$ , are partially interpreted via the theoretical postulates,  $T$ , and – a finite set of postulates that has come to be known as – the *correspondence rules*,  $C$ . The latter are mixed sentences of  $L$ , that is, they are constructed with at least one term from each of the two classes  $V_T$  and  $V_O$ . (The reader could consult *Suppe* [2.1] for a detailed exposition of the RV, but also for a detailed philosophical study of the developments that the RV underwent under the weight of several criticisms until it reached, what Suppe calls, the “final version of the RV”).

We could synopsise how scientific theories are conceived according to the RV as follows: The scientific laws, which as noted constitute the axioms of the theory, specify relations holding between the theoretical terms. Via a set of *correspondence rules*, theoretical terms are reduced to, or defined by, observation terms. Observation terms refer to objects and relations of the physical world and thus are interpreted. Hence, a scientific theory, according to the RV, is a formal axiomatic system having as point of departure a set of theoretical postulates, which when augmented with a set of correspondence rules has deductive consequences that stretch all the way to terms, and sentences consisting of such terms, that refer to the directly observable physical objects. Since according to this view, the backbone of a scientific theory is the set of theoretical postulates,  $T$ , and a partial interpretation of  $L$  is given via the set of correspondence rules,  $C$ , let  $TC$  (i. e., the union set of  $T$  and  $C$ ) designate the scientific theory.

From this sketch, it can be inferred that the RV implies several philosophically interesting things. For the purposes of this chapter, it suffices to limit the discussion only to those implications of the RV that are relevant to the criticisms that have contributed to its downfall. These implications, which in one way or another relate to the difficulty in characterizing  $V_T$  terms, are:

1. It relies on an observational–theoretical distinction of the terms of  $L$ .
2. It embodies an analytic–synthetic distinction of the sentences of  $L$ .

3. It employs the obscure notion of correspondence rules to account for the interpretation of theoretical terms and to account for theory application.
4. It does not assign a representational function to models.
5. It assigns a deductive status to the relation between empirical theories and experiment.
6. It commits to a theory consistency condition and to a meaning invariance condition.

### 2.1.1 The Observation–Theory Distinction

The separation of  $L$  into  $V_O$  and  $V_T$  terms implies that the RV requires an observational–theoretical distinction in the terms of the vocabulary of the theory. This idea was criticized in two ways. The first kind of objection to the observation–theory distinction relied on a twofold argument. On the one hand, the critics claim that an observation–theory distinction of scientific terms cannot be drawn; and on the other, that a classification of terms following such a distinction would give rise to a distinction of observational–theoretical statements, which also cannot be drawn for scientific languages. The second kind of objection to the distinction relies on attempts to establish accounts of *observation* that are incompatible with the observation–theory distinction and on showing that observation statements are theory laden.

#### The Untenability of the Observation–Theory Distinction

The argument of the first kind that focuses on the untenability of the observation–theory distinction is due to *Achinstein* [2.2, 3] and *Putnam* [2.4]. Achinstein explores the sense of observation relevant to science, that is, “the sense in which observing involves visually attending to something,” and he claims that this sense exhibits the following characteristics:

1. Observation involves attention to the various aspects or features of an item depending on the observer’s concerns and knowledge.
2. Observation does not necessarily involve recognition of the item.
3. Observation does not imply that whatever is observed is in the visual field or in the line of sight of the observer.
4. Observation could be achieved indirectly.
5. The description of what one observes can be done in different ways (The reader could refer to *Achinstein* [2.3, pp. 160–165] for an explication of these characteristics of observation by the use of specific examples).

If now one urges an observation–theory distinction by simply constructing lists of observable and unobservable terms (as proponents of the RV according to Achinstein do), the distinction becomes untenable. For example, according to typical lists of unobservables, *electron* is a theoretical term. But based on points (3) and (4) above, Achinstein claims, this could be rejected. Similarly based on point (5), Achinstein also rejects the tenability of such a distinction at the level of statements, because “what scientists as well as others observe is describable in many different ways, using terms from both vocabularies” [2.3, p. 165].

Furthermore, if, as proponents of the RV have often claimed, (For instance, *Hempel* [2.5], *Carnap* [2.6] and [2.7]), items in the observational list are directly observable whereas those in the theoretical list are not, then *Achinstein* [2.3, pp. 172–177] claims that a close construal of *directly observable* reveals that the desired classification of terms into the two lists fails. He explains that *directly observable* could mean that it can be observed without the use of instruments. If this is what advocates of the RV require, then it does not warrant the distinction. First, it is not precise enough to classify things seen by images and reflections. Second, if something is not observable without instruments means that no aspect of it is observable without instruments then things like temperature and mass would be observables, since some aspects of them are detected without instruments. If however directly observable means that no instruments are required to detect its presence, then it would be insufficient because one cannot talk about the presence of temperature. Finally, if it means that no instruments are required to measure it or its properties, then such terms as volume, weight, etc. would have to be classified as theoretical terms. Hence, Achinstein concludes that the notion of direct observability is unclear and thus fails to draw the desired observation–theory distinction.

Along similar lines, *Putnam* [2.4] argues that the distinction is completely *broken-backed* mainly for three reasons. First, if an observation term is one that only refers to observables then there are no observation terms. For example, the term *red* is in the observable class but it was used by Newton to refer to a theoretical term, namely red corpuscles. Second, many terms that refer primarily to the class of unobservables are not theoretical terms. Third, some theoretical terms, that are of course the outcome of a scientific theory, refer primarily to observables. For example, the theory of evolution, as put forward by Darwin, referred to observables by employing theoretical terms.

What these arguments accomplish is to highlight the fact that scientific languages employ terms that cannot clearly and easily be classified into observational or

theoretical. They do not however show the untenability of the observation–theory distinction as employed by the RV. As *Suppe* [2.8] correctly observes, what they show is that the RV needs a sufficiently rich artificial language for science, no matter how complex it may turn out to be. Such a language, in which presumably the observation–theory distinction is tenable, must have a plethora of terms, such that, to use his example, the designated term *red<sub>o</sub>* will refer to the observable occurrences of the predicate *red*, and the designated term *red<sub>t</sub>* will refer to the unobservable occurrences.

### The Theory–Ladenness of Observation

Hanson’s argument is a good example of the second kind, in which an attempt is made to show that there is no theory-neutral observation language and that observation is theory-laden and thus establish an account of *observation* that is incompatible with the observation–theory distinction required by the RV (*Hanson* [2.9, pp. 4–30], *Hanson* [2.10, pp. 59–198]. Also *Suppe* [2.1, pp. 151–166]). He does this by attempting to establish that an observation language that intersubjectively can be given a theory-independent semantic interpretation, as the RV purports, cannot exist.

He begins by asking whether two people see the same things when holding different theories. We could follow his argument by reference to asking whether Kepler and Tycho Brahe see the same thing when looking at the sun rising. Kepler, of course, holds that the earth revolves around the sun, while Tycho holds that the sun revolves around the earth. Hanson addresses this question by considering ambiguous figures, that is, figures that sometimes can be seen as one thing and other times as another. The most familiar example of this kind is the duck–rabbit figure.

When confronted with such figures, viewers see either a duck or a rabbit depending on the perspective they take, but in both cases they see the same distal object (i. e., the object that emits the rays of light that impinge the retina). Hanson uses this fact to develop a sequence of arguments to counter the standard interpretations of his time. There were two standard interpretations at the time. The first was that the perceptual system delivers the same visual representation and then cognition (thought) interprets this either as a duck or as a rabbit. The other was that the perceptual system outputs both representations and then cognition chooses one of the two. Both interpretations are strongly linked with the idea that the perceptual process and the cognitive process function independently of one another, that is, the perceptual system delivers its output independent of any cognitive influences. However, Hanson challenges the assumption that the two observers see the same thing and via thought they in-

interpret it differently. He claims that perception does not deliver either a duck or a rabbit, or an ambiguous figure, and then via some other independent process thought chooses one or the other. On the contrary, the switch from seeing one thing to seeing the other seems to take place spontaneously and moreover a process of back and forth seeing without any thinking seems to be involved. He goes on to ask, what could account for the difference in what is seen? His answer is that what changes is the organization of the ambiguous figure as a result of the conceptual background of each viewer. This entails that what one sees, the percept, depends on the conceptual background that results from one's experience and knowledge, which means that thought affects the formation of the percept; thus perception and cognition become intertwined. When Tycho and Kepler look at the sun, they are confronted with the same distal object but they see different things because their conceptual organizations of their experiences are vastly different. In other words, Hanson's view is that the percept depends on background knowledge, which means that cognition influences perceptual processing. Consequently, observation is theory laden, namely, observation is conditional on background knowledge.

By this argument, Hanson undermines the RV's position, which entails that Kepler and Brahe see the same thing but interpret it differently; and also establishes that conceptual organizations are features of *seeing* that are indispensable to scientific observation and thus that Kepler and Brahe see two different things because perception inherently involves interpretation, since the former is conditional on background knowledge. It is, however, questionable whether Hanson's arguments are conclusive. Fodor [2.11–13], Pylyshyn [2.14, 15], and Raftopoulos [2.16–18], for example, have extensively argued on empirical grounds that perception, or at least a part of it, is theory independent and have proposed explanations of the ambiguous figures that do not invoke cognitive effects in explaining the percept and the switch between the two interpretations of the figure. This debate, therefore, has not yet reached its conclusion; and many today would argue that fifty or so years after Hanson the arguments against the theory ladenness of observation are much more tenable.

### 2.1.2 The Analytic–Synthetic Distinction

The RV's dependence on the observation–theory distinction is intimately linked to the requirement for an analytic–synthetic distinction. An argument to defend this claim is given by Suppe [2.1, pp. 68–80]. Here is a sketch of that argument. The analytic–synthetic distinction is embodied in the RV, because (as suggested by Carnap [2.19]) implicit in *TC* are meaning postu-

lates (or semantical rules) that specify the meanings of sentences in *L*. However, if meaning specification were the only function of *TC* then *TC* would be analytic, and in such case it would not be subject to empirical investigation. *TC* must therefore have a factual component, and the meaning postulates must separate the meaning from the factual component. This would imply an analytic–synthetic separation, if those sentences in *L* that are logical truths or logical consequences of the meaning postulates are analytic and all nonanalytic sentences are understood to be synthetic. Moreover, any nonanalytic sentence in *L* taken in conjunction with the class of meaning postulates would have certain empirical consequences. If the conjunction is refuted or confirmed by directly observable evidence, this will reflect only on the truth value of the conjunction and not on the meaning postulates. Hence such conjunctive sentences can only be synthetic. Thus every nonanalytic sentence of *L<sub>0</sub>* and every sentence of *L* constituted by a mixed vocabulary is synthetic. So the observation–theory distinction supports an analytic–synthetic distinction of the sentences of *L*.

The main criticism against the analytic–synthetic distinction consists of attempts to show its untenability. Quine [2.20] points out that there are two kinds of analytic statements: (a) logical truths, which remain true under all interpretations, and (b) statements that are true by virtue of the meaning of their nonlogical terms, for example, *No bachelor is married*. He then argues that the analyticity of statements of the second kind cannot be established without resort to the notion of synonymy, and that the latter notion is just as problematic as the notion of analyticity. The argument runs roughly as follows. Given that meaning (or intension) is clearly distinguished from its extension, that is, the class of entities to which it refers, a theory of meaning is primarily concerned with cognitive synonymy (i. e., the synonymy of linguistic forms). For example, to say that *bachelor* and *unmarried man* are cognitively synonymous is to say that they are interchangeable in all contexts without change of truth value. If such were the case then the statement *No bachelor is married* would become *No unmarried man is married*, which would be a logical truth. In other words, statements of kind (b) are reduced to statements of kind (a) if only we could interchange synonyms for synonyms. But as Quine argues, the notion of interchangeability *salva veritate* is an extensional concept and hence does not help with analyticity. In fact, no analysis of the interchangeability *salva veritate* account of synonymy is possible without recourse to analyticity, thus making such an effort circular, unless interchangeability is “[...] relativized to a language whose extent is specified in relevant respects” [2.20, p. 30]. That is to say,

we first need to know what statements are analytic in order to decide which expressions are synonymous; hence appeal to synonymy does not help with the notion of analyticity.

Similarly White [2.21] argues that an artificial language,  $L_1$ , can be constructed with appropriate definitional rules, in which the predicates  $P_1$  and  $Q_1$  are synonymous whereas  $P_1$  and  $Q_2$  are not; hence making such sentences as  $\forall x (P_1(x) \rightarrow Q_1(x))$  logical truths and such sentences as  $\forall x (P_1(x) \rightarrow Q_2(x))$  synthetic. In a different artificial language  $L_2$ ,  $P_1$  could be defined to be synonymous to  $Q_2$  and not to  $Q_1$ , hence making the sentence  $\forall x (P_1(x) \rightarrow Q_2(x))$  a logical truth and the sentence  $\forall x (P_1(x) \rightarrow Q_1(x))$  synthetic. This relies merely upon convention. However, he asks, in a natural language what rules are there that dictate what choice of synonymy can be made such that one formula is a synthetic truth rather than analytic? The key point of the argument is therefore that in a natural language or in a scientific language, which are not artificially constructed and which do not contain definitional rules, the notion of analyticity is unclear.

Nevertheless, it could be argued that such arguments as the above are not entirely conclusive, primarily because the RV is not intended as a description of actual scientific theories. Rather, the RV is offered as a *rational reconstruction* of scientific theories, that is, an explication of the structure of scientific theories. It does not aim to describe how actual theories are formulated, but only to indicate a logical framework (i. e., a canonical linguistic formulation) in which theories can be essentially reformulated. Therefore, all that proponents of the RV, needed to show was that the analytic–synthetic distinction is tenable in some artificial language (with meaning postulates) in which scientific theories could potentially be reformulated. In view of this, in order for the RV to overcome the obscurity of the notion of analyticity, pointed out by Quine and White, it would require the conclusion of a project that Carnap begun: To spell out a clear way by which to characterize meaning postulates for a specified theoretical language (This is clearly Carnap’s intention in his [2.19]).

### 2.1.3 Correspondence Rules

In order to distinguish the character and function of theoretical terms from speculative metaphysical ones (e.g., *unicorn*), logical positivists sought for a connection of theoretical to observational terms by giving an analysis of the empirical nature of theoretical terms contrary to that of metaphysical terms. This connection was formulated in what we can call, following Achinstein [2.22], the *Thesis of Partial Interpretation*, which is basically the following: As indicated above, in the brief sketch of

the main features of the RV, the RV allows that a complete empirical semantic interpretation in terms of directly observables is given to  $V_O$  terms and to sentences that belong to  $L_O$ . However, no such interpretation is intended for  $V_T$  terms and consequently for sentences of  $L$  containing them. It is  $TC$  as a whole that supplies the empirical content of  $V_T$  terms. Such terms receive a partial observational meaning indirectly by being related to sets of observation terms via correspondence rules. To use one of Achinstein’s examples [2.22, p. 90]:

“it is in virtue of [a correspondence-rule] which connects a sentence containing the theoretical term *electron* to a sentence containing the observational term *spectral line* that the former theoretical term gains empirical meaning within the Bohr theory of the atom”

Correspondence rules were initially introduced to serve three functions in the RV:

1. To define theoretical terms.
2. To guarantee the cognitive significance of theoretical terms.
3. To specify the empirical procedures for applying theory to phenomena.

In the initial stages of logical positivism it was assumed that if observational terms were cognitively significant, then theoretical terms were cognitively significant if and only if they were explicitly defined in terms of observational terms. The criteria of explicit definition and cognitive significance were abandoned once proponents of the RV became convinced that dispositional terms, which are cognitively significant, do not admit of explicit definitions (Carnap [2.23, 24], also Hempel [2.25, pp. 23–29], and Hempel [2.5]). Consider, for example, the dispositional term *tearable* (let us assume all the necessary conditions for an object to be torn apart hold), if we try to explicitly define it in terms of observables we end up with something like this:

“An object  $x$  is tearable if and only if, if it is pulled sharply apart at time  $t$  then it will tear at  $t$  (assuming for simplicity that pulling and tearing occur simultaneously).”

The above definition could be rendered as  $\forall x (T(x) \leftrightarrow \forall t (P(x, t) \rightarrow Q(x, t)))$ , where,  $T$  is the theoretical term *tearable*,  $P$  is the observational term *pulled apart*, and  $Q$  is the observational term *tears*. But this does not correctly define the actual dispositional property *tearable*, because the right-hand side of the biconditional will be true of objects that are never pulled apart. As a result, some objects that are not tearable and have never being pulled apart will by definition have the property *tearable*.



Because of this, Carnap [2.23, 24] proposed to replace the construal of correspondence rules as explicit definitions, by *reduction sentences* that partially determine the observational content of theoretical terms. A reduction sentence defined the dispositional property tearable as follows:  $\forall x \forall t (P(x, t) \rightarrow (Q(x, t) \leftrightarrow T(x)))$ . That is, (Carnap calls such sentences *bilateral reduction sentences* [2.23, 24]):

“If an object  $x$  is pulled-apart at time  $t$ , then it tears at time  $t$  if and only if it is tearable.”

Unlike the explicit definition case, if  $a$  is a non-tearable object that is never pulled apart then it is not implied that  $T(a)$  is true. What will be implied, in such case, is that  $\forall t (P(a, t) \rightarrow (Q(a, t) \leftrightarrow T(a)))$ , is true. Thus the above shortcoming of explicit definitions is avoided, because a reduction sentence does not completely define a disposition term. In fact, this is also the reason why correspondence rules supply only partial observational content, since many other reduction sentences can be used to supply other empirical aspects of the term *tearable*, for example, being torn by excessively strong shaking. Consequently, although correspondence rules were initially meant to provide explicit definitions and cognitive significance to  $V_T$  terms, these functions were abandoned and substituted by *reduction sentences* and *partial interpretation* (A detailed explication of the changes in the use of correspondence rules through the development of the RV can be found in [2.1]).

Therefore, in its most defensible version the RV could be construed to assign the following functions to correspondence rules: First, they specify empirical procedures for the application of theory to phenomena and second, as a constitutive part of  $TC$ , they supply  $V_T$  and  $L_T$  with partial interpretation. Partial interpretation in the above sense is all the RV needs since, given its goal of distinguishing theoretical from speculative metaphysical terms, it only needs a way to link the  $V_T$  terms to the  $V_O$  terms. The version of the RV that employs correspondence rules for these two purposes motivated two sorts of criticisms. The first concerns the idea that correspondence rules provide partial interpretation to  $V_T$  terms, and the second concerns the function of correspondence rules for providing theory application.

The thesis of partial interpretation came under attack from Putnam [2.4] and Achinstein [2.3, 22]. The structure of their arguments is similar. They both think that partial interpretation is unclear and they attempt to clarify the concept. They do so by suggesting plausible explications for *partial interpretation*. Then they show that for each plausible explication that each of them suggests partial interpretation is either an incoherent notion or inadequate for the needs of the RV. Thus,

they both conclude that any attempt to elucidate the notion of partial interpretation is problematic and that partial interpretation of  $V_T$  terms cannot be adequately explicated. For example, Putnam gives the following plausible explications for *partial interpretation*:

1. To partially interpret  $V_T$  terms is to specify a class of intended models.
2. To partially interpret a term is to specify a verification–refutation procedure that applies only to a proper subset of the extension of the term.
3. To partially interpret a formal language  $L$  is to interpret only part of the language.

In similar spirit, Achinstein gives three other plausible explications. One of Putnam’s counterexamples is that (1) above cannot meet its purpose because the class of intended models, that is, the semantic structures or interpretations that satisfy  $TC$  and which are so intended by scientists, is not well defined (A critical assessment of these arguments can be found in [2.1]).

The other function of correspondence rules, that of specifying empirical procedures for theory application to phenomena, also came under criticism. Suppe [2.1, pp. 102–109] argued that the account of correspondence rules inherent in the RV is inadequate for understanding actual science on the following three grounds:

1. They are mistakenly viewed as components of the theory rather than as auxiliary hypotheses.
2. The sorts of connections (e.g., explanatory causal chains) that hold between theories and phenomena are inadequately captured.
3. They oversimplify the ways in which theories are applied to phenomena.

The first argument is that the RV considers  $TC$  as postulates of the theory. Hence  $C$  is assumed to be an integral part of the theory. But, if a new experimental procedure is discovered it would have to be incorporated into  $C$ , and the result would be a new set of rules  $C'$  that subsequently leads to a new theory  $TC'$ . But obviously the theory does not undergo any change. When new experimental procedures are discovered we only improve our knowledge of how to apply theory to phenomena. So we must think of correspondence rules as auxiliary hypotheses distinct from theory.

The second argument is based upon Schaffner’s [2.26] observation that there is a way in which theories are applied to phenomena, which is not captured by the RV’s account of correspondence rules. This is the case when various auxiliary theories (independent of  $T$ ) are used to describe a *causal sequence*, which obtains between the states described by  $T$  and the observation reports. These causal sequences are descriptions of the mechanisms involved within physical systems to

cause the measurement apparatus to behave as it does. Thus, they supplement theoretical explanations of the observed behavior of the apparatus by linking the theory to the observation reports via a causal story. For example, such auxiliary hypotheses are used to establish a causal link between the motion of an electron ( $V_T$  term) and the spectral line ( $V_O$  term) in a spectrometer photograph. Schaffner's point is that the relation between theory and observation reports is frequently achieved by the use of these auxiliary hypotheses that establish explanations of the behavior of physical systems via causal mechanisms. Without recognizing the use of these auxiliaries the RV may only describe a type of theory application whereby theoretical states are just correlated to observational states. If these kinds of auxiliaries were to be viewed as part of  $C$  then it is best that  $C$  is dissociated from the core theory and is regarded as a separate set of auxiliary hypotheses required for establishing the relation between theory and experiment, because such auxiliaries are obviously not theory driven, but if they are not to be considered part of  $C$  then  $C$  does not adequately explain the theory–experiment relation.

Finally, the third argument is based on Suppes' [2.27, 28] analysis of the complications involved in relating theoretical predictions to observation reports. Suppes observes that in order to reach the point where the two can meaningfully be compared, several epistemologically important modifications must take place on the side of the observation report. For example, Suppes claims, on the side of theory we typically have predictions derived from continuous functions, and on the side of an observation report we have a set of discrete data. The two can only be compared after the observation report is modified accordingly. Similarly, the theory's predictions may be based on the assumption that certain idealizing conditions hold, for example, no friction. Assuming that in the actual experiment these conditions did not hold, it would mean that to achieve a reasonable comparison between theory and experiment the observational data will have to be converted into a corresponding set that reflects the result of an *ideal* experiment. In other words, the actual observational data must be converted into what they would have been had the idealizing conditions obtained. According to Suppes, these sorts of conversion are obtained by employing appropriate *theories of data*. So, frequently, there will not be a direct comparison between theory and observation, but a comparison between theory and observation-altered-by-theory-of-data.

By further developing Suppes' analysis, Suppe [2.8] argues that because of its reliance on the observation–theory distinction, the RV employs correspondence rules in such a way as to blend together unrelated as-

pects of the scientific enterprise. Such aspects are the design of experiments, the interpretation of theories, the various calibration procedures, the employment of results and procedures of related branches of science, etc. All these unrelated aspects are compounded into the correspondence rules. Contrary to the implications of the RV, Suppe claims, in applying a theory to phenomena we do not have any direct link between theoretical terms and observational terms. In a scientific experiment we collect data about the phenomena, and often enough the process of collecting the data involves rather sophisticated bodies of theory. Experimental design and control, instrumentation, and reliability checks are necessary for the collection of data. Moreover, sometimes generally accepted laws or theories are also employed in collecting these data. All these features of experimentation and data collection are then employed in ways as to structure the data into forms (which Suppe calls, *hard data*) that allow meaningful comparison to theoretical predictions. In fact, theory application according to Suppe involves contrasting theoretical predictions to *hard data*, and not to something directly observed [2.8, p. 11]:

“Accordingly, the correspondence rules for a theory should not correlate direct-observation statements with theoretical statements, but rather should correlate *hard data* with theoretical statements.”

In a nutshell, although both Suppes' and Suppe's arguments do not establish with clarity how the theory–experiment relation is achieved they do make the following point: Actual scientific practice, and in particular theory–application, is far more complex than the description given by the RV's account of correspondence rules.

#### 2.1.4 The Cosmetic Role of Models According to the RV

The objection that the RV obscures several epistemologically important features of scientific theories is implicitly present in all versions of the SV of theories. Suppe, however, brings this out explicitly in the form of a criticism (Suppe [2.1, 29, 30]). To clarify the sort of criticism presented by Suppe, we need to make use of some elements of the alternative picture of scientific theories given by the SV, which we shall explore in detail in Sect. 2.2.

The reasoning behind Suppe's argument is the following. Science, he claims, has managed so far to go about its business without involving the observation–theory distinction and all the complexities that it gives rise to. Since, he suggests, the distinction is not required by science, it is important to ask not only whether an

analysis of scientific theories that employs the distinction is adequate or not, that is, the issue on which (as we have seen so far) many of the criticisms of the RV have focused, but whether or not the observation–theory distinction which leads to the notion of correspondence rules subsequently steers toward obscuring epistemological aspects of scientific theorizing.

The sciences, he argues, do not deal with all the detailed features of phenomena and not with phenomena in all their complexity. Rather they isolate a certain number of physical parameters by abstraction and idealization and use these parameters to characterize *physical systems* (Suppe’s terminology is idiosyncratic, he uses the term *physical system* to refer to the abstract entity that an idealized model of the theory represents and not to the actual target physical system), which are highly abstract and idealized replicas of phenomena. A classical mechanical description of the earth–sun system of our solar system, would not deal with the actual system, but with a physical system in which some relevant parameters are abstracted (e.g., mass, displacement, velocity) from the complex features of the actual system. And in which some other parameters are ignored, for example, the intensity of illumination by the sun, the presence of electromagnetic fields, the presence of organic life. In addition, these abstracted parameters are not used in their full complexity to characterize the physical system. Indeed, the description would idealize the physical system by ignoring certain factors or features of the actual system that may plausibly be causally relevant to the actual system. For instance, it may assume that the planets are point masses, or that their gravitational fields are uniform, or that there are no disturbances to the system by external factors and that the system is in a vacuum. What scientific theories do is attempt to characterize the behavior of such physical systems not the behavior of directly observable phenomena.

Although this is admittedly a rough sketch of Suppe’s view, it is not hard to see that the aim of the argument is to lead to the conclusion that the directly observable phenomena are connected to a scientific theory via the physical system. That is to say, (if we put together this idea with the one presented at the end of Sect. 2.1.3 above) the connection between the theory and the phenomena, according to Suppe, requires an analysis of theories and of theory–application that involves a two-stage move. The first move involves the connection between raw phenomena and the *hard data* about the particular target system in question. The second move involves the connection between the *physical system* that represents the *hard data* and the theoretical postulates of the theory. According to Suppe’s understanding of the theory–experiment

relation, the physical system plays the intermediate role between phenomena and theory and this role, which is operative in theory–application, is what needs to be illuminated. The RV implies that the correspondence rules “[...] amalgamate together the two sorts of moves [...] so as to eliminate the physical system” [2.29, p. 16], thus obscuring this important epistemological feature of scientific theorizing.

So, according to Suppe, correspondence rules must give way to this two-stage move, if we are to identify and elucidate the epistemic features of *physical systems*. Suppe’s suggestion is that the only way to accommodate physical systems into our understanding of how theories relate to phenomena is to give models of the theory their representational status. The representational means of the RV are linguistic entities, for example, sentences. Models, within the RV, are denied any representational function. They are conceived exclusively as interpretative devices of the formal calculus, that is, as structures that satisfy subsets of sentences of the theory. This reduces models to meta-mathematical entities that are employed in order to make intelligible the abstract calculus, which amounts to treating them as more or less *cosmetic* aspects of science. But this understanding of the role of models leads to the incapacity of the RV to elucidate the epistemic features of physical systems, and thus obscures – what Suppe considers to be – epistemologically important features of scientific theorizing.

### 2.1.5 Hempel’s Provisos Argument

In one of his last writings, *Hempel* [2.31] raises a problem that suggests a flaw in interpreting the link between empirical theories and experimental reports as mere deduction. Assuming that a theory is a formal axiomatic system consisting of  $T$  and  $C$ , as we did so far, consider Hempel’s example. If we try to apply the theory of magnetism for a simple case we are faced with the following inferential situation. From the observational sentence *b is a metal bar to which iron filings are clinging* ( $S_{O1}$ ), by means of a suitable correspondence rule we infer the theoretical sentence *b is a magnet* ( $S_{T1}$ ). Then by using the theoretical postulates in  $T$ , we infer *if b is broken into two bars, then both are magnets and their poles will attract or repel each other* ( $S_{T2}$ ). Finally using further correspondence rules we derive the observational sentence *if b is broken into two shorter bars and these are suspended, by long thin threads, close to each other at the same distance from the ground, they will orient themselves so as to fall into a straight line* ( $S_{O2}$ ) ([2.31, p. 20]). If the inferential structure is assumed to be deductive then the above structure can be read as follows:  $S_{O1}$  in

combination with the theory deductively implies  $S_{O2}$ . Hempel concludes that this deductivist construal faces a difficulty, which he calls *the problem of provisos*.

To clarify the problem of provisos, we must look into the third inferential step from  $S_{T2}$  to  $S_{O2}$ . What is necessary here is for the theory of magnetism to provide correspondence rules that would turn this step into a deductive inference. The theory however, as Hempel points out, clearly does not do this. In fact, the theory allows for the possibility that the magnets orient themselves in a way other than a straight line, for example, if an external magnetic field of suitable strength and direction is present. This leads to recognizing that the third inferential step presupposes the additional assumption that there are no disturbing influences to the system of concern. Hempel uses the term *provisos*, “[...] to refer to assumptions [of this kind] [...], which are essential, but generally unstated, presuppositions of theoretical inferences” [2.31, p. 23]. Therefore, provisos are presupposed in the application of a theory to phenomena (The problem we saw in Sect. 2.1.3 which Suppes raises, namely that in science theoretical predictions are not confronted with raw observation reports but with observation-altered-by-theory-of-data reports, neighbors this problem but it is not the same. Hempel’s problem of provisos concerns whether it is possible to deductively link theory to observational statements no matter how the latter are constructed).

What is the character of provisos? Hempel suggests we may view provisos as *assumptions of completeness*. For example, in a theoretical inference from a sentence  $S_1$  to another  $S_2$ , a proviso is required that asserts that in a given case “[...] no factors other than those specified in  $S_1$  are present that could affect the event described by  $S_2$ ” [2.31, p. 29]. As, for example, is the case in the application of the Newtonian theory to a two-body system, where it is presupposed that their mutual gravitational attraction are the only forces the system is subjected to. It is clear that [2.31, p. 26]:

“[...] a proviso as here understood is not a clause that can be attached to a theory as a whole and vouchsafe its deductive potency by asserting that in all particular situations to which the theory is applied, disturbing factors are absent. Rather, a proviso has to be conceived as a clause that pertains to some particular application of a given theory and asserts that in the case at hand, no effective factors are present other than those explicitly taken into account.”

Thus, if a theory is conceived as a deductively closed set of statements and its axioms conceived as empirical universal generalizations, as the RV purports, then to apply theory to phenomena, that is, to de-

ductively link theoretical to observational statements, provisos are required. However, in many theory applications there would be an indefinitely large number of provisos, thus trivializing the concept of scientific laws understood as empirical universal generalizations. In other cases, some provisos would not even be expressible in the language of the theory, thus making the deductive step impossible. Hempel’s challenge is that theory–applications presuppose provisos and this does not cohere with the view that theory relates to observation sentences deductively (For an interesting discussion of Hempel’s problem of provisos, see [2.32–35]).

### 2.1.6 Theory Consistency and Meaning Invariance

Feyerabend criticized the logical positivist conception of scientific theories on the ground that it imposes on them a *meaning invariance condition* and a *consistency condition*. By the consistency condition he meant that [2.36, p. 164]

“[...] only such theories are [...] admissible in a given domain which either *contain* the theories already used in this domain, or which are at least *consistent* with them inside the domain.”

By the condition of meaning invariance he meant that [2.36, p. 164]:

“[...] meanings will have to be invariant with respect to scientific progress; that is, all future theories will have to be framed in such a manner that their use in explanations [or reductions] does not affect what is said by the theories, or factual reports to be explained”

Feyerabend’s criticisms are not aimed directly at the RV, but rather at two other claims of logical positivism that are intimately connected to the RV, namely the theses of *the development of theories by reduction* and *the covering law model of scientific explanation*.

A brief digression, in order to look into the aforementioned theses, would be helpful. The development of theories by reduction involves the reduction of one theory (secondary) into a second more inclusive theory (primary). In such developments, the former theory may employ [2.37, p. 342]

“[...] in its formulations [...] a number of distinctive descriptive predicates that are not included in the basic theoretical terms or in the associated rules of correspondence of the primary [theory] [...].”

That is to say, the  $V_T$  terms of the secondary theory are not necessarily all included in the theoretical vocabulary of the primary theory. Nagel builds up his

case based on the example of the reduction of thermodynamics to statistical mechanics. There are several requirements that have to be satisfied for theory reduction to take place, two of which are: (1) the  $V_T$  terms for both theories involved in the reduction must have unambiguously fixed meanings by codified rules of usage or by established procedures appropriate to each discipline, for example, theoretical postulates or correspondence rules. (2) for every  $V_T$  term in the secondary theory that is absent from the theoretical vocabulary of the primary theory, assumptions must be introduced that postulate suitable relations between these terms and corresponding theoretical terms in the primary theory. (See Nagel [2.37, pp. 345–358]. In fact Nagel presents a larger set of conditions that have to hold in order for reduction to take place [2.37, pp. 336–397], but these are the only two relevant to Feyerabend's arguments).

The covering law model of scientific explanation is, in a nutshell, explanation in terms of a deductively valid argument. The sentence to be explained (explanandum) is a logical consequence of a set of law-premises together with a set of premises consisting of initial conditions or other particular facts involved (explanans). For the special case when the explanandum is a scientific theory,  $T'$ , the covering law model can be formulated as follows: A theory  $T$  explains  $T'$  if and only if  $T$  together with initial conditions constitute a deductively valid inference with consequence  $T'$ . In other words, if  $T'$  is derivable from  $T$  together with statements of particular facts involved then  $T'$  is explained by  $T$ . It seems that reduction and explanation of theories go hand in hand, that is, if  $T'$  is reduced to  $T$ , then  $T$  explains  $T'$  and conversely.

Feyerabend points out that Nagel's two assumptions – (1) and (2) above – for theory reduction respectively impose a condition of meaning invariance and a consistency condition to scientific progress. The thesis of development of theories by reduction condemns science to restrict itself to theories that are mutually consistent. But the consistency condition requires that terms in the admissible theories for a domain must be used with the same meanings. Similarly, it can be shown that the covering law model of explanation also imposes these two conditions. In fact, the consistency condition follows from the requirement that the explanandum must be a logical consequence of the explanans, and since the meanings of the terms and statements in a logically valid argument must remain constant, an obvious demand for explanation – imposed

by the covering law model – is that meanings must be invariant. Feyerabend objects to the meaning invariance and the consistency conditions and argues his case inductively by drawing from historical examples of theory change. For example, the concept of mass does not have the same meaning in relativity theory as it does in classical mechanics. Relativistic mass is a relational concept between an object and its velocity, whereas in classical mechanics mass is a monadic property of an object. Similarly, Galileo's law asserts that acceleration due to gravity is constant, but if Newton's law of gravitation is applied to the surface of the earth it yields a variable acceleration due to gravity. Hence, Galileo's law cannot be derived from Newton's law. By such examples, he attempts to undermine Nagel's assumptions (1) and (2) above and establish that neither meaning invariance nor the related notion of theory consistency characterize actual science and scientific progress (see Feyerabend [2.36, 38–40]. Numerous authors have criticized Feyerabend's views. For instance, objections to his views have been raised based on his idiosyncratic analysis of *meaning*, on which his arguments rely. His views are hence not presented here as conclusive criticisms of the RV; but only to highlight that they cast doubt on the adequacy of the theses of theory development by reduction and the covering law model of explanation).

### 2.1.7 General Remark on the Received View

The RV is intended as an explicative and not a descriptive view of scientific theories. We have seen that even as such it is vulnerable to a great deal of criticism. One way or another, all these criticisms rely on one weakness of the RV: Its inability to clearly spell out the nature of theoretical terms (and how they acquire their meaning) and its inability to specify how sentences consisting of such terms relate to experimental reports. This is a weakness that has been understood by the RV's critics to stem from the former's focus on syntax. By shifting attention away from the representational function of models and attempting to characterize theory structure in syntactic terms, the RV makes itself vulnerable to such objections. Despite all of the above criticisms pointing to the difficulty in explicating how theoretical terms relate to observation, I do not think that any one of them is conclusive in the ultimate sense of rebutting the RV. Nevertheless, the subsequent result was that under the weight of all of these criticisms together the RV eventually made room for its successor.

## 2.2 The Semantic View of Scientific Theories

The SV has for the last few decades been the standard-bearer of the view that theories are families of models. The slogan *theories are families of models* was meant by the philosophers that originally put forward the SV to stand for the claim that it is more suitable – for understanding scientific theorizing – that the structure of theory is identified with, or presented as, classes of models. A logical consequence of identifying theory structure with classes of models is that models and modeling are turned into crucial components of scientific theorizing. Indeed, this has been one of the major contributions of the SV, since it unquestionably assisted in putting models and modeling at the forefront of philosophical attention. However, identifying theory structure with classes of models is not a logical consequence of the thesis that models (and modeling) are important components of scientific theorizing. Some philosophers who came to this conclusion have since defended the view that although models are crucial to scientific theorizing, the relation between theory and models is much more complex than that of set-theoretical inclusion. I shall proceed in this section by articulating the major features of the SV; in the process I shall try to clarify the notion of model inherent in the view and also explain – what I consider to be – the main difference among its proponents, and finally I will briefly discuss the criticisms against it, which, nevertheless, do not undermine the importance of models in science.

Patrick Suppes was the first to attempt a model-theoretic account of theory structure. He was one of the major denouncers of the attempts by the logical positivists to characterize theories as first-order calculi supplemented by a set of correspondence rules. (See [2.27, 28, 41–43]; much of the work developed in these papers is included in [2.44]). His objections to the RV led him on the one hand to suggest that in scientific practice the theory–experiment relation is more sophisticated than what is implicit in the RV and that theories are not confronted with raw experimental data (as we have seen in Sect. 2.1) but with, what has since been dubbed, *models of data*. On the other hand, he proposed that theories be construed as collections of models. The models are possible realizations (in the Tarskian sense) that satisfy sets of statements of theory, and these models, according to Suppes, are entities of the appropriate set-theoretical structure. Both of these insights have been operative in shaping the SV.

Suppes urged against standard formalizations of scientific theories. First, no substantive example of a scientific theory is worked out in a formal calculus, and second the [2.28, p. 57]

“[...] very sketchiness [of standard formalizations] makes it possible to omit both important properties of theories and significant distinctions that may be introduced between different theories.”

He opts for set-theoretical axiomatization as the way by which to overcome the shortcomings of standard formalization. As mentioned by *Gelfert*, Chap. 1, Suppe’s example of a set-theoretical axiomatization is classical particle mechanics (CPM). Three axioms of kinematics and four axioms of dynamics (explicitly stated in Chap. 1 of this volume: *The Ontology of Models*) are articulated by the use of predicates that are defined in terms of set theoretical notions. The structure  $\wp = \langle P, T, s, m, f, g \rangle$  can then be understood to be a model of CPM if and only if it satisfies those axioms [2.41, p. 294]. Such a structure is what logicians would label a (semantic) model of the theory, or more accurately a class of models. In general, the model–theoretic notion of a structure,  $S$ , is that of an entity consisting of a nonempty set of individuals,  $D$ , and a set of relations defined upon the former,  $R$ , that is,  $S = \langle D, R \rangle$ . The set  $D$  specifies the domain of the structure and the set  $R$  specifies the relations that hold between the individuals in  $D$ . (Note that as far as the notion of a structure is concerned, it only matters how many individuals are there and not what they are, and it only matters that the relations in  $R$  hold between such and such individuals of  $D$  and not what the relations are. For more on this point and a detailed analysis of the notion of structure *Frigg* and *Nguyen*, Chap. 3).

Models of data, according to Suppes, are possible realizations of the experimental data. It is to models of data that models of the theory are contrasted. The RV would have it that the theoretical predictions have a *direct analogue* in the observation statements. This view however, is, according to Suppes, a distorting simplification. As we have seen in Sect. 2.1.3, Suppes defends the claim that by the use of theories of experimental design and other auxiliary theories, the raw data are regimented into a structural form that bears a relation to the models of the theory. To structure the data, as we saw earlier, various influencing factors that the theory does not account for, but are known to influence the experimental data, must be accommodated by an appropriate conversion of the data into canonical form. This regimentation results in a finished product that Suppes dubbed *models of data*, which are structures that could reasonably be contrasted to the models of the theory. Suppes’ picture of science as an enterprise of theory construction and empirical testing of theories involves establishing a *hierarchy of models*,

roughly consisting of the general categories of models of the theory and models of the data. Furthermore, since the theory–experiment relation is construed as no more than a comparison (i. e., a mapping) of mathematical structures, he invokes the mathematical notion of *isomorphism* of structure to account for the link between theory and experiment. (An isomorphism between structures  $U$  and  $V$  exists, if there is a function that maps each element of  $U$  onto each element of  $V$ ). Hence, Suppes can be read as urging the thesis that defining the models of the theory and checking for isomorphism with models of data, is a rational reconstruction that does more justice to actual science than the RV does.

The backbone of Suppes’ account is the sharp distinction between models of theory and models of data. In his view, the traditional syntactic account of the relation between theory and evidence, which could be captured by the schema:  $(T\&A) \rightarrow E$  (where,  $T$  stands for theory,  $A$  for auxiliaries,  $E$  for empirical evidence), is replaced by theses (1), (2), and (3) below:

1.  $M_T \subseteq TS$ , where  $M_T$  stands for model of the theory  $TS$  for the theory structure, and  $\subseteq$  for the relation of inclusion
2.  $(A\&E\&D) \mapsto M_D$ , where  $M_D$  stands for model of data,  $A$  for auxiliary theories,  $E$  for theories of experimental design etc.,  $D$  for raw empirical data, and  $\mapsto$  for *... used in the construction of...*
3.  $M_T \approx M_D$ , where  $\approx$  stands for mapping of the elements and relations of one structure onto the other.

$M_T \subseteq TS$  expresses Suppes’ view that by defining a theory structure a class of models is laid down for the representation of physical systems.  $(A\&E\&D) \mapsto M_D$  is meant to show how Suppes distances himself from past conceptions of the theory–experiment relation, by claiming that theories are not directly confronted with raw experimental data (collected from the target physical systems) but rather that the latter are used, together with much of the rest of the scientific inventory, in the construction of data structures,  $M_D$ . These data structures are then contrasted to a theoretical model, and the theory–experiment relation consists in an isomorphism, or more generally in a mapping of a data onto a theoretical structure, that is,  $M_T \approx M_D$ . The proponents of the SV would, I believe, concur to the above three general theses. Furthermore, they would concur with two of the theses’ corollaries: that scientific representation of phenomena can be explicated exclusively by mapping of structures, and that all scientific models constructed within the framework of a particular scientific theory are united under a common mathematical or relational structure. We shall look into these two contentions of

the SV toward the end of this section. For now, let me turn our attention to some putative differences between the various proponents of the SV.

Despite agreeing about focusing on the mathematical structure of theories for giving a unitary account of models, it is not hard to notice in the relevant literature that different proponents of the SV have spelled out the details of thesis (1) in different ways. This is because different proponents of the SV have chosen different mathematical entities with which to characterize theory structure. As we saw above, Suppes chooses set theoretical predicates a choice that seems to be shared by *da Costa* and *French* [2.45, 46]. *Van Fraassen* [2.47] on the other hand prefers state-spaces, and *Suppe* [2.30] uses relational systems.

Let us, by way of example, briefly look into van Fraassen’s state-space approach. The objects of concern of scientific theories are physical systems. Typically, mathematical models represent physical systems that can generally be conceived as admitting of a certain set of states. *State-spaces* are the mathematical spaces the elements of which can be used to represent the states of physical systems. It is a generic notion that refers to what, for example, physicists would label as phase space in classical mechanics or Hilbert space in quantum mechanics. A simple example of a state-space would be that of an  $n$ -particle system. In CPM, the state of each particle at a given time is specified by its position  $\mathbf{q} = (q_x, q_y, q_z)$  and momentum  $\mathbf{p} = (p_x, p_y, p_z)$  vectors. Hence the state-space of an  $n$ -particle system would be a Euclidean  $6n$ -dimensional space, whose points are the  $6n$ -tuples of real numbers

$$\langle q_{1x}, q_{1y}, q_{1z}, \dots, q_{nx}, q_{ny}, q_{nz}, \\ p_{1x}, p_{1y}, p_{1z}, \dots, p_{nx}, p_{ny}, p_{nz} \rangle.$$

More generally, a state-space is the collection of mathematical entities such as, vectors, functions, or numbers, which is used to specify the set of possible states for a particular physical system. A model, in van Fraassen’s characterization of theory structure, is a particular sequence of states of the state-space over time, that is, the state of the modeled physical system evolves over time according to the particular sequence of states admitted by the model. State-spaces unite clusters of models of a theory, and they can be used to single out the class of intended models just as set-theoretical predicates would in Suppes’ approach. The presentation of a scientific theory, according to van Fraassen, consists of a description of a *class of state-space types*. As *van Fraassen* explains [2.47, p. 44]:

“[w]henver certain parameters are left unspecified in the description of a structure, it would be more

accurate to say [...] that we described a structure type.”

The Bohr model of the atom, for example, does not refer to a single structure, but to a structure type. Once the necessary characteristics are specified, it gives rise to a structure for the hydrogen atom, a structure for the helium atom, and so forth.

The different choices of different authors on how theory structure is characterized, however, belong to the realm of personal preference and do not introduce any significant differences on the substance of thesis (1) of the SV, which is that all models of the theory are united under an all-inclusive theory structure. So, irrespective of the particular means used to characterize theory structure, the SV construes models as structures (or structure types) and theories as collections of such structures. Neither have disagreements been voiced regarding thesis (2). On the contrary, there seems to be a consensus among adherents of the SV that models of theory are confronted with models of data and not the direct result of an experimental setup (Not much work has been done to convincingly analyze particular scientific examples and to show the details of the use of *models of data* in science; rather, adherents of the SV repeatedly use the notion with reference to something very general with unclear applications in actual scientific contexts).

### 2.2.1 On the Notion of Model in the SV

An obvious objection to thesis (1) would be that a standard formalization could be used to express the theory and subsequently define the class of semantic models metamathematically, as the class of structures that satisfy the sentences of the theory, despite Suppes suggestion that such a procedure would be unnecessarily complex and tedious.

In fact, proponents of the SV have often encouraged this objection. *Van Fraassen* and *Suppe* are notable examples as the following quotations suggest [2.48, p. 326]:

“There are natural interrelations between the two approaches [i. e., the RV and the SV]: An axiomatic theory may be characterized by the class of interpretations which satisfy it, and an interpretation may be characterized by the set of sentences which it satisfies; though in neither case is the characterization unique. These interrelations [...] would make implausible any claim of philosophical superiority for either approach. But the questions asked and methods used are different, and with respect to fruitfulness and insight they may not be on a par with specific contexts or for special purposes.”

*Suppe* [2.30, p. 82]:

“This suggests that theories be construed as propounded abstract structures serving as models for sets of interpreted sentences that constitute the linguistic formulations. These structures are meta-mathematical models of their linguistic formulations, where the same structure may be the model for a number of different, and possibly nonequivalent, sets of sentences or linguistic formulations of the theory.”

From such remarks, one is justifiably led to believe that propounding a theory as a class of models directly defined, without recourse to its syntax, only aims at convenience in avoiding the hustle of constructing a standard formalization, and at easier adaptability of our reconstruction with common scientific practices. Epigrammatically, the difference – between the SV and the RV – would then be methodological and heuristic. Reasons such as this have led some authors to question the *logical* difference between defining the class of models directly as opposed to metamathematically.

Examples are *Friedman* and *Worrall* who in their separate reviews of *van Fraassen* [2.47] ask whether the class of models that constitutes the theory, according to the proponents of the SV, is to be identified with an elementary class, that is, a class that contains all the models (structures) that satisfy a first-order theory. They both notice that not only does *van Fraassen* and other proponents of the SV offer no reason to oppose such a supposition, but also they even encourage it (as in the above quotations). But if that is the case [2.49, p. 276]:

“[t]hen the completeness theorem immediately yields the equivalence of *van Fraassen*’s account and the traditional syntactic account [i. e., that of the RV].”

In other words [2.50, p. 71]:

“So far as logic is concerned, syntax and semantics go hand-in-hand – to every consistent set of first-order sentences there corresponds a nonempty set of models, and to every normal (elementary) set of models there corresponds a consistent set of first-order sentences.”

If we assume (following *Friedman* and *Worrall*) that the proponents of the SV are referring to the elementary class of models then the preceding argument is sound. The SV, in agreement with the logical positivists, retains formal methods as the primary tool for philosophical analysis of science. The only new elements of its own would be the suggestions that first it is more convenient that rather than developing these methods



using proof–theory we should instead use formal semantics (model–theory), and second we should assign to models (i. e., the semantic interpretations of sets of sentences) a representational capacity.

Van Fraassen, however, resists the construal of the class of models of the SV with an elementary class (See *van Fraassen* [2.51, pp. 301–303] and his [2.52]). Let me rehearse his argument. The SV claims that to present a theory is to define a class  $M$  of models. This is the class of structures the theory makes available for modeling its domain. For most scientific theories, the real number continuum would be included in this class. Now his argument goes, if we are able to formalize what is meant to be conveyed by  $M$  in some appropriate language, then we will be left with a class  $N$  of models of the language, that is, the class of models in which the axioms and theorems of the language are satisfied. Our hope is that every structure in  $M$  occurs in  $N$ . However, the real number continuum is infinite and [2.52, p. 120]:

“[t]here is no elementary class of models of a denumerable first-order language each of which includes the real numbers. As soon as we go from mathematics to metamathematics, we reach a level of formalization where many mathematical distinctions cannot be captured.”

Furthermore, “[t]he Löwenheim–Skolem theorems [...] tell us [...] that  $N$  contains many structures not isomorphic to any member of  $M$ ” [2.51, p. 302]. Van Fraassen relies, here, on the following reasoning: The Löwenheim–Skolem theorem tells us that all satisfiable first-order theories that admit infinite models will have models of all different infinite cardinalities. Now models of different cardinality are nonisomorphic. Consequently, every theory that makes use of the real number continuum will have models that are not isomorphic to the intended models (i. e., nonstandard interpretations) but which satisfy the axioms of the theory. So van Fraassen is suggesting that  $M$  is the intended class of models, and since the limitative meta-theorems tell us that it cannot be uniquely determined by any set of first-order sentences we can only define it directly. Here is his concluding remark [2.51, p. 302]:

“The set  $N$  contains [...] [an] image  $M^*$  of  $M$ , namely, the set of those members of  $N$  which consist of structures in  $M$  accompanied by interpretations therein of the syntax. But, moreover, [...]  $M^*$  is not an elementary class.”

Evidently, van Fraassen’s argument aims to establish that the directly defined class of models is not an elementary class. It is hard, however, to see that defining the models of the theory directly without resort to formal syntax yields only the intended models of theory

(i. e., excludes all nonstandard models), despite the possibility that one could see the prospect of the SV being heuristically superior to the RV. (Of course, we must not forget that this superiority would not necessarily be the result of thesis (1) of the SV, but it could be the result of its consequence of putting particular emphasis on the significance of scientific models that, as noted earlier, does not logically entail thesis (1)).

Let us, for the sake of argument, ignore the Friedman–Worrall argument. Now, according to the SV, models of theory have a dual role. On the one hand, they are devices by which phenomena are represented, and on the other, they are structures that would satisfy a formal calculus were the theory formalized. The SV requires this dual role. First because the representational role of models is the way by which the SV accounts for scientific representation without the use of language; and second because the role of interpreting a set of axioms ensures that a unitary account of models is given. Now, *Thompson-Jones* [2.53] notices that the notion of model implicit in the SV is either that of an interpretation of a set of sentences or a mathematical structure (the disjunction is of course inclusive). He analyzes the two possible notions and argues that the SV becomes more tenable if the notion of model is only understood as that of a mathematical structure that functions as a representation device. If that were the case then the adherents of the SV could possibly claim that defining the class of structures directly indeed results in something distinct from the metamathematical models of a formal syntax. *Thompson-Jones*’ suggestion, however, would give rise to new objections. Here is one. It would give rise to the following question: How could a theory be identified with a class of models (i. e., mathematical structures united under an all-inclusive theory structure) if the members of such a class do not attain membership in the class because they are interpretations of the same set of theory axioms? In other words, the proponents of the SV would have to explain what it is that *unites* the mathematical models other than the satisfaction relation they have to the theoretical axioms. To my knowledge, proponents of the SV have not offered an answer to this question. If *Thompson-Jones*’ suggestion did indeed offer a plausible way to overcome the Friedman–Worrall argument then the SV would have to abandon the quest of giving a unitary account of models. Given the dual aim of the SV, namely to give a unitary account of models and to account for scientific representation by means of structural relations, it seems that the legitimate notion of model integral to this view must have these two-hard to reconcile-roles; namely, to function both as an interpretation of sets of sentences and as a representation of phenomena. (Notice that this dual function of models is

an aspect of all versions of the SV, independent of how one chooses to characterize theory structure and of how one chooses to interpret that structure).

### 2.2.2 The Difference Between Various Versions of the SV

The main difference among the various versions of the SV relates to two intertwined issues that relate to thesis (3), namely how the theory structure is construed and how the theory–experiment mapping relation is construed. To a first approximation we could divide the different versions of the SV, from the perspective of these two issues, into two sorts. Those in which particular emphasis is given to the presence of abstraction and idealization in scientific theorizing for explicating the theory–experiment (or model–experiment) relation, and those in which the significance of this nature of scientific theorizing is underrated.

#### Idealization and Abstraction Underrated

*Van Fraassen* (Suppes most probably could be placed in this group too), for example, seems to be a clear case of this sort. Here is how he encapsulates his conception of scientific theories and of how theory relates to experiment [2.47, p. 64]:

“To present a theory is to specify a family of structures, its *models*; and secondly, to specify certain parts of those *models* (*the empirical substructures*) as candidates for the direct representation of observable phenomena. The structures which can be described in experimental and measurement reports we can call *appearances*: The theory is empirically adequate if it has some model such that all appearances are isomorphic to empirical substructures of that model.”

Appearances (which is van Fraassen’s term for *models of data*) are relational structures of measurements of observable aspects of the target physical system, for example, relative distances and velocities. For example, in the Newtonian description of the solar system, as *van Fraassen* points out, the relative motions of the planets “[...] form relational structures defined by measuring relative distances, time intervals, and angles of separation” [2.47, p. 45]. Within the theoretical model for this physical system, “[...] we can define structures that are meant to be exact reflections of those appearances [...]” [2.47, p. 45]. Van Fraassen calls these *empirical substructures*. When a theory structure is defined each of its models, which are candidates for the representation of phenomena, includes empirical substructures. So within representational models we could specify a division between observable/nonobservable features

(albeit this division is not drawn in linguistic terms), and the empirical substructures of such models are assumed to be isomorphic to the observable aspects of the physical system. In other words, the theory structure is interpreted as having distinctly divided observable and nonobservable features, and the theory–experiment relation is interpreted as being an isomorphic relation between the data model and the observable parts of the theoretical model. Now, the state-space is a class of models, it thus includes – for CPM – many models in which the world is a Newtonian mechanical system. In fact, it seems that the state-space includes (unites) all logically possible models, as the following dictum suggests ([2.52, p. 111], [2.54, p. 226]):

“In one such model, nothing except the solar system exists at all; in another the fixed stars also exist, and in a third, the solar system exists and dolphins are its only rational inhabitants.”

According to van Fraassen, the theory is empirically adequate if we can find a model of the theory in which we can specify empirical substructures that are isomorphic to the data model. The particular view of scientific representation that resides within this idea is this: *A model represents its target if and only if it is isomorphic to a data model constructed from measurements of the target*. Not much else seems to matter for a representation relation to hold but the *isomorphism condition*. Many would argue, however, that such a condition for the representation relation is too strong to explicate how actual scientific models relate to experimental results and would object to this view on the ground that for isomorphism to occur it would require that target physical systems occur under highly idealized conditions or in isolated circumstances. (Admittedly, it would not be such a strong requirement for models that would only describe observable aspects of the world. In such cases isomorphism could be achieved, but at the expense of the model’s epistemic significance. I do not think, for instance, that such models would be of much value to a science like Physics as, more often than not, they would be useless in predicting the future behavior of their targets).

#### Idealization and Abstraction Highlighted

In the second camp of the SV, we encounter several varieties. One of these is *Suppe* [2.30], who interprets theory structure and the theory–experiment relation as follows. Theories characterize particular classes of target systems. However, target systems are not characterized in their full complexity, as already mentioned in Sect. 2.1.4. Instead, Suppe’s understanding is that certain parameters are abstracted and employed in this characterization. In the case of CPM, these are the posi-

tion and momentum vectors. These two parameters are abstracted from all other characteristics that target systems may possess. Furthermore, once the factors, which are assumed to influence the class of target systems in the theory's intended scope, have been abstracted the characterization of physical systems (as mentioned in Sect. 2.1.4, physical systems in Suppe's terminology refer to the abstract entities that models of the theory represent and not to the actual target systems) still does not *fully* account for target systems. Physical systems are not concerned with the actual values of the parameters the particulars possess, for example, actual velocities, but with the values of these parameters under certain conditions that obtain only within the physical system itself. Thus in CPM, where the behavior of dimensionless point-masses are studied in isolation from outside interactions, physical systems characterize this behavior only by reference to the positions and momenta of the point-masses at given times.

An example can serve to demonstrate Suppe's idea in bit more detail. The linear harmonic oscillator, that is, a *mathematical instrument*, is expressed by the following equation of motion  $\ddot{x} + (k/m)x = 0$ , which is the result of applying Newton's second law to a linear restoring force. The mathematical model is interpreted (and thus characterizes a *physical system*) as follows: Periodic oscillations are assumed to take place with respect to time,  $x$  is the displacement of an oscillating mass-point, and  $k$  and  $m$  are constant coefficients that may be replaced by others. When the mathematical parameters in the above equation are linked to features of a specific object, the equation can be used to model for instance the torsion pendulum, that is, an elastic rod connected to a disk that oscillates about an equilibrium position. This sort of linking of mathematical terms to features of objects could be understood to be a manifestation of what Giere calls identification. Giere introduces a useful distinction between *interpretation* and *identification* [2.55, p. 75]:

“[...] [Interpretation] is the linking of the mathematical symbols with *general terms*, or concepts, such as *position*[...] [Identification] is the linking of a mathematical symbol with some feature of a *specific object*, such as *the position of the moon*.”

In the torsion pendulum model,  $x$  is identified with the angle of twist,  $k$  with the torsion constant, and  $m$  with the moment of inertia. By linking the mathematical symbols of a model to features of a target system we can reasonably assume, according to Suppe, that the model could be associated with an actual system of the world; the model characterizes, as Suppe would say in his own jargon, “a causally possible physical system.”

However, even when a certain mathematical product of theory is identified with a causally possible physical system, we still know that typically the situation described by the physical system does not obtain. The actual torsion pendulum apparatus is subject to a number of different factors (or may have a number of different characteristics) that may or may not influence the process of oscillation. Some influencing factors are the amplitude of the angle of oscillation, the mass distribution of the rod and disc, the nonuniformity of the gravitational field of the earth, the buoyancy of the rod and disc, the resistance of the air and the stirring up of the air due to the oscillations. In modeling the torsion pendulum by means of the linear harmonic oscillator the physical system is abstracted from factors assumed to influence the oscillations in the same manner as from those assumed not to. Therefore, the replicating relation between the physical system,  $P$ , and the target system,  $S$ , which Suppe urges cannot be understood as one of identity or isomorphism. *Suppe* is explicit about this [2.30, p. 94]:

“The attributes in  $P$  determine a sequence of states over time and thus indicate a possible behavior of  $S$  [...] Accordingly,  $P$  is a kind of *replica* of  $S$ ; however, it need not replicate  $S$  in any straight-forward manner. For the state of  $P$  at  $t$  does not indicate what attributes the particulars in  $S$  possess at  $t$ ; rather, it indicates what attributes they *would have* at  $t$  were the abstracted parameters the only ones influencing the behavior of  $S$  and were certain idealized conditions met. In order to see how  $P$  replicates  $S$  we need to investigate these abstractive and idealizing conditions holding between them.”

In summary, the replicating relation is counterfactual: If the conditions assumed to hold for the description of the physical system were to hold for the target system, then the target system would behave in the way described by the physical system. The behavior of actual target systems, however, may be subject to other unselected parameters or other conditions, for which the theory does not account.

The divergence of Suppe's view from that of van Fraassen is one based primarily on the representation relation of theory to phenomena. Suppe understands the theory structure as being a highly abstract and idealized representation of the complexities of the real world. Van Fraassen disregards this because he is concerned with the observable aspects of theories and assumes that these can, to a high degree of accuracy, be captured by experiments. Thus van Fraassen regards theories as containing empirical substructures that stand in isomorphic relations to the observable aspects of the world. Suppe's

understanding of theory structure, however, points to a significant drawback present in van Fraassen's view: How can isomorphism obtain between a data model and an empirical substructure of the model, given that the model is abstract and idealized? Suppe's difference with van Fraassen's view of the representation relation and of the epistemic inferences that can be drawn from it is this, if indeed it is the case that isomorphism obtains between a data model and an empirical substructure, then it is so for either of two reasons: (1) the experiment is highly idealized, or (2) the data model is converted to what the measurements *would have been* if the influences that are not accounted by the theory did not have any effect on the experimental setup. This is a significantly different claim from what van Fraassen would urge, to wit that the world or some part of it is isomorphic to the model. According to Suppe's understanding of theory structure, no part of the world is or can be isomorphic to a model of the theory, because abstraction and idealization are involved in scientific theorizing.

*Geire* [2.55] is another example of a version of the SV that places the emphasis on abstraction and idealization. Following Suppes and van Fraassen, Geire understands theories as classes of models. He does not have any special preference about the mathematical entities by which theory structure is characterized, but he is interested in looking at the characteristics of actual science and how these could be captured by the SV. This leads him to a similar claim as Suppe. He claims that although he does not see any logical reason why a real target system could not be isomorphic to a model, nevertheless for the examples of models found in mechanics texts, typically, no claim of isomorphism is made, indeed "[...] the texts often explicitly note respects in which the model fails to be isomorphic to the real system" [2.55, p. 80]. He attributes this to the abstract and idealized nature of models of the theory. His solution is to substitute the strict criterion of isomorphism, as a way by which to explicate the theory–experiment relation, with that of similarity in relevant respects and degrees between the model and its target.

Finally, there is another example of a version of the SV that also gives attention to idealization and abstraction, namely the version advocated by *da Costa* and *French* in [2.45, 46, 56]. They do this indirectly by interpreting theories as partial structures, that is, structures consisting of a domain of individuals and a set of partial relations defined on the domain, where a partial relation is one that is not defined for all the  $n$ -tuples of individuals of the domain for which it presumably holds. If models of theory are interpreted in this manner and if it is assumed that models of data are also partial structures, then the theory–experiment relation

is explicated by *da Costa* and *French* [2.46] as a partial isomorphism. A partial isomorphism between two partial structures  $U$  and  $V$  exists when a partial substructure of  $U$  is isomorphic to a partial substructure of  $V$ . In other words, partial isomorphism exists when some elements of the set of relations in  $U$  are mapped onto elements of the set of relations in  $V$ . If a model of theory is partially isomorphic to a data model then, *da Costa* and *French* claim, the model is partially true. The notion of partial truth is meant to convey a pragmatic notion of truth, which plausibly could avoid the problems of correspondence or complete truth, and capture the commonplace idea that theories (or models) are incomplete or imperfect or abstract or idealized descriptions of target systems.

In conclusion, if we could speak of *different* versions of the SV and not just different formulations of the same idea, if, in other words, the proposed versions of the semantic conception of theories can be differentiated in any significant way amongst them, it is on the basis of how thesis (3) is conceived: There are those that understand the representation relation,  $M_T \approx M_D$ , as a strict isomorphic relation, and those that construe it more liberally, for example, as a similarity relation. In particular, van Fraassen prefers an isomorphic relation between theory and experiment, whereas Suppe and others understand theories as being abstract and idealized representations of phenomena. It would seem therefore that particular criticisms would not necessarily target both versions. This has not been the case however, as we shall examine in the next two subsections. Critics of the SV have either targeted theses (1) and (2) and the unitary account of models implicit in the SV, or thesis (3) and the representation relation however the latter is conceived. The arguments against the unitary account of scientific models, which obviously aim indiscriminately at all versions of the SV, will be explored in Sect. 2.2.4. The arguments against the nature of the representation relation implied by the SV, which shall be explored in Sect. 2.2.3, if properly adapted affect both versions of the SV.

### 2.2.3 Scientific Representation Does not Reduce to a Mapping of Structures

*Suarez* [2.57] presents five arguments against the idea that scientific representation can be explicated by appealing to a structural relation (like isomorphism or similarity) that may hold between the representational device and the represented target. (*Suarez* [2.57] also develops his arguments for other suggested interpretations of theses (3), such as partial isomorphism). These arguments, which are summarized below, imply that

the representational capacity of scientific models cannot derive from having a structural relation with its target. Suarez's first argument is that in science many disparate things act as representational devices, for example, a mathematical equation, or a Feynman diagram, or an architect's model of a building, or the double helix macro-model of the DNA molecule. Neither isomorphism nor similarity can be applied to such disparate representational devices in order to explicate their representational function. A similar point is also made by *Downes* [2.58], who by also exploring some examples of scientific models, argues that models in science relate to their target systems in various ways, and that attempts to explicate this relation by appeal to isomorphism or similarity does little to serve the purpose of understanding the theory–experiment relation.

The second argument concerns the logical properties of representation vis-a-vis those of isomorphism and similarity. Suarez explains that representation is nonsymmetric, nonreflexive and nontransitive. If scientific representation is a type of representation then any attempt to explicate scientific representation cannot imply different logical features from representation. But appeal to a structural relation does not accomplish this, because “[...] similarity is reflexive and symmetric, and isomorphism is reflexive, symmetric and transitive” [2.55, p. 233].

His third argument is that any explication of representation must allow for misrepresentation or inaccurate representation. Misrepresentation, he explains, occurs either when the target of a representation is mistaken or when a representation is inaccurate because it is either incomplete or idealized. Neither isomorphism nor similarity allows for the first kind of misrepresentation and isomorphism does not allow for the second kind. Although, similarity does account for the second kind of representation, Suarez argues, it does so in a restrictive sense. That is, if we assume that an incomplete representation is given according to theory *X* then similarity does account for misrepresentation. However, if a complete representation were given according to theory *X* (i. e., if we have similarity in all relevant respects that *X* dictates) but the predictions of this representation still diverge from measurements of the values of the target's attributes then similarity does not account for this kind of misrepresentation.

The fourth argument is that neither isomorphism nor similarity is necessary for representation. Our intuitions about the notion of representation allow us to accept the representational device derived from theory *X* as a *representation* of its target, even though we may know that isomorphism or similarity does not obtain because, for example, an alternative theory *Y* not only gives us better predictions about the target but

also tells us why *X* fails to produce representational devices that are isomorphic or similar to their targets. A different argument but with the same conclusion is given by *Portides* [2.59], who argues that isomorphism, or other forms of structural mapping, is not necessary for representation because it is possible to explicate the representational function of some successful quantum mechanical models, which are not isomorphic to their targets. Suarez's final argument is that neither isomorphism nor similarity is sufficient for representation. In other words, even though there may not be a representation relation between *A* and *B*, *A* and *B* may, however, be isomorphic or similar.

Aiming at the same feature of the SV as Suarez, *Frigg* [2.60] reiterates some of the arguments above and gives further reasons to fortify them, but he also presents two more arguments that undermine the notion of representation as dictated by thesis (3) of the SV. Employed in his first argument is a particular notion of abstractness of concepts advocated by *Cartwright* [2.61]. A concept is considered abstract in relation to a set of more concrete concepts if for the former to apply it is necessary that one of its concrete instances apply. One of *Frigg's* intuitive examples is that the concept of traveling is more abstract than the concept of sitting in a moving train. So according to this sense of *abstractness* the concept of traveling applies whenever one is sitting in a moving train and that the abstract concept does not apply if one is not performing some action that belongs to the set of concrete instances of traveling. *Frigg* then claims, “[...] that possessing a structure is abstract in exactly this sense and it therefore does not apply without some more concrete concepts applying as well” [2.60, p. 55]. He defends this claim with the following argument. Since to have a structure means to consist of a set of individuals which enter into some relations, then it follows that whenever the concept of possessing a structure applies to *S* the concept of being an individual applies to members of a set of *S* and the concept of being in a relation applies to some parts of that set. The concepts of being an individual and being in a relation are abstract in the above sense. For example, given the proper context, for *being an individual* to apply, *occupying a certain space-time region* has to apply. Similarly, given the proper context, for *being in a relation* to apply it must be the case that *being greater than* applies. Therefore, both being an individual and being in a relation are abstract. Thus *Frigg* concludes, *possessing a structure* is abstract; hence for it to apply, it must be the case that a concrete description of the target applies. Because, the claim that the representation relation can be construed as an isomorphism (or similarity) of structures presupposes that the target possesses a structure, *Frigg* concludes that such a claim “[...] pre-

supposes that there is a more concrete description that is true of the [target] system” [2.60, p. 56]. This argument shows that to reduce the representation relation to a mapping of structures the proponents of the SV need to invoke nonstructural elements into their account of representation, so pure and simple reduction fails.

Frigg’s second argument, as he states, is inductive. He examines several examples of systems from different contexts in order to support the claim that a target system does not have a unique structure. For a system to have a structure it must be made of individuals and relations, but slicing up the physical systems of the world into individuals and relations is dependent on how we conceptualize the world. The world itself does not provide us with a unique slicing. “Because different conceptualizations may result in different structures there is no such thing as the one and only structure of a system” [2.60, p. 57]. One way that Frigg’s argument could be read is this: Thesis (2) of the SV implies that the measurements of an experiment are structured to form a data model. But, according to Frigg, this structuring is not unique. So the claim of thesis (3), that there is, for example, an isomorphism between a theoretical model and a data model is not epistemically informative since there may be numerous other structures that could be constructed from the data that are not isomorphic to the theoretical model.

### 2.2.4 A Unitary Account of Models Does not Illuminate Scientific Modeling Practices

The second group of criticisms against the SV consists of several heterogeneous arguments stemming from different directions and treating a variety of features and functions of models. Despite this heterogeneity, they can be grouped together because they all indirectly undermine the idea that the unitary account of scientific models given by employing a set theoretical (or other mathematical) characterization of theory structure is adequate for understanding the notion of representational model and the model–experiment relation. This challenge to the SV is indirect because the main purpose of these arguments is to illuminate particular features of actual scientific models. In highlighting these features, these arguments illustrate that actual representational models in science are constructed in ways that are incompatible with the SV, they function in ways that the SV does not adequately account for and they represent in ways that is incompatible with the SV’s account of representation; furthermore, they indicate that models in science are complex entities that cannot be thoroughly understood by unitary accounts such as set-theoretical inclusion. In other words, a conse-

quence of most of these arguments is that the unitary account of models that the SV provides through thesis (1) that all models are constitutive parts of theory structure, obscures the particular features that representational scientific models demonstrate.

One such example is *Morrison* [2.62], who argues that models are partially autonomous from the theories that may be responsible for instigating their construction. This partial autonomy is something that may derive from the way they function but also from the way they are constructed. She discusses Prandtl’s hydrodynamic model of the boundary layer in order to mark out that the inability of theory to provide an explanation of the phenomenon of fluid flow did not hinder scientific modeling. Prandtl constructed the model with little reliance on high-level theory and with a conceptual apparatus that was partially independent from the conceptual resources of theory. This partial independence in construction, according to Morrison, gives rise to functional independence and renders the model partially autonomous from theory. Furthermore, *Morrison* raises another issue (see [2.62], as well as [2.63]); that theories, and hence theoretical models as direct conceptual descendants of theory, are highly abstract and idealized descriptions of phenomena, and therefore they represent only the general features of phenomena and do not explain the specific mechanisms at work in physical systems. In contrast, actual representational scientific models – that she construes as partially autonomous mediators between theories and phenomena – are constructed in ways that allow them to function as explanations of the specific mechanisms and thus function as sources of knowledge about corresponding target systems and their constitutive parts. (As she makes clear in *Morrison* [2.64], to regard a model as partially independent from theory does not mean that theory plays an unimportant role in its construction). This argument, in which representational capacity is correlated to the explanatory power of models, is meant to achieve two goals. Firstly, to offer a way by which to go beyond the narrow understanding of scientific representation as a mapping relation of structure, and second, to offer a general way to understand the representational function of both kinds of models that physicists call theory-driven and phenomenological (In *Portides* [2.65] a more detailed contrast between Morrison’s view of the representation relation and that of the SV is offered). *Cartwright* et al. [2.66] and *Portides* [2.67] have also argued that by focusing exclusively on theory-driven models and the mapping relation criterion, the SV obscures the representational function of phenomenological models and also many aspects of scientific theorizing that are the result of phenomenological methods.

It is noteworthy that the unitary account that the SV offers may be applicable to theory-driven models. Whether that is helpful or not is debatable. However, more often than not representation in science is achieved by the use of phenomenological models or phenomenological elements incorporated into theory-driven models. One aspect of Morrison's argument is that if we are not to dismiss the representational capacity of such models we should give up unitary accounts of models. Cartwright makes a similar point but her approach to the same problem is from another angle.

*Cartwright* [2.61, 68] claims that theories are highly abstract and thus do not and cannot represent what happens in actual situations. Cartwright's observation seems similar to versions of the SV such as Suppe's, however her approach is much more robust. To claim that theories represent what happens in actual situations, she argues, is to overlook that the concepts used in them – such as, *force functions* and *Hamiltonians* – are abstract. Such abstract concepts could only apply to the phenomena whenever more concrete descriptions (as those present in models) can stand-in for them and for this to happen the bridge principles of theory must mediate. Hence the abstract terms of theory apply to actual situations via bridge principles, and this makes bridge principles an operative aspect of theory-application to phenomena. It is only when bridge principles sanction the use of theoretical models that we are led to the construction of a model – with a relatively close relation to theory – that represents the target system. But Cartwright observes that there are only a small number of such theoretical models that can be used successfully to construct representations of physical systems and this is because there are only a handful of theory bridge principles. In most other cases, where no bridge principles exist that enable the use of a theoretical model, concrete descriptions of phenomena are achieved by constructing phenomenological models. Phenomenological models are constructed with minimal aid from theory, and surely there is no deductive (or structural) relation between them and theory. The relation between the two should be sought in the nature of the abstract–concrete distinction between scientific concepts, according to Cartwright. Models in science, whether constructed phenomenologically or by the use of available bridge principles, encompass descriptions that are in some way independent from theory because they are made up of more concrete conceptual ingredients. A weak reading of this argument is that the SV could be a plausible suggestion for understanding the structure of scientific theories for use in foundational work. But in the context of utilizing the theory to construct representations of phenomena, focusing on the structure of theory does not illuminate

much because it is not sufficient as to account for the abstract–concrete distinction that exists between theory and models. A stronger reading of the argument is that the structure of theories is completely irrelevant to how theories represent the world, because they just do not represent it at all. Only models represent pieces of the world and they are partially independent from theory because they are constituted by concrete concepts that apply only to particular physical systems.

Other essays in the volume by *Morgan* and *Morrison* [2.69] discuss different aspects of partial independence of models from theory. Here are two brief examples that aim to show the partial independence of model construction from theory. *Suarez* [2.70] explains how simplifications and approximations that are introduced into representational models (such as the London brothers model of superconductivity) are decided independently of theory and of theoretical requirements. This process gives rise to a model that mediates in the sense that the model itself is the means by which corrections are established that may be incorporated into theory in order to facilitate its applications. But even in cases of models that are strongly linked to theory such as the MIT-bag model of quark confinement, *Hartmann* [2.71] argues, many parts of the model are not motivated by theory but by an accompanying *story* about quarks. From the empirical fact that quarks were not observed physicists were eventually led to the hypothesis that quarks are confined. But confinement is not something that follows from theory. Nevertheless, via the proper amalgam of theory and *story* about quarks the MIT-bag model was constructed to account for quark confinement.

I mentioned earlier in Sect. 2.2.2 that *Giere* [2.55] is also an advocate of the SV. However, his later writings [2.72, 73] suggest that he makes a gradual shift from his earlier conception of representational models in science to a view that neighbors that of Morrison and Cartwright. Even in *Giere* [2.55] the reader notices that he, unlike most other advocates of the SV, is less concerned with the attempt to give a unitary account of models and more concerned with the importance of models in actual scientific practices. But in [2.72] and [2.73] this becomes more explicit. *Giere* [2.55] espouses the idea that the laws of a theory are definitional devices of theoretical models. This view is compatible with the use of scientific laws in the SV. However, in *Giere* [2.72, p. 94] he suggests that scientific laws “[...] should be understood as rules devised by humans to be used in building models to represent specific aspects of the natural world.” It is patent that operating as rules for building models is quite a different thing from understanding laws to be the means by which models are defined. The latter view is in line with the three

theses of the SV; the former however is only in line with the view that models are important in scientific theorizing. Moreover, in *Giere* [2.73] he makes a more radical step in distinguishing between the abstract models (which he calls *abstract objects*) defined by the laws and those models used by scientists to represent physical systems (which he calls *representational models*). The latter [2.73, p. 63]

“[...] are designed for use in representing aspects of the world. The abstract objects defined by scientific principles [i. e., scientific laws] are, on my view, not intended directly to represent the world.”

Giere points to the important difference between the SV and its critics. The SV considers the models that the theory directly delivers representations of target systems of the world. Its critics do not think that; they argue that many successful representational models are constructed by a variety of conceptual ingredients and thus have a degree of autonomy from theory. But if each representational model is partially autonomous from the theory that prompted its construction then a unitary account of representational models does not seem to be much enlightening in enhancing our understanding of why models are so important in scientific theorizing.

### 2.2.5 General Remark on the Semantic View

Just like its predecessor the SV employs formal methods for the philosophical analysis of scientific theories. In the SV, models of the theory are directly defined by the laws of the theory, and are thus united under a common mathematical structure. Of course, mathematical equations satisfy a structure, no one disputes that mathematically formulated theories can be presented in terms of mathematical structures. Nonetheless, keen to overcome the philosophical problems associated with the RV and its focus on the syntactic elements of theories, the proponents of the SV take the idea of presenting theories structurally one step further. They claim that the SV not only offers a canonical structural formulation for theories, into which any theory can be given an equivalent reformulation (an idea that, no doubt, is useful for the philosophy of mathematics), but they also contend that a scientific theory represents phenomena *because* this structure can be linked to empirical data. To defend this assertion, the proponents of the SV assume that in science there is a sharp distinction between models of theory and models of data and argue

that scientific representation is no more than a mapping relation between these two kinds of structures. As we have seen, serious arguments against the idea that representation can be reduced to structural mapping have surfaced; and these arguments counter the SV independently of how the details of the mapping relation is construed.

Furthermore, the SV implies that by defining a theory structure an indefinite number of models that are thought to be antecedently available for modeling the theory's domain are laid down. Neither this position has gone unnoticed. Critics of the SV claim that this idea does not do justice to actual science because it undervalues the complexities involved in actual scientific model construction and the variety of functions that models have in science, but more importantly because it obscures the features of representational models that distinguish them from the models that are direct descendants of theory.

I claimed that the SV employs a notion of model that has two functions – interpretation and representation. In addition, it requires models that have this dual role to be united under a common structure. It is hard to reconcile these two ideas and do justice to actual science. The devices by which the theoretical models are defined, according to the SV, are the laws of the theory. Hence the laws of the theory provide the constraints that determine the structure of these models. Now, it is not hard to see that models viewed as interpretations are indeed united under a common structure determined by the laws of the theory. What is problematic, however, is that the SV assumes that models that are interpretations also function as representations and this means that models functioning as representations can be united under a common structure. The truth value of the conjunction *models are interpretations and representations* is certainly not a trivial issue. When scientists construct representational models, they continuously impose constraints that alter their initial structure. The departure of the resulting constructs from the initial structure is such that it is no longer easily justified to think of them all as united under a common theory structure. Indeed, in many scientific cases this departure of individual representational models is such that they end up having features that may be incompatible with other models that are also instigated by the same theory. These observations lead to the thought that the model-theory and the model-experiment relations may in the end be too complex for our formal tools to capture.



## References

- 2.1 F. Suppe: The search for philosophic understanding of scientific theories. In: *The Structure of Scientific Theories*, ed. by F. Suppe (Univ. Illinois Press, Urbana 1974) pp. 1–241
- 2.2 P. Achinstein: The problem of theoretical terms, *Am. Philos. Q.* **2**(3), 193–203 (1965)
- 2.3 P. Achinstein: *Concepts of Science: A Philosophical Analysis* (Johns Hopkins, Baltimore 1968)
- 2.4 H. Putnam: What theories are not. In: *Logic, Methodology and Philosophy of Science*, ed. by E. Nagel, P. Suppes, A. Tarski (Stanford Univ. Press, Stanford 1962) pp. 240–251
- 2.5 Theoretician's dilemma: A study in the logic of theory construction. In: *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, ed. by C. Hempel, C. Hempel (Free Press, New York 1958) pp. 173–226
- 2.6 R. Carnap: The methodological character of theoretical concepts. In: *Minnesota Studies in the Philosophy of Science: The Foundations of Science and the Concepts of Psychology and Psychoanalysis*, Vol. 1, ed. by H. Feigl, M. Scriven (Univ. Minnesota Press, Minneapolis 1956) pp. 38–76
- 2.7 R. Carnap: *Philosophical Foundations of Physics* (Basic Books, New York 1966)
- 2.8 F. Suppe: Theories, their formulations, and the operational imperative, *Synthese* **25**, 129–164 (1972)
- 2.9 N.R. Hanson: *Patterns of Discovery: An Inquiry into the Conceptual Foundations of Science* (Cambridge Univ. Press, Cambridge 1958)
- 2.10 N.R. Hanson: *Perception and Discovery: An Introduction to Scientific Inquiry* (Freeman, San Francisco 1969)
- 2.11 J. Fodor: *The Modularity of Mind* (MIT, Cambridge 1983)
- 2.12 J. Fodor: Observation reconsidered, *Philos. Sci.* **51**, 23–43 (1984)
- 2.13 J. Fodor: The modularity of mind. In: *Meaning and Cognitive Structure*, ed. by Z. Pylyshyn, W. Demopoulos (Ablex, Norwood 1986)
- 2.14 Z. Pylyshyn: Is vision continuous with cognition?, *Behav. Brain Sci.* **22**, 341–365 (1999)
- 2.15 Z. Pylyshyn: *Seeing and Visualizing: It's Not What You Think* (MIT, Cambridge 2003)
- 2.16 A. Raftopoulos: Is perception informationally encapsulated?, *The issue of the theory-ladenness of perception*, *Cogn. Sci.* **25**, 423–451 (2001)
- 2.17 A. Raftopoulos: Reentrant pathways and the theory-ladenness of observation, *Phil. Sci.* **68**, 187–200 (2001)
- 2.18 A. Raftopoulos: *Cognition and Perception* (MIT, Cambridge 2009)
- 2.19 R. Carnap: Meaning postulates, *Philos. Stud.* **3**(5), 65–73 (1952)
- 2.20 W.V. Quine: Two dogmas of empiricism. In: *From a Logical Point of View*, (Harvard Univ. Press, Massachusetts 1980) pp. 20–46
- 2.21 M.G. White: The analytic and the synthetic: An untenable dualism. In: *Semantics and the Philosophy of Language*, ed. by L. Linsky (Univ. Illinois Press, Urbana 1952) pp. 272–286
- 2.22 P. Achinstein: Theoretical terms and partial interpretation, *Br. J. Philos. Sci.* **14**, 89–105 (1963)
- 2.23 R. Carnap: Testability and meaning, *Philos. Sci.* **3**, 420–468 (1936)
- 2.24 R. Carnap: Testability and meaning, *Philos. Sci.* **4**, 1–40 (1937)
- 2.25 C. Hempel: *Fundamentals of Concept Formation in Empirical Science* (Univ. Chicago Press, Chicago 1952)
- 2.26 K.F. Schaffner: Correspondence rules, *Philos. Sci.* **36**, 280–290 (1969)
- 2.27 P. Suppes: Models of data. In: *Logic, Methodology and Philosophy of Science*, ed. by E. Nagel, P. Suppes, A. Tarski (Stanford Univ. Press, Stanford 1962) pp. 252–261
- 2.28 P. Suppes: What is a scientific theory? In: *Philosophy of Science Today*, ed. by S. Morgenbesser (Basic Books, New York 1967) pp. 55–67
- 2.29 F. Suppe: What's wrong with the received view on the structure of scientific theories?, *Philos. Sci.* **39**, 1–19 (1972)
- 2.30 F. Suppe: *The Semantic Conception of Theories and Scientific Realism* (Univ. Illinois Press, Urbana 1989)
- 2.31 C. Hempel: Provisos: A problem concerning the inferential function of scientific theories. In: *The Limitations of Deductivism*, ed. by A. Grünbaum, W.C. Salmon (Univ. California Press, Berkeley 1988) pp. 19–36
- 2.32 M. Lange: Natural laws and the problem of provisos, *Erkenntnis* **38**, 233–248 (1993)
- 2.33 M. Lange: Who's afraid of ceteris paribus laws?, or: How I learned to stop worrying and love them, *Erkenntnis* **57**, 407–423 (2002)
- 2.34 J. Earman, J. Roberts: Ceteris paribus, there is no problem of provisos, *Synthese* **118**, 439–478 (1999)
- 2.35 J. Earman, J. Roberts, S. Smith: Ceteris paribus lost, *Erkenntnis* **57**, 281–301 (2002)
- 2.36 P.K. Feyerabend: Problems of empiricism. In: *Beyond the Edge of Certainty*, ed. by R.G. Colodny (Prentice-Hall, New Jersey 1965) pp. 145–260
- 2.37 E. Nagel: *The Structure of Science* (Hackett Publishing, Indianapolis 1979)
- 2.38 P.K. Feyerabend: Explanation, reduction and empiricism. In: *Minnesota Studies in the Philosophy of Science: Scientific Explanation, Space and Time*, Vol. 3, ed. by H. Feigl, G. Maxwell (Univ. Minnesota Press, Minneapolis 1962) pp. 28–97
- 2.39 P.K. Feyerabend: How to be a good empiricist – A plea for tolerance in matters epistemological. In: *Philosophy of Science: The Delaware Seminar*, Vol. 2, ed. by B. Baumrin (Interscience, New York 1963) pp. 3–39
- 2.40 P.K. Feyerabend: Problems of empiricism, Part II. In: *The Nature and Function of Scientific Theories*, ed. by R.G. Colodny (Univ. Pittsburgh Press, Pittsburgh 1970) pp. 275–353
- 2.41 P. Suppes: *Introduction to Logic* (Van Nostrand, New York 1957)

- 2.42 P. Suppes: A Comparison of the meaning and uses of models in mathematics and the empirical sciences. In: *The Concept and the Role of the Model in Mathematics and the Natural and Social Sciences*, ed. by H. Freudenthal (Reidel, Dordrecht 1961) pp. 163–177
- 2.43 P. Suppes: *Set-Theoretical Structures in Science* (Stanford Univ., Stanford 1967), mimeographed lecture notes
- 2.44 P. Suppes: *Representation and Invariance of Scientific Structures* (CSLI Publications, Stanford 2002)
- 2.45 N.C.A. Da Costa, S. French: The model-theoretic approach in the philosophy of science, *Philos. Sci.* **57**, 248–265 (1990)
- 2.46 N.C.A. Da Costa, S. French: *Science and Partial Truth, a Unitary Approach to Models and Scientific Reasoning* (Oxford Univ. Press, Oxford 2003)
- 2.47 B.C. Van Fraassen: *The Scientific Image* (Oxford Univ. Press, Oxford 1980)
- 2.48 B.C. Van Fraassen: On the extension of beth's semantics of physical theories, *Philos. Sci.* **37**, 325–339 (1970)
- 2.49 M. Friedman: Review of Bas C. van Fraassen: The scientific image, *J. Philos.* **79**, 274–283 (1982)
- 2.50 J. Worrall: Review article: An unreal image, *Br. J. Philos. Sci.* **35**, 65–80 (1984)
- 2.51 B.C. Van Fraassen: *An Introduction to the Philosophy of Time and Space*, 2nd edn. (Columbia Univ. Press, New York 1985)
- 2.52 B.C. Van Fraassen: The semantic approach to scientific theories. In: *The Process of Science*, ed. by N.J. Nersessian (Martinus Nijhoff, Dordrecht 1987) pp. 105–124
- 2.53 M. Thompson-Jones: Models and the semantic view, *Philos. Sci.* **73**, 524–535 (2006)
- 2.54 B.C. Van Fraassen: *Laws and Symmetry* (Oxford Univ. Press, Oxford 1989)
- 2.55 R.N. Giere: *Explaining Science: A Cognitive Approach* (The Univ. Chicago Press, Chicago 1988)
- 2.56 S. French: The structure of theories. In: *The Routledge Companion to the Philosophy of Science*, ed. by S. Psillos, M. Curd (Routledge, London 2008) pp. 269–280
- 2.57 M. Suarez: Scientific representation: Against similarity and isomorphism, *Int. Stud. Philos. Sci.* **17**(3), 225–244 (2003)
- 2.58 S.M. Downes: The importance of models in theorising: A deflationary semantic view, *PSA 1992*, Vol. 1, ed. by D. Hull, M. Forbes, K. Okruhlik (Philosophy of Science Association, Chicago 1992) pp. 142–153
- 2.59 D. Portides: Scientific models and the semantic view of scientific theories, *Philos. Sci.* **72**(5), 1287–1298 (2005)
- 2.60 R. Frigg: Scientific representation and the semantic view of theories, *Theoria* **55**, 49–65 (2006)
- 2.61 N.D. Cartwright: *The Dappled World: A Study of the Boundaries of Science* (Cambridge Univ. Press, Cambridge 1999)
- 2.62 M.C. Morrison: Models as autonomous agents. In: *Models as Mediators*, ed. by M.S. Morgan, M. Morrison (Cambridge Univ. Press, Cambridge 1999) pp. 38–65
- 2.63 M.C. Morrison: Modelling nature: Between physics and the physical world, *Philos. Naturalis* **35**, 65–85 (1998)
- 2.64 M.C. Morrison: Where have all the theories gone?, *Philos. Sci.* **74**, 195–228 (2007)
- 2.65 D. Portides: Models. In: *The Routledge Companion to the Philosophy of Science*, ed. by S. Psillos, M. Curd (Routledge, London 2008) pp. 385–395
- 2.66 N.D. Cartwright, T. Shomar, M. Suarez: The tool-box of science. In: *Theories and Models In Scientific Processes*, Poznan Studies, Vol. 44, ed. by E. Herfel, W. Krajewski, I. Niiniluoto, R. Wojcicki (Rodopi, Amsterdam 1995) pp. 137–149
- 2.67 D. Portides: Seeking representations of phenomena: Phenomenological models, *Stud. Hist. Philos. Sci.* **42**, 334–341 (2011)
- 2.68 N.D. Cartwright: Models and the limits of theory: Quantum hamiltonians and the BCS models of superconductivity. In: *Models as Mediators*, ed. by M.S. Morgan, M. Morrison (Cambridge Univ. Press, Cambridge 1999) pp. 241–281
- 2.69 M.S. Morgan, M. Morrison (Eds.): *Models as Mediators: Perspectives on Natural and Social Science* (Cambridge Univ. Press, Cambridge 1999)
- 2.70 M. Suarez: The role of models in the application of scientific theories: Epistemological implications. In: *Models as Mediators: Perspectives on Natural and Social Science*, ed. by M.S. Morgan, M. Morrison (Cambridge Univ. Press, Cambridge 1999) pp. 168–196
- 2.71 S. Hartman: Models and stories in hadron physics. In: *Models as Mediators: Perspectives on Natural and Social Science*, ed. by M.S. Morgan, M. Morrison (Cambridge Univ. Press, Cambridge 1999) pp. 326–346
- 2.72 R. Giere: *Science Without Laws* (Univ. Chicago Press, Chicago 1999)
- 2.73 R. Giere: *Scientific Perspectivism* (Univ. Chicago Press, Chicago 2006)

# Models and Representation

## 3. Models and Representation

Roman Frigg, James Nguyen

Models are of central importance in many scientific contexts. We study models and thereby discover features of the phenomena they stand for. For this to be possible models must be representations: they can instruct us about the nature of reality only if they represent the selected parts or aspects of the world we investigate. This raises an important question: In virtue of what do scientific models represent their target systems? In this chapter we first disentangle five separate questions associated with scientific representation and offer five conditions of adequacy that any successful answer to these questions must meet. We then review the main contemporary accounts of scientific representation – similarity, isomorphism, inferentialist, and fictionalist accounts – through the lens of these questions. We discuss each of their attributes and highlight the problems they face. We finally outline our own preferred account, and suggest that it provides the most promising way of addressing the questions raised at the beginning of the chapter.

3.1	<b>Problems Concerning Model-Representation</b> .....	51
3.2	<b>General Griceanism and Stipulative Fiat</b> .....	55
3.3	<b>The Similarity Conception</b> .....	57
3.3.1	Similarity and ER-Problem .....	58
3.3.2	Accuracy and Style .....	62
3.3.3	Problems of Ontology .....	64
3.4	<b>The Structuralist Conception</b> .....	66
3.4.1	Structures and the Problem of Ontology ..	66
3.4.2	Structuralism and the ER-Problem .....	68
3.4.3	Accuracy, Style and Demarcation .....	70
3.4.4	The Structure of Target Systems .....	71
3.5	<b>The Inferential Conception</b> .....	76
3.5.1	Deflationary Inferentialism .....	76
3.5.2	Inflating Inferentialism: Interpretation ...	80
3.5.3	The Denotation, Demonstration, and Interpretation Account .....	82
3.6	<b>The Fiction View of Models</b> .....	83
3.6.1	Models and Fiction .....	84
3.6.2	Direct Representation .....	86
3.6.3	Parables and Fables .....	88
3.6.4	Against Fiction .....	89
3.7	<b>Representation-as</b> .....	91
3.7.1	Exemplification and Representation-as ..	91
3.7.2	From Pictures to Models: The Denotation, Exemplification, Keying-up and Imputation Account .....	93
3.8	<b>Envoi</b> .....	96
	<b>References</b> .....	96

Models play a central role in contemporary science. Scientists construct models of atoms, elementary particles, polymers, populations, genetic trees, economies, rational decisions, airplanes, earthquakes, forest fires, irrigation systems, and the world's climate – there is hardly a domain of inquiry without models. Models are essential for the acquisition and organization of scientific knowledge. We often study a model to discover features of the thing it stands for. How does this work? The answer is that a model can instruct us about the

nature of its subject matter if it represents the selected part or aspect of the world that we investigate. So if we want to understand how models allow us to learn about the world, we have to come to understand how they represent.

The problem of representation has generated a sizeable literature, which has been growing fast in particular over the last decade. The aim of this chapter is to review this body of work and assess the strengths and weaknesses of the different proposals. This enterprise

faces an immediate difficulty: Even a cursory look at the literature on scientific representation quickly reveals that there is no such thing as *the* problem of scientific representation. In fact, we find a cluster of interrelated problems. In Sect. 3.1 we try to untangle this web and get clear on what the problems are and on how they relate to one another (for a historical introduction to the issue, see [3.1]). The result of this effort is a list with five problems and five conditions of adequacy, which provides the analytical lens through which we look at the different accounts. In Sect. 3.2 we discuss Griceanism and *stipulative fiat*. In Sect. 3.3 we look at the time-honored similarity approach, and in Sect. 3.4 we examine its modern-day cousin, the structuralist approach. In Sect. 3.5 we turn to inferentialism, a more recent family of conceptions. In Sect. 3.6 we discuss the fiction view of models, and in Sect. 3.7 we consider the conception of representation-as.

Before delving into the discussion, a number of caveats are in order. The first is that our discussion in no way presupposes that models are the sole unit of scientific representation, or that all scientific representation is model-based. Various types of images have their place in science, and so do graphs, diagrams, and drawings (*Perini* [3.2–4] and *Elkins* [3.5] provide discussions of visual representation in the sciences). In some contexts scientists use what *Warmbröd* [3.6] calls *natural forms of representation* and what *Peirce* [3.7] would have classified as indices: tree rings, fingerprints, disease symptoms. These are related to thermometer readings and litmus paper indications, which are commonly classified as measurements. Measurements also provide representations of processes in nature, sometimes together with the subsequent condensation of measurement results in the form of charts, curves, tables and the like (*Tal* [3.8] provides a discussion of measurement). And, last but not least, many would hold that theories represent too. At this point the vexing problem of the nature of theories and the relation between theories and models rears its head again. We refer the reader to *Portides'* contribution to this volume, Chap. 2, for a discussion of this issue. Whether these other forms of scientific representation have features in common with how models represent is an interesting question, but this is a problem for another day. Our aim here is a more modest one: to understand how models represent. To make the scope of our investigation explicit we call the kind of representation we are interested in *model-representation*.

The second point to emphasize is that our discussion is not premised on the claim that *all* models are representational; nor does it assume that representation is the only (or even primary) function of models. It has been emphasized variously that models

perform a number of functions other than representation. To mention but few: *Knuuttila* [3.9, 10] points out that the epistemic value of models is not limited to their representational function and develops an account that views models as epistemic artifacts that allow us to gather knowledge in diverse ways; *Morgan* and *Morrison* [3.11] emphasize the role models play in the mediation between theories and the world; *Hartmann* [3.12] discusses models as tools for theory construction; *Peschard* [3.13] investigates the way in which models may be used to construct other models and generate new target systems; and *Bokulich* [3.14] and *Kennedy* [3.15] present nonrepresentational accounts of model explanation (*Woody* [3.16] and *Reiss* [3.17] provide general discussions of the relation between representation and explanation). Not only do we not see projects like these as being in conflict with a view that sees some models as representational; we think that the approaches are in fact complementary.

Finally, there is a popular myth according to which a representation is a mirror image, a copy, or an imitation of the thing it represents. In this view representation is ipso facto realistic representation. This is a mistake. Representations can be realistic, but they need not. And representations certainly need not be copies of the real thing. This, we take it, is the moral of the satire about the cartographers who produce maps as large as the country itself only to see them abandoned. The story has been told by Lewis Carroll in *Sylvie and Bruno* and Jorge Luis Borges in *On Exactitude in Science*. Throughout this review we encounter positions that make room for nonrealistic representation and hence testify to the fact that representation is a much broader notion than mirroring.

There is, however, a sense in which we presuppose a minimal form of realism. Throughout the discussion we assume that target systems exist independently of human observers, and that they are how they are irrespective of what anybody thinks about them. That is, we assume that the targets of representation exist independently of the representation. This is a presupposition not everybody would share. Constructivists (and other kinds of metaphysical antirealists) assume that there is no phenomenon independent of its representation: representations constitute the phenomena they represent (this view is expounded for instance by *Lynch* and *Wooglar* [3.18]; *Giere* [3.19] offers a critical discussion). It goes without saying that an assessment of the constructivist program is beyond the scope of this review. It is worth observing, though, that many of the discussions to follow are by no means pointless from a constructivist perspective. What in the realist idiom is conceptualized as the representation of an object in the world by a model would, from the constructivist

perspective, turn into the study of the relation between a model and another representation, or an object constituted by another representation. This is because even from a constructivist perspective, models and their tar-

gets are not identical, and the fact that targets are representationally constituted would not obliterate the differences between a target representation and scientific model.

### 3.1 Problems Concerning Model-Representation

In this section we say what questions a philosophical account of model-representation has to answer and reflect on what conditions such an answer has to satisfy. As one would expect, different authors have framed the problem in different ways. Nevertheless, recent discussions about model-representation have tended to cluster around a relatively well-circumscribed set of issues. The aim of this section is to make these issues explicit and formulate five problems that an account of model-representation has to answer. These problems will help us in structuring the discussion in later sections and put views and positions into perspective. In the course of doing so we also articulate five conditions of adequacy that every account of model-representation has to satisfy.

Models are representations of a selected part or aspect of the world. This is the model's *target system*. The first and most fundamental question about a model therefore is: In virtue of what is a model a representation of something else? Attention has been drawn to this issue by Frigg ([3.20, p. 17], [3.21, p. 50]), Morrison [3.22, p. 70], and Suárez [3.23, p. 230]. To appreciate the thrust of this question it is instructive to briefly ponder the same problem in the context of pictorial representation. When seeing, say, Soutine's *The Groom or the Bellboy* we immediately realize that it depicts a man in a red dress. Why is this? Per se the painting is a plane surface covered with pigments. How does an arrangement of pigments on a surface represent something outside the picture frame? Likewise, models, before being representations of atoms, populations, or economies, are equations, structures, fictional scenarios, or mannerly physical objects. The problem is: what turns equations and structures, or fictional scenarios and physical objects into representations of something beyond themselves? It has become customary to phrase this problem in terms of necessary and sufficient conditions and throughout this review we shall follow suit (some may balk at this, but it's worth flagging that the standard arguments against such an analysis, e.g., those surveyed in Laurence and Margolis [3.24], lose much of their bite when attention is restricted to core cases as we do here). The question then is: What fills the blank in *M is a model-representation of T iff \_\_\_\_\_*, where *M* stands for model and *T* for target system?

To spare ourselves difficulties further down the line, this formulation needs to be adjusted in light of

a crucial condition of adequacy that any account of model-representation has to meet. The condition is that models represent in a way that allows us to form hypotheses about their target systems. We can generate claims about a target system by investigating a model that represents it. Many investigations are carried out on models rather than on reality itself, and this is done with the aim of discovering features of the things models stands for. Every acceptable theory of scientific representation has to account for how reasoning conducted on models can yield claims about their target systems. Let us call this the *surrogate reasoning condition*.

The term *surrogate reasoning* was introduced by Swoyer [3.25, p. 449], and there seems to be widespread agreement on this point (although Callender and Cohen [3.26], whose views are discussed in Sect. 3.3, provide a noteworthy exception). To mention just a few writers on the subject: Bailer-Jones [3.27, p. 59] emphasizes that models “*tell us* something about certain features of the world” (original emphasis). Boliskna [3.28] and Contessa [3.29] both call models *epistemic representations*; Frigg ([3.21, p. 51], [3.30, p. 104]) sees the potential for learning as an essential explanandum for any theory of representation; Liu [3.31, p. 93] emphasizes that the main role for models in science and technology is epistemic; Morgan and Morrison [3.11, p. 11] regard models as *investigative tools*; Suárez ([3.23, p. 229], [3.32, p. 772]) submits that models license specific inferences about their targets; and Weisberg [3.33, p. 150] observes that the “model-world relation is the relationship in virtue of which studying a model can tell us something about the nature of a target system”. This distinguishes models from lexicographical representations such as words. Studying the internal constitution of a model can provide information about the target. Not so with words. The properties of a word (consisting of so and so many letters and syllables, occupying this or that position in a dictionary, etc.) do not matter to its functioning as a word; and neither do the physical properties of the ink used to print words on a piece of paper. We can replace one word by another at will (which is what happens in translations from one language to another), and we can print words with other methods than ink on paper. This is possible because the properties of a word as an object do not matter to its semantic function.

This gives rise to a problem for the schema *M is a model-representation of T iff \_\_\_\_*. The problem is that any account of representation that fills the blank in a way that satisfies the *surrogative reasoning condition* will almost invariably end up covering other kinds of representations too. Geographical maps, graphs, diagrams, charts, drawings, pictures, and photographs often provide epistemic access to features of the items they represent, and hence are likely to fall under an account of representation that explains this sort of reasoning. This is a problem for an analysis of model-representation in terms of necessary and sufficient conditions because if something that is not *prima facie* a model (for instance a map or a photograph) satisfies the conditions of an account of model-representation, then one either has to conclude that the account fails because it does not provide necessary conditions, or that first impressions are wrong and other representations (such as maps or photographs) are in fact model-representations.

Neither of these options is appealing. To avoid this problem we follow a suggestion of *Contessa's* [3.29] and broaden the scope of the investigation. Rather than analyzing the relatively narrow category of model-representation, we analyze the broader category of *epistemic representation*. This category comprises model-representations, but it also includes other representations that allow for surrogative reasoning. The task then becomes to fill the blank in *M is an epistemic representation of T iff \_\_\_\_*. For brevity we use  $R(M, T)$  as a stand in for *M is an epistemic representation of T*, and so the biconditional becomes  $R(M, T) \text{ iff } \_\_\_\_\_\_$ . We call the general problem of figuring out in virtue of what something is an epistemic representation of something else the *epistemic representation problem (ER-problem, for short)*, and the above biconditional the *ER-scheme*. So one can say that the ER is to fill the blank in the ER-scheme. *Frigg* [3.21, p. 50] calls this the “enigma of representation” and in *Suárez's* [3.23, p. 230] terminology this amounts to identifying the *constituents* of a representation (although he questions whether both necessary *and* sufficient conditions can be given; see Sect. 3.5 for further discussion on how his views fit into the ER-framework).

Analyzing the larger category of epistemic representation and placing model-representations in that category can be seen as giving rise to a demarcation problem for scientific representations: How do scientific model-representations differ from other kinds of epistemic representations? We refer to this question as the *representational demarcation problem*. *Callender and Cohen* [3.26, p. 69] formulate this problem, but then voice skepticism about our ability to solve it [3.26, p. 83]. The representational demarcation problem has

received little, if any, attention in the recent literature on scientific representation, which would suggest that other authors either share Callender and Cohen's skepticism, or regard it as a nonissue to begin with. The latter seems to be implicit in approaches that discuss scientific representation alongside pictorial representation such as *Elgin* [3.34], *French* [3.35], *Frigg* [3.21], *Suárez* [3.32], and *van Fraassen* [3.36]. But a dismissal of the problem is in no way a neutral stance. It amounts to no less than the admission that model-representations are not fundamentally different from other epistemic representations, or that we are unable to pin down what the distinguishing features are. Such a stance should be made explicit and, ideally, justified.

Two qualifications concerning the ER-scheme need to be added. The first concerns its flexibility. Some might worry that posing the problem in this way pre-judges what answers can be given. The worry comes in a number of variants. A first variant is that the scheme presupposes that representation is an intrinsic relation between *M* and *T* (i. e., a relation that only depends on intrinsic properties of *M* and *T* and on how they relate to one another rather than on how they relate to other objects) or even that it is naturalisable (a notion further discussed in Sect. 3.3). This is not so. In fact, *R* might depend on any number of factors other than *M* and *T* themselves, and on ones that do not qualify as natural ones. To make this explicit we write the ER-scheme in the form  $R(M, T) \text{ iff } C(M, T, x_1, \dots, x_n)$ , where *n* is a natural number and *C* is an (*n* + 2)-ary relation that grounds representation. The *x<sub>i</sub>* can be anything that is deemed relevant to epistemic representation, for instance a user's intentions, standards of accuracy, and specific purposes. We call *C* the *grounding relation* of an epistemic representation.

Before adding a second qualification, let us introduce the next problem in connection with model-representation. Even if we restrict our attention to scientific epistemic representations (if they are found to be relevantly different to nonscientific epistemic representations as per the demarcation problem above), not all representations are of the same kind. In the case of visual representations this is so obvious that it hardly needs mention: An Egyptian mural, a two-point perspective ink drawing, a pointillist oil painting, an architectural plan, and a road map represent their respective targets in different ways. This pluralism is not limited to visual representations. Model-representations do not all seem to be of the same kind either. *Woody* [3.37] argues that chemistry as a discipline has its own ways to represent molecules. But differences in style can also appear in models from the same discipline. *Weizsäcker's* liquid drop model represents the nucleus of an atom in a manner that seems to be different from the one of the shell

model. A scale model of the wing of a plane represents the wing in a way that is different from how a mathematical model of its cross section does. Or Phillips and Newlyn's famous hydraulic machine and Hicks' mathematical models both represent a Keynesian economy but they seem to do so in different ways. This gives rise to the question: What styles are there and how can they be characterized? This is the *problem of style* [3.21, p. 50]. There is no expectation that a *complete* list of styles be provided in response. Indeed, it is unlikely that such a list can ever be drawn up, and new styles will be invented as science progresses. For this reason a response to the problem of style will always be open-ended, providing a taxonomy of what is currently available while leaving room for later additions.

With this in mind we can now turn to the second qualification concerning the ER-scheme. The worry is this: The scheme seems to assume that representation is a monolithic concept and thereby make it impossible to distinguish between different kinds of representation. The impression is engendered by the fact the scheme asks us to fill a blank, and blank is filled only once. But if there are different kinds of representations, we should be able to fill the blank in different ways on different occasions because a theory of representation should not force upon us the view that the different styles are all variations of one overarching concept of representation.

The ER-scheme is more flexible than it appears at first sight. There are at least three ways in which different styles of representations can be accommodated. For the sake of illustration, and to add some palpability to an abstract discussion, let us assume that we have identified two styles: analogue representation and idealized representation. The result of an analysis of these relations is the identification of their respective grounding relations. Let  $C_A(M, T, \dots)$  and  $C_1(M, T, \dots)$  be these relations. The first way of accommodating them in the ER-scheme is to fill the blank with the disjunction of the two:  $R(M, T) \text{ iff } C_A(M, T, \dots) \text{ or } C_1(M, T, \dots)$ . In plain English:  $M$  represents  $T$  if and only if  $M$  is an analogue representation of  $T$  or  $M$  is an idealized representation of  $T$ . This move is possible because, first appearances notwithstanding, nothing hangs on the grounding relation being homogeneous. The relation can be as complicated as we like and there is no prohibition against disjunctions. In the above case we have  $C = [C_A \text{ or } C_1]$ . Furthermore, the grounding relation could even be an open disjunction. This would help accommodating the above observation that a list of styles is potentially open-ended. In that case there would be a grounding relation for each style and the scheme could be written as  $R(M, T) \text{ iff } C_1(M, T, \dots) \text{ or } C_2(M, T, \dots) \text{ or } C_3(M, T, \dots) \text{ or } \dots$ , where the  $C_i$  are the grounding relations for different styles. This

is not a new scheme; it's the old scheme where  $C = [C_1 \text{ or } C_2 \text{ or } C_3 \text{ or } \dots]$  is spelled out.

Alternatively one could formulate a different scheme for every kind of representation. This would amount to changing the scheme slightly in that one does not analyze epistemic representation per se. Instead one would analyze different kinds of epistemic representations. Consider the above example again. Let  $R_1(M, T)$  stand for  $M$  is an analogue epistemic representation of  $T$  and  $R_2(M, T)$  for  $M$  is an idealized epistemic representation of  $T$ . The response to the ER-problem then consists in presenting the two biconditionals  $R_1(M, T) \text{ iff } C_A$  and  $R_2(M, T) \text{ iff } C_1$ . This generalizes straightforwardly to the case of any number of styles, and the open-endedness of the list of styles can be reflected in the fact that an open-ended list of conditionals of the form  $R_i(M, T) \text{ iff } C_i$  can be given, where the index ranges over styles.

In contrast with the second option, which pulls in the direction of more diversity, the third aims for more unity. The crucial observation here is that the grounding relation can in principle be an abstract relation that can be concretized in different ways, or a determinable that can have different determinates. On the third view, then, the concept of representation is like the concept of force (which is abstract in that in a concrete situation force is gravity or electromagnetic attraction or some other specific force), or like color (where a colored object must be blue or green or \_\_\_\_). This view would leave  $R(M, T) \text{ iff } C(M, T, x_1, \dots, x_n)$  unchanged and take it as understood that  $C$  is an abstract relation.

At this point we do not adjudicate between these options. Each has its own pros and cons, and which one is the most convenient to work with depends on one's other philosophical commitments. What matters is that the ER-scheme does have the flexibility to accommodate different representational styles, and that it can in fact accommodate them in at least three different ways.

The next problem in line for the theory of model-representation is to specify standards of accuracy. Some representations are accurate; others aren't. The Schrödinger model is an accurate representation of the hydrogen atom; the Thomson model isn't. On what grounds do we make such judgments? In Morrison's words: "how do we identify what constitutes an accurate representation?" [3.22, p. 70]. We call this the problem of *standards of accuracy*. Answering this question might make reference to the purposes of the model and model user, and thus it is important to note that by *accuracy* we mean something that can come in degrees and may be context dependent. Providing a response to the problem of accuracy is a crucial aspect of an account of epistemic representation.

This problem goes hand in hand with a second condition of adequacy: the *possibility of misrepresentation*. Asking what makes a representation an accurate representation already presupposes that inaccurate representations are representations too. And this is how it should be. If  $M$  does not accurately portray  $T$ , then it is a misrepresentation but not a nonrepresentation. It is therefore a general constraint on a theory of epistemic representation that it has to make misrepresentation possible. This can be motivated by a brief glance at the history of science, but is plausibly also part of the concept of representation, and as such is found in discussions of other kinds of representation (*Stitch and Warfield* [3.38, pp. 6–7], for instance, suggest that a theory of mental representation should be able to account for misrepresentation, as do *Sterelny and Griffiths* [3.39, p. 104] in their discussion of genetic representation). A corollary of this requirement is that representation is a wider concept than accurate representation and that representation cannot be analyzed in terms of accurate representation.

A related condition concerns models that misrepresent in the sense that they lack target systems. Models of ether, phlogiston, four-sex populations, and so on, are all deemed scientific models, but ether, phlogiston, and four-sex populations don't exist. Such models lack (actual) target systems, and one hopes that an account of epistemic representation would allow us to understand how these models work. We call this the problem of targetless models (or models without targets).

The fourth condition of adequacy for an account of model-representation is that it must account for the directionality of representation. Models are about their targets, but (at least in general) targets are not about their models. So there is an essential directionality to representations, and an account of model-representation has to identify the root of this directionality. We call this the *requirement of directionality*.

Many scientific models are highly mathematized, and their mathematical aspects are crucial to their cognitive as well as their representational function. This forces us to reconsider a time-honored philosophical puzzle: the applicability of mathematics in the empirical sciences. Even though the problem can be traced back at least to Plato's *Timaeus*, its canonical modern expression is due to *Wigner*, who famously remarked that “the enormous usefulness of mathematics in the natural sciences is something bordering on the mysterious and that there is no explanation for it” [3.40, p. 2]. One need not go as far as seeing the applicability of mathematics as an inexplicable miracle, but the question remains: How does mathematics hook onto the world?

The recent discussion of this problem has taken place in a body of literature that grew out of the philos-

ophy of mathematics (see *Shapiro* [3.41, Chap. 8] for a review). But, with the exception of *Bueno and Colyvan* [3.42], there has been little contact with the literature on scientific modeling. This is a regrettable state of affairs. The question of how a mathematized model represents its target implies the question of how mathematics applies to a physical system. So rather than separating the question of model-representation from the problem of the applicability of mathematics and dealing with them in separate discussions, they should be seen as the two sides of the same coin and be dealt with in tandem. For this reason, our fifth and final condition of adequacy is that an account of representation has to explain how mathematics is applied to the physical world. We call this the *applicability of mathematics condition*.

In answering the above questions one invariably runs up against a further problem, the *problem of ontology*: What kinds of objects are models? Are they structures in the sense of set theory, fictional entities, descriptions, equations or yet something else? Or are there no models at all? While some authors develop an ontology of models, others reject an understanding of models as *things* and push a program that can be summed up in the slogan *modeling without models* [3.43]. There is also no presupposition that all models be of the same kind. Some models are material objects, some are things that one holds in one's head rather than one's hands (to use *Hacking's* phrase [3.44, p. 216]). For the most part, the focus in debates about representation has been on nonmaterial models, and we will follow this convention. It is worth emphasizing, however, that also the seemingly straightforward material models raise interesting philosophical questions: *Rosenblueth and Wiener* [3.45] discuss the criteria for choosing an object as a model; *Ankeny and Leonelli* [3.46] discuss issues that arise when using organisms as models; and the contributors to [3.47] discuss representation in the laboratory.

A theory of representation can recognize different kinds of models, or indeed no models at all. The requirement only asks us to be clear on our commitments and provide a list with things, if any, that we recognize as models and give an account of what they are in case these entities raise questions (what exactly do we mean by something that one holds in one's head rather than one's hands?).

In sum, an account of model-representation has to do the following:

1. Provide an answer to the *epistemic representation problem* (filling the blank in ER-scheme:  $M$  is an epistemic representation of  $T$  iff . . .).
2. Take a stand on the *representational demarcation problem* (the question of how scientific epistemic



representations differ from other kinds of epistemic representations).

3. Respond to the *problem of style* (what styles are there and how can they be characterized?).
4. Formulate *standards of accuracy* (how do we identify what constitutes an accurate representation?).
5. Address the *problem of ontology* (what kinds of objects are models?).

Any satisfactory answer to these five issues will have to meet the following five conditions of adequacy:

1. *Surrogative reasoning condition* (models represent their targets in a way that allows us to generate hypotheses about them).
2. *Possibility of misrepresentation* (if  $M$  does not accurately represent  $T$ , then it is a misrepresentation but not a nonrepresentation).

3. *Targetless models* (what are we to make of scientific representations that lack targets?).
4. *Requirement of directionality* (models are about their targets, but targets are not about their models).
5. *Applicability of mathematics condition* (how the mathematical apparatus used in  $M$  latches onto the physical world).

To frame the problem in this way is not to say that these are separate and unrelated issues, which can be dealt with one after the other in roughly the same way in which we first buy a ticket, walk to the platform and then take a train. This division is analytical, not factual. It serves to structure the discussion and to assess proposals; it does not imply that an answer to one of these questions can be dissociated from what stance we take on the other issues.

## 3.2 General Griceanism and Stipulative Fiat

*Callender* and *Cohen* [3.26] submit that the entire debate over scientific representation has started on the wrong foot. They claim that scientific representation is not different from “artistic, linguistic, and culinary representation” and in fact “there is no special problem about scientific representation” [3.26, p. 67]. Underlying this claim is a position *Callender* and *Cohen* call *General Griceanism* (GG). The core of GG is the reductive claim that most representations we encounter are “derivative from the representational status of a privileged core of representations” [3.26, p. 70]. GG then comes with a practical prescription about how to proceed with the analysis of a representation [3.26, p. 73]:

“The General Gricean view consists of two stages. First, it explains the representational powers of derivative representations in terms of those of fundamental representations; second, it offers some other story to explain representation for the fundamental bearers of content.”

Of these stages only the second requires serious philosophical work, and this work is done in the philosophy of mind because the fundamental form of representation is mental representation.

Scientific representation is a derivative kind of representation [3.26, pp. 71,75] and hence falls under the first stage of the above recipe. It is reduced to mental representation by an act of stipulation [3.26, pp. 73–74]:

“Can the salt shaker on the dinner table represent Madagascar? Of course it can, so long as you stipulate that the former represents the latter. [...] Can your left hand represent the Platonic form of

beauty? Of course, so long as you stipulate that the former represents the latter. [...] On the story we are telling, then, virtually anything can be stipulated to be a representational vehicle for the representation of virtually anything [...]; the representational powers of mental states are so wide-ranging that they can bring about other representational relations between arbitrary relata by dint of mere stipulation. The upshot is that, once one has paid the admittedly hefty one-time fee of supplying a metaphysics of representation for mental states, further instances of representation become extremely cheap.”

So explaining any form of representation other than mental representation is a triviality – all it takes is an act of “stipulative fiat” [3.26, p. 75]. This supplies their answer to the ER-problem:

### *Definition 3.1 Stipulative fiat*

A scientific model  $M$  represents a target system  $T$  iff a model user stipulates that  $M$  represents  $T$ .

On this view, scientific representations are cheap to come by. The question therefore arises why scientists spend a lot of time constructing and studying complex models if they might just as well take a salt shaker and turn it into a representation of, say, a Bose–Einstein condensate by an act of fiat. *Callender* and *Cohen* admit that there are useful and not so useful representations, and that salt shakers belong the latter group. However, they insist that this has nothing to do with representation [3.26, p. 75]:

“The questions about the utility of these representational vehicles are questions about the pragmatics of things that are representational vehicles, not questions about their representational status per se.”

So, in sum, scientific representation [3.26, p. 78]

“is constituted in terms of a stipulation, together with an underlying theory of representation for mental states, isomorphism, similarity, and inference generation are all idle wheels.”

The first question we are faced with when assessing this account is the relation between GG and stipulative fiat (Definition 3.1). Callender and Cohen do not comment on this issue, but that they mention both in the same breath would suggest that they regard them as one and the same doctrine, or at least as the two sides of the same coin. This is not so. Stipulative fiat (Definition 3.1) is just one way of fleshing out GG, which only requires that there be *some* explanation of how derivative representations relate to fundamental representations; GG does not require that this explanation be of a particular kind, much less that it consists of nothing but an act of stipulation ([3.48, pp. 77–78], [3.49, p. 244]). Even if GG is correct, it doesn’t follow that stipulative fiat is a satisfactory answer to the ER-problem. Model-representation can, in principle, be *reduced* to fundamental representation in many different ways (some of which we will encounter later in this chapter). Conversely, the failure of stipulative fiat does not entail that we must reject GG: one can uphold the idea that an appeal to the intentions of model users is a crucial element in an account of scientific representation even if one dismisses stipulative fiat (Definition 3.1).

Let us now examine stipulative fiat (Definition 3.1). Callender and Cohen emphasize that anything can be a representation of anything else [3.26, p. 73]. This is correct. Things that function as models don’t belong to a distinctive ontological category, and it would be a mistake to think that that some objects are, intrinsically, representations and other are not. This point has been made by others too (including Frigg [3.50, p. 99], Giere [3.51, p. 269], Suárez [3.32, p. 773], Swoyer [3.25, p. 452], and Teller [3.52, p. 397]) and, as we shall see, it is a cornerstone of several alternative accounts of representation.

But just because anything can, in principle, be a representation of anything else, it doesn’t follow that a mere act of stipulation suffices to turn *M* into a representation of *T*. Furthermore, it doesn’t follow that an object elevated to the status of a representation by an act of fiat represents its target in a way that can appropriately be characterized as an instance of epistemic

representation. We discuss both concerns in reverse order.

Stipulative fiat (Definition 3.1) fails to meet the surrogative reasoning condition: it fails to provide an account of how claims about Madagascar could be extracted from reasoning about the salt shaker. Even if we admit that stipulative fiat (Definition 3.1) establishes that models denote their targets (and as we will see soon, there is a question about this), denotation is not sufficient for epistemic representation. Both the word *Napoleon* and Jacques-Louis David’s portrait of Napoleon serve to denote the French general. But this does not imply that they represent him in the same way, as noted by Toon [3.48, pp. 78–79]. *Bueno* and *French* [3.53, pp. 871–874] gesture in the same direction when they point to Peirce’s distinction between icon, index and symbol and dismiss Callender and Cohen’s views on grounds that they cannot explain the obvious differences between different kinds of representations.

Supporters of stipulative fiat (Definition 3.1) could try to mitigate the force of this objection in two ways. First, they could appeal to additional facts about the object, as well as its relation to other items, in order to account for surrogative reasoning. For instance, the salt shaker being to the right of the pepper mill might allow us to infer that Madagascar is to the east of Mozambique. Moves of this sort, however, invoke (at least tacitly) a specifiable relation between features of the model and features of the target (similarity, isomorphism, or otherwise), and an invocation of this kind goes beyond mere stipulation. Second, the last quotation from Callender and Cohen suggests that they might want to relegate surrogative reasoning into the realm of pragmatics and deny that it is part of the relation properly called epistemic representation. This, however, in effect amounts to a removal of the surrogative reasoning condition from the desiderata of an account of scientific representation, and we have argued in Sect. 3.1 that surrogative reasoning is one of the hallmarks of scientific representation. And even if it were *pragmatics*, we still would want an account of how it works.

Let us now turn to our first point, that a mere act of stipulation is insufficient to turn *M* into a representation of *T*. We take our cue from a parallel discussion in the philosophy of language, where it has been pointed out that it is not clear that stipulation is sufficient to establish a denotational relationship (which is weaker than epistemic representation). A position similar to stipulative fiat (Definition 3.1) faces what is known as the *Humpty Dumpty problem*, named in reference to Lewis Carroll’s discussion of Humpty using the word *glory* to mean *a nice knockdown argument* [3.54, 55] (it’s worth noting that this debate concerns meaning,

rather than denotation, but it's plausible that it can be reconstructed in terms of the latter). If stipulation is all that matters, then as long as Humpty simply stipulates that *glory* means *a nice knockdown argument*, then it does so. And this doesn't seem to be the case. Even if the utterance *glory* could mean *a nice knockdown argument* – if, for example, Humpty was speaking a different language – in the case in question it doesn't, irrespective of Humpty's stipulation. In the contemporary philosophy of language the discussion of this problem focuses more on the denotation of demonstratives rather than proper names, and work in that field focuses on propping up existing accounts so as to ensure that a speaker's intentions successfully establish the denotation of demonstratives uttered by the speaker [3.56]. Whatever the success of these endeavors, their mere existence shows that successfully establishing denotation requires moving beyond a bare appeal to stipulation, or brute intention. But if a brute appeal to intentions fails in the case of demonstratives – the sorts of terms that such an account would most readily be applicable to – then we find it difficult to see how stipulative fiat (Definition 3.1) will establish a representational relationship between models and their targets. Moreover, this whole discussion supposed that an intention-based account of denotation is the correct one. This is controversial – see Reimer and Michaelson [3.57] for an overview of discussions of denotation in the philosophy of language. If this is not the correct way to think about denotation,

then stipulative fiat (Definition 3.1) will fail to get off the ground at all.

It now pays that we have separated GG from stipulative fiat (Definition 3.1). Even though stipulative fiat (Definition 3.1) does not provide an adequate answer to the ER-problem, one can still uphold GG. As Callender and Cohen note, all that it requires is that there is a privileged class of representations (they take them to be mental states but are open to the suggestion that they might be something else [3.26, p. 82]), and that other types of representations owe their representational capacities to their relationship with the primitive ones. So philosophers need an account of how members of this privileged class of representations represent, and how derivative representations, which includes scientific models, relate to this class.

This is a plausible position, and when stated like this, many recent contributors to the debate on scientific representation can be seen as falling under the umbrella of GG. As we will see below, the more developed versions of the similarity (Sect. 3.3) and isomorphism (Sect. 3.4) accounts of scientific representation make explicit reference to the intentions and purposes of model users, even if their earlier iterations did not. And so do the accounts discussed in the latter sections, where the intentions of model users (in a more complicated manner than that suggested by stipulative fiat (Definition 3.1)) are invoked to establish epistemic representation.

### 3.3 The Similarity Conception

Moving on from the Gricean account we now turn to the similarity conception of scientific representation (in aesthetics the term *resemblance* is used more commonly than *similarity*, but there does not seem to be a substantive difference between the notions, and we use the terms as synonyms throughout). Similarity and representation initially appear to be two closely related concepts, and invoking the former to ground the latter has a philosophical lineage stretching back at least as far as Plato's *The Republic*.

In its most basic guise the similarity conception of scientific representation asserts that scientific models represent their targets in virtue of being similar to them. This conception has universal aspirations in that it is taken to account for epistemic representation across a broad range of different domains. Paintings, statues, and drawings are said to represent by being similar to their subjects, (see Abell [3.58] and Lopes [3.59] for relatively current discussions of similarity in the context of visual representation). And recently Giere, one of the

view's leading contemporary proponents, proclaimed that it covers scientific models alongside “words, equations, diagrams, graphs, photographs, and, increasingly, computer-generated images” [3.60, p. 243] (see also Giere [3.61, p. 272], and for further discussion Toon [3.49, pp. 249–250]). So the similarity view repudiates the demarcation problem and submits that the same mechanism, namely similarity, underpins different kinds of representation in a broad variety of contexts. (Sometimes the similarity view is introduced by categorizing models as icons in Peirce's sense, and, as Kraleman and Lattmann point out, icons represent “on the basis of a similarity relation between themselves and their objects” [3.62, p. 3398].)

The view also offers an elegant account of surrogate reasoning. Similarities between model and target can be exploited to carry over insights gained in the model to the target. If the similarity between  $M$  and  $T$  is based on shared properties, then a property found in  $M$  would also have to be present in  $T$ ; and if the similar-

ity holds between properties themselves, then  $T$  would have to instantiate properties similar to  $M$  (however, it is worth noting that this kind of knowledge transfer can cause difficulties in some contexts, Frigg et al. [3.63] discuss these difficulties in the context of nonlinear dynamic modeling).

However, appeal to similarity in the context of representation leaves open whether similarity is offered as an answer to the ER-problem, the problem of style, or whether it is meant to set standards of accuracy. Proponents of the similarity account typically have offered little guidance on this issue. So we examine each option in turn and ask whether similarity offers a viable answer. We then turn to the question of how the similarity view deals with the problem of ontology.

### 3.3.1 Similarity and ER-Problem

Understood as response to the ER-problem, a similarity view of representation amounts to the following:

#### *Definition 3.2 Similarity 1*

A scientific model  $M$  represents a target  $T$  iff  $M$  and  $T$  are similar.

A well-known objection to this account is that similarity has the wrong logical properties. Goodman [3.64, pp. 4–5] submits that similarity is symmetric and reflexive yet representation isn't. If object  $A$  is similar to object  $B$ , then  $B$  is similar to  $A$ . But if  $A$  represents  $B$ , then  $B$  need not (and in fact in most cases does not) represent  $A$ : the Newtonian model represents the solar system, but the solar system does not represent the Newtonian model. And everything is similar to itself, but most things do not represent themselves. So this account does not meet our third condition of adequacy for an account of scientific representation insofar as it does not provide a direction to representation. (Similar problems also arise in connection with other logical properties, e.g., transitivity; see Frigg [3.30, p. 31] and Suárez [3.23, pp. 232–233].)

Yaghmaie [3.65] argues that this conclusion – along with the third condition itself – is wrong: epistemic representation is symmetric and reflexive (he discusses this in the context of the isomorphism view of representation, which we turn to in the next section, but the point applies here as well). His examples are drawn from mathematical physics, and he presents a detailed case study of a symmetric representation relation between quantum field theory and statistical mechanics. His case raises interesting questions, but even if one grants that Yaghmaie has identified a case where representation is reflexive and symmetrical it does not follow that representation *in general* is. The photograph in

Jane's passport represents Jane; but Jane does not represent her passport photograph; and the same holds true for myriads of other representations. Goodman is correct in pointing out that typically representation is not symmetrical and reflexive: a target  $T$  does not represent model  $M$  just because  $M$  represents  $T$ .

A reply diametrically opposed to Yaghmaie's emerges from the writings of Tversky and Weisberg. They accept that representation is not symmetric, but dispute that similarity fails on this count. Using a gradual notion of similarity (i. e., one that allows for statements like *A is similar to B to degree d*), Tversky found that subjects in empirical studies judged that North Korea was more similar to China than China was to North Korea [3.66]; similarly Poznic [3.67, Sect. 4.2] points out with reference to the characters in a Polanski movie that the similarity relation between a baby and the father need not be symmetric.

So allowing degrees into ones notion of similarity makes room for an asymmetry (although degrees by themselves are not sufficient for asymmetry; metric-based notions are still symmetric). This raises the question of how to analyze similarity. We discuss this thorny issue in some detail in the next subsection. For now we concede the point and grant that similarity need not always be symmetrical. However, this does not solve Goodman's problem with reflexivity (as we will see on Weisberg's notion of similarity everything is maximally similar to itself); nor does it, as will see now, solve other problems of the similarity account.

However the issue of logical properties is resolved, there is another serious problem: similarity is too inclusive a concept to account for representation. In many cases neither one of a pair of similar objects represents the other. Two copies of the same book are similar but neither represents the other. Similarity between two items is not enough to establish the requisite relationship of representation; there are many cases of similarity where no representation is involved. And this won't go away even if similarity turns out to be non-symmetric. That North Korea is similar to China (to some degree) does not imply that North Korea represents China, and that China is not similar to North Korea to the same degree does not alter this conclusion.

This point has been brought home in a now-classical thought experiment due to Putnam [3.68, pp. 1–3] (but see also Black [3.69, p. 104]). An ant is crawling on a patch of sand and leaves a trace that happens to resemble Winston Churchill. Has the ant produced a picture of Churchill? Putnam's answer is that it didn't because the ant has never seen Churchill and had no intention to produce an image of him. Although *someone else* might see the trace as a depiction of Churchill, the trace itself does not represent Churchill. This, Putnam concludes,

shows that “[s]imilarity [...] to the features of Winston Churchill is not sufficient to make something represent or refer to Churchill” [3.68, p. 1]. And what is true of the trace and Churchill is true of every other pair of similar items: similarity on its own does not establish representation.

There is also a more general issue concerning similarity: it is too easy to come by. Without constraints on what counts as similar, any two things can be considered similar to any degree [3.70, p. 21]. This, however, has the unfortunate consequence that anything represents anything else because any two objects are similar in some respect. Similarity is just too inclusive to account for representation. An obvious response to this problem is to delineate a set of relevant respects and degrees to which  $M$  and  $T$  have to be similar. This suggestion has been made explicitly by *Giere* [3.71, p. 81] who suggests that models come equipped with what he calls *theoretical hypotheses*, statements asserting that model and target are similar in relevant respects and to certain degrees. This idea can be molded into the following definition:

#### **Definition 3.3 Similarity 2**

A scientific model  $M$  represents a target  $T$  iff  $M$  and  $T$  are similar in relevant respects and to the relevant degrees.

On this definition one is free to choose one’s respects and degrees so that unwanted similarities drop out of the picture. While this solves the last problem, it leaves the others untouched: similarity in relevant respects and to the relevant degrees is reflexive (and symmetrical, depending on one’s notion of similarity); and presumably the ant’s trace in the sand is still similar to Churchill in the relevant respects and degrees but without representing Churchill. Moreover, similarity 2 (Definition 3.3) introduces three new problems.

First, a misrepresentation is one that portrays its target as having properties that are not similar in the relevant respects and to the relevant degrees to the true properties of the target. But then, on similarity 2 (Definition 3.3),  $M$  is not a representation at all. *Ducheyne* [3.72] embraces this conclusion when he offers a variant of a similarity account that explicitly takes the *success* of the hypothesized similarity between a model and its target to be a necessary condition on the model representing the target. In Sect. 3.2 we argued that the possibility of misrepresentation is a condition of adequacy for any acceptable account of representation and so we submit that misrepresentation should not be conflated with nonrepresentation ([3.20, p. 16], [3.23, p. 235]).

Second, similarity in relevant respects and to the relevant degrees does not guarantee that  $M$  represents the right target. As *Suárez* points out [3.23, pp. 233–234], even a regimented similarity can obtain with no corresponding representation. If John dresses up as Pope Innocent X (and he does so perfectly), then he resembles Velázquez’s portrait of the pope (at least in as far as the pope himself resembled the portrait). In cases like these, which Suárez calls *mistargeting*, a model represents one target rather than another, despite the fact that both targets are relevantly similar to the model. Like in the case of Putnam’s ant, the root cause of the problem is that the similarity is accidental. In the case of the ant, the accident occurs at the representation end of the relation, whereas in the case of John’s dressing up the accidental similarity occurs at the target end. Both cases demonstrate that similarity 2 (Definition 3.3) cannot rule out accidental representation.

Third, there may simply be nothing to be similar to because some representations represent no actual object [3.64, p. 26]. Some paintings represent elves and dragons, and some models represent phlogiston and the ether. None of these exist. As *Toon* points out, this is a problem in particular for the similarity view [3.49, pp. 246–247]: models without objects cannot represent what they seem to represent because in order for two things to be similar to each other both have to exist. If there is no ether, then an ether model cannot be similar to the ether.

It would seem that at least the second problem could be solved by adding the requirement that  $M$  denote  $T$  (as considered, but not endorsed, by *Goodman* [3.64, pp. 5–6]). Amending the previous definition accordingly yields:

#### **Definition 3.4 Similarity 3**

A scientific model  $M$  represents a target  $T$  iff  $M$  and  $T$  are similar in relevant respects and to the relevant degrees and  $M$  denotes  $T$ .

This account would also solve the problem with reflexivity (and symmetry), because denotation is directional in a way similarity is not. Unfortunately similarity 3 (Definition 3.4) still suffers from the first and the third problems. It would still lead to the conflation of misrepresentations with nonrepresentations because the first conjunct (similar in the relevant respects) would still be false. And a nonexistent system cannot be denoted and so we have to conclude that models of, say, the ether and phlogiston represent nothing. This seems an unfortunate consequence because there is a clear sense in which models without targets are about something. Maxwell’s writings on the ether provide a detailed and intelligible account of a number of properties of the

ether, and these properties are highlighted in the model. If ether existed then similarity 3 (Definition 3.4) could explain why these were important by appealing to them as being relevant for the similarity between an ether model and its target. But since ether does not, no such explanation is offered.

A different version of the similarity view sets aside the moves made in similarity 3 (Definition 3.4) and tries to improve on similarity 2 (Definition 3.3). The crucial move is to take the very act of *asserting* a specific similarity between a model and a target as constitutive of the scientific representation.

#### Definition 3.5 Similarity 4

A scientific model  $M$  represents a target system  $T$  if and only if a theoretical hypotheses  $H$  asserts that  $M$  and  $T$  are similar in certain respects and to certain degrees.

This comes close to the view *Giere* advocated in *Explaining Science* [3.71, p. 81] (something like this is also found in *Cartwright* ([3.73, pp. 192–193], [3.74, pp. 261–262]) who appeals to a “loose notion of resemblance”; her account of modeling is discussed in more detail in Sect. 3.6.3). This version of the similarity view avoids problems with misrepresentation because, being hypotheses, there is no expectation that the assertions made in  $H$  are true. If they are, then the representation is accurate (or the representation is accurate to the extent that they hold). If they are not, then the representation is a misrepresentation. It resolves the problem of mistargeting because hypotheses identify targets before asserting similarities with  $M$  (that is, the task of picking the right target is now placed in the court of the hypothesis and is no longer expected to be determined by the similarity relation). Finally it also resolves the issue with directionality because  $H$  can be understood as introducing a directionality that is not present in the similarity relation. However, it fails to resolve the problem with representation without a target. If there is no ether, no hypotheses can be asserted about it.

Let us set the issue of nonexistent targets aside for the moment and have a closer look at the notion of representation proposed in similarity 4 (Definition 3.5). A crucial point remains understated in similarity 4 (Definition 3.5). Hypotheses don’t assert themselves; hypotheses are put forward by those who work with representations, in the case of models, scientists. So the crucial ingredient – users – is left implicit in similarity 4 (Definition 3.5).

In a string of recent publications *Giere* made explicit the fact that “scientists are intentional agents with goals and purposes” [3.60, p. 743] and proposed to build this insight explicitly into an account of epistemic

representation. This involves adopting an agent-based notion of representation that focuses on “the activity of representing” [3.60, p. 743]. Analyzing epistemic representation in these terms amounts to analyzing schemes like “ $S$  uses  $X$  to represent  $W$  for purposes  $P$ ” [3.60, p. 743], or in more detail [3.51, p. 274]:

“Agents (1) intend; (2) to use model,  $M$ ; (3) to represent a part of the world  $W$ ; (4) for purposes,  $P$ . So agents specify which similarities are intended and for what purpose.”

This conception of representation had already been proposed half a century earlier by *Apostel* when he urged the following analysis of model-representation [3.75, p. 4]:

“Let then  $R(S, P, M, T)$  indicate the main variables of the modeling relationship. The subject  $S$  takes, in view of the purpose  $P$ , the entity  $M$  as a model for the prototype  $T$ .”

Including the intentions of model agents in the definition of scientific representation is now widely accepted, as we discuss in more detail in Sect. 3.4 (although *Rusanen* and *Lappi* disagree with this, and claim that “the semantics of models as scientific representations should be based on the mind-independent model-world relation” [3.76, p. 317]).

*Giere*’s proposal, in our own terminology comes down to:

#### Definition 3.6 Similarity 5

A scientific model  $M$  represents a target system  $T$  iff there is an agent  $A$  who uses  $M$  to represent a target system  $T$  by proposing a theoretical hypothesis  $H$  specifying a similarity (in certain respects and to certain degrees) between  $M$  and  $T$  for purpose  $P$ .

This definition inherits from similarity 4 (Definition 3.5) the resolutions of the problems of directionality, misrepresentation, and mistargeting; and for the sake of argument we assume that the problem with nonexistent targets can be resolved in one way or other.

A crucial thing to note about similarity 5 (Definition 3.6) is that, by invoking an active role for the purposes and actions of scientists in constituting epistemic representation, it marks a significant change in emphasis for similarity-based accounts. *Suárez* [3.23, pp. 226–227], drawing on *van Fraassen* [3.77] and *Putnam* [3.78], defines *naturalistic* accounts of representation as ones where “whether or not representation obtains depends on facts about the world and does not in any way answer to the personal purposes, views or interests of enquirers”. By building the purposes of model

users directly into an answer to the ER-problem, similarity 5 (Definition 3.6) is explicitly not a naturalistic account (in contrast, for example, to similarity 1 (Definition 3.2)). As noted in Sect. 3.2 we do not demand a naturalistic account of model-representation (and as we will see later, many of the more developed answers to the ER-problem are also not naturalistic accounts).

Does this suggest that similarity 5 (Definition 3.6) is a successful similarity-based solution to the ER-problem? Unfortunately not. A closer look at similarity 5 (Definition 3.6) reveals that the role of similarity has shifted. As far as offering a solution to the ER-problem is concerned, all the heavy lifting in similarity 5 (Definition 3.6) is done by the appeal to agents and similarity has in fact become an idle wheel. *Giere* implicitly admits this when he writes [3.60, p. 747]:

“How do scientists use models to represent aspects of the world? What is it about models that makes it possible to use them in this way? One way, perhaps the most important way, but probably not the only way, is by exploiting similarities between a model and that aspect of the world it is being used to represent. Note that I am not saying that the model itself represents an aspect of the world because it is similar to that aspect. There is no such representational relationship. [footnote omitted] Anything is similar to anything else in countless respects, but not anything represents anything else. It is not the model that is doing the representing; it is the scientist using the model who is doing the representing.”

But if similarity is not the only way in which a model can be used as a representation, and if it is the use by a scientist that turns a model into a representation (rather than any mind-independent relationship the model bears to the target), then similarity has become otiose in a reply to the ER-problem. A scientist could invoke any relation between  $M$  and  $T$  and  $M$  would still represent  $T$ . Being similar in the relevant respects to the relevant degrees now plays the role either of a representational style, or of a normative criterion for accurate representation, rather than of a grounding of representation. We assess in the next section whether similarity offers a cogent reply to the issues of style and accuracy.

A further problem is that there seems to be a hidden circularity in the analysis. As *Toon* [3.49, pp. 251–252] points out, having a scientist form a theoretical hypothesis about the similarity relation between two objects  $A$  and  $B$  and exploit this similarity for a certain purpose  $P$  is not sufficient for representation.  $A$  and  $B$  could be two cars in a showroom and an engineer inspects car  $A$  and then use her knowledge about similarities to make assertions about  $B$  (for instance if both cars are of the same brand she can infer something about  $B$ 's quality

of manufacturing). This, *Toon* submits, is not a case of representation: neither car is representational. Yet, if we delete the expression *to represent* on the right hand side of the biconditional in similarity 5 (Definition 3.6), the resulting condition provides an accurate description of what happens in the showroom. So the only difference between the nonrepresentational activity of comparing cars and representing  $B$  by  $A$  is that in one case  $A$  is *used to represent* and in the other it's only *used*. So representation is explained in terms of *to represent*, which is circular. So similarity 5 (Definition 3.6) does not provide nontrivial conditions for something to be used *as a representation*.

One way around the problem would be to replace *to represent* by *to denote*. This, however, would bring the account close to similarity 3 (Definition 3.4), and it would suffer from the same problems.

*Mäki* [3.79] suggested an extension of similarity 5 (Definition 3.6), which he explicitly brands as “a (more or less explicit) version” of *Giere*'s. *Mäki* adds two conditions to *Giere*'s: the agent uses the model to address an audience  $E$  and adds a commentary  $C$  [3.79, p. 57]. The role of the commentary is to specify the nature of the similarity. This is needed because [3.79, p. 57]:

“representation does not require that all parts of the model resemble the target in all or just any arbitrary respects, or that the issue of resemblance legitimately arises in regard to all parts. The relevant model parts and the relevant respects and degrees of resemblance must be delimited.”

What these relevant respects and degrees of resemblance are depends on the purposes of the scientific representation in question. These are not determined *in the model* as it were, but are pragmatic elements. From this it transpires that in effect  $C$  plays the same role as that played by theoretical hypotheses in *Giere*'s account. Certain aspects of  $M$  are chosen as those relevant to the representational relationship between  $M$  and  $T$ .

The addition of an audience, however, is problematic. While models are often shared publicly, this does not seem to be a necessary condition for the representational use of a model. There is nothing that precludes a lone scientist from coining a model  $M$  and using it representationally. That some models are easier to grasp, and therefore serve as more effective tools to drive home a point in certain public settings, is an indisputable fact, but one that has no bearing on a model's status as a representation. The pragmatics of communication and the semantics of modeling are separate issues.

The conclusion we draw from this discussion is that similarity does not offer a viable answer to the ER-problem.

### 3.3.2 Accuracy and Style

Accounting for the possibility of misrepresentation resulted in a shift of the division of labor for the more developed similarity-based accounts. Rather than being the relation that grounds representation, similarity should be considered as setting a standard of accuracy or as providing an answer to the question of style (or both). The former is motivated by the observation that a proposed similarity between  $M$  and  $T$  could be wrong, and hence if the model user's proposal does in fact hold (and  $M$  and  $T$  are in fact similar in the specified way) then  $M$  is an accurate representation of  $T$ . The latter transpires from the simple observation that a judgment of accuracy in fact presupposes a choice of respects in which  $M$  and  $T$  are claimed to be similar. Simply proposing that they are similar in some unspecified respect is vacuous. But delineating relevant properties could potentially provide an answer to the problem of style. For example, if  $M$  and  $T$  are proposed to be similar with respect to their causal structure, then we might have a style of causal modeling; if  $M$  and  $T$  are proposed to be similar with respect to structural properties, then we might have a style of structural modeling; and so on and so forth. So the idea is that if  $M$  representing  $T$  involves the claim that  $M$  and  $T$  are similar in a certain respect, the respect chosen specifies the style of the representation; and if  $M$  and  $T$  are in fact similar in that respect (and to the specified degree), then  $M$  accurately represents  $T$  within that style.

In this section we investigate both options. But before delving into the details, let us briefly step back and reflect on possible constraints on viable answers. Taking his cue from Lopes' [3.59] discussion of pictures, Downes [3.80, pp. 421–422] proposes two constraints on allowable notions of similarity. The first, which he calls the *independence challenge*, requires that a user must be able to specify the relevant representation-grounding similarity *before* engaging in a comparison between  $M$  and  $T$ . Similarities that are recognizable only with hindsight are an unsound foundation of a representation. We agree with this requirement, which in fact is also a consequence of the surrogative reasoning condition: a model can generate novel hypotheses only if (at least some of the) similarity claims are not known only *ex post facto*.

Downes' second constraint, the *diversity constraint*, is the requirement that the relevant notion of similarity has to be identical in all kinds of representation and across all representational styles. So all models must bear the same similarity relations to their targets. Whatever its merits in the case of pictorial representation, this observation does not hold water in the case

of scientific representation. Both Giere and Teller have insisted – rightly, in our view – that there need not be a substantive sense of similarity uniting all representations (see also Callender and Cohen [3.26, p. 77] for a discussion). A proponent of the similarity view is free to propose different kinds of similarity for different representations and is under no obligation to also show that they are special cases of some overarching conception of similarity.

We now turn to the issue of style. A first step in the direction of an understanding of styles is the explicit analysis of the notion of similarity. Unfortunately the philosophical literature contains surprisingly little explicit discussion about what it means for something to be similar to something else. In many cases similarity is taken to be primitive, possible worlds semantics being a prime example. The problem is then compounded by the fact that the focus is on comparative overall similarity instead rather than on similarity in respect and degrees; for a critical discussion see [3.81]. Where the issue is discussed explicitly, the standard way of cashing out what it means for an object to be similar to another object is to require that they co-instantiate properties. This is the idea that Quine [3.82, pp. 117–118] and Goodman [3.83, p. 443] had in mind in their influential critiques of the notion. They note that if all that is required for two things to be similar is that they co-instantiate *some* property, then everything is similar to everything else, since any pair of objects have at least one property in common.

The issue of similarity seems to have attracted more attention in psychology. In fact, the psychological literature provides formal accounts to capture it directly in more fully worked out accounts. The two most prominent suggestions are the *geometric* and *contrast* accounts (see [3.84] for an up-to-date discussion). The former, associated with Shepard [3.85], assigns objects a place in a multidimensional space based on values assigned to their properties. This space is then equipped with a metric and the degree of similarity between two objects is a function of the distance between the points representing the two objects in that space.

This account is based on the strong assumptions that values can be assigned to all features relevant to similarity judgments, which is deemed unrealistic. This problem is supposed to be overcome in Tversky's *contrast account* [3.86]. This account defines a graded notion of similarity based on a weighted comparison of properties. Weisberg ([3.33, Chap. 8], [3.87]) has recently introduced this account into the philosophy of science where it serves as the starting point for his so-called *weighted feature matching account of model world-relations*. This account is our primary interest here.



The account introduces a set  $\Delta$  of relevant properties. Let then  $\Delta_M \subseteq \Delta$  be the set of properties from  $\Delta$  that are instantiated by the model  $M$ ; likewise  $\Delta_T$  is the set of properties from  $\Delta$  instantiated by the target system. Furthermore let  $f$  be a ranking function assigning a real number to every subset of  $\Delta$ . The simplest version of a ranking function is one that assigns to each set the number of properties in the set, but rankings can be more complex, for instance by giving important properties more weight. The level of similarity between  $M$  and  $T$  is then given by the following equation [3.87, p. 788] (the notation is slightly amended)

$$S(M, T) = \theta f(\Delta_M \cap \Delta_T) - \alpha f(\Delta_M - \Delta_T) - \beta f(\Delta_T - \Delta_M),$$

where  $\alpha$ ,  $\beta$  and  $\theta$  are weights, which can in principle take any value. This equation provides “a similarity score that can be used in comparative judgments of similarity” [3.87, p. 788]. The score is determined by weighing the properties the model and target have in common against those they do not. (Thus we note that this account could be seen as a quantitative version of *Hesse’s* [3.88] theory of analogy in which properties that  $M$  and  $T$  share are the *positive analogy* and ones they don’t share are the *negative analogy*.) In the above formulation the similarity score  $S$  can in principle vary between any two values (depending on the choice of the ranking function and the value of the weights). One can then use standard mathematical techniques to renormalize  $S$  so that it takes values in the unit interval  $[0, 1]$  (these technical moves need not occupy us here and we refer the reader to *Weisberg* for details [3.33, Chap. 8]).

The obvious question at this point is how the various blanks in the account can be filled. First in line is the specification of a property set  $\Delta$ . *Weisberg* is explicit that there are no general rules to rely on and that “the elements of  $\Delta$  come from a combination of context, conceptualization of the target, and theoretical goals of the scientist” [3.33, p. 149]. Likewise, the ranking function as well as the values of weighting parameters depend on the goals of the investigation, the context, and the theoretical framework in which the scientists operate. *Weisberg* further divides the elements of  $\Delta$  into *attributes* and *mechanisms*. The former are the “the properties and patterns of a system” while the latter are the “underlying mechanism[s] that generates these properties” [3.33, p. 145]. This distinction is helpful in the application to concrete cases, but for the purpose of our conceptual discussion it can be set aside.

Irrespective of these choices, the similarity score  $S$  has a number of interesting features. First, it is asymmetrical for  $\alpha \neq \beta$ , which makes room for the pos-

sibility of  $M$  being similar to  $T$  to a different degree than  $T$  is similar to  $M$ . So  $S$  provides the asymmetrical notion of similarity mentioned in Sect. 3.3.1. Second,  $S$  has a property called *maximality*: everything is maximally similar to itself and every other nonidentical object is equally or less similar. Formally:  $S(A, A) \geq S(A, B)$  for all objects  $A$  and  $B$  as long as  $A \neq B$  [3.33, p. 154].

What does this account contribute to a response to the question of style? The answer, we think, is that it has heuristic value but does not provide substantive account. In fact, stylistic questions stand outside the proposed framework. The framework can be useful in bringing questions into focus, but eventually the substantive stylistic questions concern inclusion criteria for  $\Delta$  (what properties do we focus on?), the weight given by  $f$  to properties (what is the relative importance of properties?) and the value of the parameters (how significant are disagreements between the properties of  $M$  and  $T$ ?). These questions have to be answered outside the account. The account is a framework in which questions can be asked but which does not itself provide answers, and hence no classification of representational styles emerges from it.

Some will say that this is old news. *Goodman* denounced similarity as “a pretender, an impostor, a quack” [3.83, p. 437] not least because he thought that it merely put a label to something unknown without analyzing it. And even some proponents of the similarity view have insisted that no general characterization of similarity was possible. Thus *Teller* submits that [3.52, p. 402]:

“[t]here can be no general account of similarity, but there is also no need for a general account because the details of any case will provide the information which will establish just what should count as relevant similarity in that case.”

This amounts to nothing less than the admission that no analysis of similarity (or even different kinds of similarity) is possible and that we have to deal with each case in its own right.

Assume now, for the sake of argument, that the stylistic issues have been resolved and full specifications of relevant properties and their relative weights are available. It would then seem plausible to say that  $S(M, T)$  provides a degree of accuracy. This reading is supported by the fact that *Weisberg* paraphrases the role of  $S(M, T)$  as providing “standards of fidelity” [3.33, p. 147]. Indeed, in response to *Parker* [3.89], *Weisberg* claims that his weighted feature matching account is supposed to answer the ER-problem and provide standards of accuracy.

As we have seen above,  $S(M, T)$  is maximal if  $M$  is a perfect replica of  $T$  (with respect to the properties

in  $\Delta$ ), and the fewer properties  $M$  and  $T$  share, the less accurate the representation becomes. This lack of accuracy is then reflected in a lower similarity score. This is plausible and Weisberg's account is indeed a step forward in the direction of quantifying accuracy.

Weisberg's account is an elaborate version of the co-instantiation account of similarity. It improves significantly on simple versions, but it cannot overcome that account's basic limitations. Niiniluoto distinguishes between two different kinds of similarities [3.90, pp. 272–274]: partial identity and likeness (which also feature in Hesse's discussion of analogies, see, for instance [3.88, pp. 66–67]). Assume  $M$  instantiates the relevant properties  $P_1, \dots, P_n$  and  $T$  instantiates the relevant properties  $Q_1, \dots, Q_n$ . If these properties are identical, i. e., if  $P_i = Q_i$  for all  $i = 1, \dots, n$ , then  $M$  and  $T$  are similar in the sense of being *partially identical*. Partial identity contrasts with what Niiniluoto calls *likeness*.  $M$  and  $T$  are similar in the sense of likeness if the properties are not identical but similar themselves:  $P_i$  is similar to  $Q_i$  for all  $i = 1, \dots, n$ . So in likeness the similarity is located at the level of the properties themselves. For example, a red post box and a red London bus are similar with respect to their color, even if they do not instantiate the exact same shade of red. As Parker [3.89, p. 273] notes, Weisberg's account (like all co-instantiation accounts) deals well with partial identity, but has no systematic place for likeness. To deal with likeness Weisberg would in effect have to reduce likeness to partial identity by introducing *imprecise* properties which encompass the  $P_i$  and the  $Q_i$ . Parker [3.89] suggests that this can be done by introducing intervals in the feature set, for instance of the form “the value of feature  $X$  lies in the interval  $[x - \varepsilon, x + \varepsilon]$ ” where  $\varepsilon$  is a parameter specifying the precision of overlap. To illustrate she uses Weisberg's example of the San Francisco bay model and claims that in order to account for the similarity between the model and the actual bay with respect to their Froude number Weisberg has to claim something like [3.89, p. 273]:

“The Bay model and the real Bay share the property of having a Froude number that is within 0.1 of the real Bay's number. It is more natural to say that the Bay model and the real Bay have *similar* Froude numbers – similar in the sense that their values differ by at most 0.1.”

In his response Weisberg accepts this and argues that he is trying to provide a reductive account of similarity that bottoms out in properties shared and those not shared [3.91, p. 302]. But such interval-valued properties have to be part of  $\Delta$  in order for the formal account to capture them. This means that another important decision regarding whether or not  $M$  and  $T$  are

similar occurs outside of the formal account itself. The inclusion criteria on what goes into  $\Delta$  now not only has to delineate relevant properties, but, at least for the quantitative ones, also has to provide an interval defining when they qualify as similar. Furthermore, it remains unclear how to account for  $M$  and  $T$  to be alike with respect to their qualitative properties. The similarity between genuinely qualitative properties cannot be accounted for in terms of numerical intervals. This is a particularly pressing problem for Weisberg, because he takes the ability to compare models and their targets with respect to their qualitative properties as a central desideratum for any account of similarity between the two [3.33, p. 136].

### 3.3.3 Problems of Ontology

Another problem facing similarity-based approaches concerns their treatment of the ontology of models. If models are supposed to be similar to their targets in the ways specified by theoretical hypotheses or commentaries, then they must be the *kind* of things that can be so similar.

Some models are homely physical objects. The Army Corps of Engineers' model of the San Francisco bay is a water basin equipped with pumps to simulate the action of tidal flows [3.33]; ball and stick models of molecules are made of metal or wood [3.92]; the Phillips–Newlyn model of an economy is system of pipes and reservoirs [3.93]; and model organisms in biology are animals like worms and mice [3.46]. For models of this kind similarity is straightforward (at least in principle) because they are of the same ontological kind as their respective targets: they are material objects.

But many interesting scientific models are not like this. Two perfect spheres with a homogeneous mass distribution that interact only with each other (the Newtonian model of the Sun–Earth system) or a single-species population isolated from its environment and reproducing at fixed rate at equidistant time steps (the logistic growth model of a population) are what *Hacking* aptly describes as “something you hold in your head rather than your hands” [3.44, p. 216]. Following Thomson–Jones [3.94] we call such models *nonconcrete models*. The question then is what kind of objects nonconcrete models are. Giere submits that they are abstract objects ([3.60, p. 747], cf. [3.51, p. 270], [3.71, p. 81]):

“Models in advanced sciences such as physics and biology should be abstract objects constructed in conformity with appropriate general principles and specific conditions.”

The appeal to abstract entities brings a number of difficulties with it. The first is that the class of abstract

objects is rather large. Numbers and other objects of pure mathematics, classes, propositions, concepts, the letter *A*, and Dante's *Inferno* are abstract objects [3.95], and *Hale* [3.96, pp. 86–87] lists no less than 12 different possible characterizations of abstract objects. At the very least this list shows that there is great variety in abstract objects and classifying models as abstract objects adds little specificity to an account of what models are. *Giere* could counter that he limits attention to those abstract objects that possess “all and only the characteristics specified in the principles” [3.60, p. 745], where principles are general rules like Newton's laws of motion. He further specifies that he takes “abstract entities to be human constructions” and that “abstract models are definitely not to be identified with linguistic entities such as words or equations” [3.60, p. 747]. While this narrows down the choices somehow, it still leaves many options and ultimately the ontological status of models in a similarity account remains unclear.

*Giere* fails to expand on this ontological issue for a reason: he dismisses the problem as one that philosophers of science can set aside without loss. He voices skepticism about the view that philosophers of science “need a deeper understanding of imaginative processes and of the objects produced by these process” [3.97, p. 250] or that “we need say much more [...] to get on with the job of investigating the functions of models in science” [3.97].

We remain unconvinced about this skepticism, not least because there is an obvious yet fundamental issue with abstract objects. No matter how the above issues are resolved (and irrespective of whether they are resolved at all), at the minimum it is clear that models are *abstract* in the sense that they have no spatiotemporal location. *Teller* [3.52, p. 399] and *Thomson-Jones* [3.98] supply arguments suggesting that this alone causes serious problems for the similarity account. The similarity account demands that models can instantiate properties and relations, since this is a necessary condition on them being similar to their targets. In particular, it requires that models can instantiate the properties and relations mentioned in theoretical hypotheses or commentaries. But such properties and relations are typically *physical*. And if models have no spatiotemporal location, then they do not instantiate any such properties or relations. *Thomson-Jones'* example of the idealized pendulum model makes this clear. If the idealized pendulum is abstract then it is difficult to see how to make sense of the idea that it has a length, or a mass, or an oscillation period of any particular time.

An alternative suggestion due to *Teller* [3.52] is that we should instead say that whilst “concrete objects HAVE properties [...] properties are PARTS of models” [3.52, p. 399] (original capitalization). It is not

entirely clear what *Teller* means by this, but our guess is that he would regard models as bundles of properties. Target systems, as concrete objects, are the sorts of things that can instantiate properties delineated by theoretical hypotheses. Models, since they are abstract, cannot. But rather than being objects instantiating properties, a model can be seen as a bundle of properties. A collection of properties is an abstract entity that is the sort of thing that can contain the properties specified by theoretical hypotheses as parts. The similarity relation between models and their targets shifts from the co-instantiation of properties, to the idea that targets instantiate (relevant) properties that are parts of the model. With respect to what it means for a model to be a bundle of properties *Teller* claims that the “[d]etails will vary with ones account of instantiation, of properties and other abstract objects, and of the way properties enter into models” [3.52].

But as *Thompson-Jones* [3.98, pp. 294–295] notes, it is not obvious that this suggestion is an improvement on *Giere's* abstract objects. A bundle view incurs certain metaphysical commitments, chiefly the existence of properties and their abstractness, and a bundle view of objects, concrete or abstract, faces a number of serious problems [3.99]. One might speculate that addressing these issues would push *Teller* either towards the kind of more robust account of abstract objects that he endeavored to avoid, or towards a fictionalist understanding of models.

The latter option has been discussed by *Giere*, who points out that a natural response to *Teller's* and *Thomson-Jones'* problem is to regard models as akin to *imaginary* or *fictional* systems of the sort presented in novels and films. It seems true to say that *Sherlock* is a smoker, despite the fact that *Sherlock* an imaginary detective, and smoking is a physical property. At times, *Giere* seems sympathetic to this view. He says [3.97, p. 249]:

“it is widely assumed that a work of fiction is a creation of human imagination [...] the same is true of scientific models. So, ontologically, scientific models and works of fiction are on a par. They are both imaginary constructs.”

And he observes that [3.51, p. 278]:

“novels are commonly regarded as works of imagination. That, ontologically, is how we should think of abstract scientific models. They are creations of scientists imaginations. They have no ontological status beyond that.”

However, these seem to be occasional slips and he recently positioned himself as an outspoken opponent of any approach to models that likens them to literary

fiction. We discuss these approaches as well as Giere's criticisms of them in Sect. 3.6.

### 3.4 The Structuralist Conception

The structuralist conception of model-representation originated in the so-called semantic view of theories that came to prominence in the second half of the 20th century (*Suppes* [3.100], *van Fraassen* [3.101], and *Da Costa and French* [3.102] provide classical statements of the view; *Byerly* [3.103], *Chakravartty* [3.104], *Klein* [3.105] and *Portides* [3.106, 107] provide critical discussions). The semantic view was originally proposed as an account of theory structure rather than model-representation. The driving idea behind the position is that scientific theories are best thought of as collections of models. This invites the questions: What are these models, and how do they represent their target systems? Defenders of the semantic view of theories take models to be structures, which represent their target systems in virtue of there being some kind of *mapping* (isomorphism, partial isomorphism, homomorphism, ...) between the two. (It is worth noting that Giere, whose account of scientific representation we discussed in the previous section, is also associated with the semantic view, despite not subscribing to either of these positions.)

This conception has two *prima facie* advantages. The first advantage is that it offers a straightforward answer to the ER-problem, and one that accounts for surrogative reasoning: the mappings between the model and the target allow scientists to convert truths found in the model into claims about the target system. The second advantage concerns the applicability of mathematics. There is time-honored position in the philosophy of mathematics that sees mathematics as the study of structures; see, for instance, *Resnik* [3.108] and *Shapiro* [3.109]. It is a natural move for the scientific structuralist to adopt this point of view, which, without further ado, provides a neat explanation of how mathematics is used in scientific modeling.

#### 3.4.1 Structures and the Problem of Ontology

Almost anything from a concert hall to a kinship system can be referred to as a *structure*. So the first task for a structuralist account of representation is to articulate what notion of structure it employs. A number of different notions of structure have been discussed in the literature (for a review see *Thomson-Jones* [3.110]), but by far the most common and widely used is the notion

In sum, the similarity view is yet to be equipped with a satisfactory account of the ontology of models.

of structure one finds in set theory and mathematical logic. A structure  $S$  in that sense (sometimes *mathematical structure* or *set-theoretic structure*) is a composite entity consisting of the following: a nonempty set  $U$  of objects called the domain (or universe) of the structure and a nonempty indexed set  $R$  of relations on  $U$ . With the exception of the caveat below regarding interpretation functions, this definition of structure is widely used in mathematics and logic; see for instance *Machover* [3.111, p. 149], *Hodges* [3.112, p. 2], and *Rickart* [3.113, p. 17]. It is convenient to write these as  $S = \langle U, R \rangle$ , where  $\langle \cdot, \cdot \rangle$  denotes an ordered tuple. Sometimes operations are also included in the definition of a structure. While convenient in some applications, operations are redundant because operations reduce to relations (see *Boolos and Jeffrey* [3.114, pp. 98–99]).

It is important to be clear on what we mean by *object* and *relation* in this context. As *Russell* [3.115, p. 60] points out, in defining the domain of a structure it is irrelevant what the objects are. All that matters from a structuralist point of view is that there are so and so many of them. Whether the object is a desk or a planet is irrelevant. All we need are dummies or placeholders whose *only* property is *objecthood*. Similarly, when defining relations one disregards completely what the relation is *in itself*. Whether we talk about *being the mother of* or *standing to the left of* is of no concern in the context of a structure; all that matters is between which objects it holds. For this reason, a relation is specified purely extensionally: as a class of ordered  $n$ -tuples. The relation literally is nothing over and above this class. So a structure consists of dummy objects between which purely extensionally defined relations hold.

Let us illustrate this with an example. Consider the structure with the domain  $U = \{a, b, c\}$  and the following two relations:  $r_1 = \{a\}$  and  $r_2 = \{\langle a, b \rangle, \langle b, c \rangle, \langle a, c \rangle\}$ . Hence  $R$  consists of  $r_1$  and  $r_2$ , and the structure itself is  $S = \langle U, R \rangle$ . This is a structure with a three-object domain endowed with a monadic property and a transitive relation. Whether the objects are books or iron rods is of no relevance to the structure; they could be literally anything one can think of. Likewise  $r_1$  could be literally any monadic property (being green, being waterproof, etc.) and  $r_2$  could be any (irreflexive) transitive relation (larger than, hotter than, more expensive than, etc.).

It is worth pointing out that this use of *structure* differs from the use one sometimes finds in logic, where linguistic elements are considered part of the model as well. Specifically, over and above  $S = \langle U, R \rangle$ , a structure is also taken to include a language (sometimes called a *signature*)  $L$ , and an interpretation function ([3.112, Chap. 1] and [3.116, pp. 80–81]). But in the context of the accounts discussed in this section, a structure is the ordered pair  $S = \langle U, R \rangle$  as introduced above and so we disregard this alternative use of *structure*.

The first basic posit of the structuralist theory of representation is that models are structures in the above sense (the second is that models represent their targets by being suitably morphic to them; we discuss morphisms in the next subsection). *Suppes* has articulated this stance clearly when he declared that “the meaning of the concept of model is the same in mathematics and the empirical sciences” [3.117, p. 12]. Likewise, *van Fraassen* posits that a “scientific theory gives us a family of models to represent the phenomena”, that “[t]hese models are mathematical entities, so all they have is structure [...]” [3.118, pp. 528–529] and that therefore [3.118, p. 516]

“[s]cience is [...] interpreted as saying that the entities stand in relations which are transitive, reflexive, etc. but as giving no further clue as to what those relations are.”

*Redhead* submits that “it is this abstract structure associated with physical reality that science aims, and to some extent succeeds, to uncover [...]” [3.119, p. 75]. Finally, *French* and *Ladyman* affirm that “the specific material of the models is irrelevant; rather it is the structural representation [...] which is important” [3.120, p. 109]. Further explicit statements of this view are offered by: *Da Costa* and *French* [3.121, p. 249], *Suppes* ([3.122, p. 24], [3.123, Chap. 2]) and *van Fraassen* ([3.101, pp. 43, 64], [3.118, pp. 516, 522], [3.124, p. 483], [3.125, p. 6]).

These structuralist accounts have typically been proposed in the framework of the so-called semantic view of theories. There are differences between them, and formulations vary from author to author. However, as *Da Costa* and *French* [3.126] point out, all these accounts share a commitment to analyzing models as structures. So we are presented with a clear answer to the problem of ontology: models are structures. The remaining issue is what structures themselves are. Are they platonic entities, equivalence classes, modal constructs, or yet something else? This is a hotly debated issue in the philosophy of logic and mathematics; for different positions see for instance *Dummett* [3.127, 295ff.], *Hellman* [3.128, 129], *Redhead* [3.119], *Resnik* [3.108], and *Shapiro* [3.109].

But philosophers of science need not resolve this issue and can pass off the burden of explanation to philosophers of mathematics. This is what usually happens, and hence we don’t pursue this matter further.

An extension of the standard conception of structure is the so-called partial structures approach (for instance, *Da Costa* and *French* [3.102] and *Bueno* et al. [3.130]). Above we defined relations by specifying between which tuples it holds. This naturally allows a sorting of all tuples into two classes: ones that belong to the relation and ones that don’t. The leading idea of partial structures is to introduce a third option: for some tuples it is indeterminate whether or not they belong to the relation. Such a relation is a *partial relation*. A structure with a set  $R$  containing partial relations is a *partial structure* (formal definitions can be found in references given above). Partial structures make room for a process of scientific investigation where one begins not knowing whether a tuple falls under the relation and then learns whether or not it does.

Proponents of that approach are more guarded as regards the ontology of models. *Bueno* and *French* emphasize that “advocates of the semantic account need not be committed to the ontological claim that models *are* structures” [3.53, p. 890] (original emphasis). This claim is motivated by the idea that the task for philosophers of science is to represent scientific theories and models, rather than to reason about them directly. *French* [3.131] makes it explicit that according to his account of the semantic view of theories, a scientific theory is *represented* as a class of models, but should not be identified with that class. Moreover, a class of models is just one way of representing a theory; we can also use an intrinsic characterization and represent the same theory as a set of sentences in order to account for how they can be objects of our epistemic attitudes [3.132].

He therefore adopts a quietist position with respect to what a theory or a model *is*, declining to answer the question [3.131, 133]. There are thus two important notions of representation at play: representation of targets by models, which is the job of scientists, and representation of theories and models by structures, which is the job of philosophers of science. The question for this approach then becomes whether or not the structuralist representation of models and epistemic representation – as partial structures and morphisms that hold between them – is an accurate or useful one. And the concerns raised below remain when translated into this context as well.

There is an additional question regarding the correct formal framework for thinking about models in the structuralist position. *Landry* [3.134] argues that in certain contexts group, rather than set, theory should

be used when talking about structures and morphisms between them, and Halvorson [3.135, 136] argues that theories should be identified with categories rather than classes or sets. Although these discussions highlight important questions regarding the nature of scientific theories, the question of how individual models represent remains unchanged. Halvorson still takes individual models to be set-theoretic structures. And Landry's paper is not an attempt to reframe the representational relationship between models and their targets (see [3.137] for her skepticism regarding how structuralism deals with this question). Thus, for reasons of simplicity we will focus on the structuralist view that identifies models with set-theoretic structures throughout the rest of this section.

### 3.4.2 Structuralism and the ER-Problem

The most basic structuralist conception of scientific representation asserts that scientific models, understood as structures, represent their target systems in virtue of being isomorphic to them. Two structures  $S_a = \langle U_a, R_a \rangle$  and  $S_b = \langle U_b, R_b \rangle$  are isomorphic iff there is a mapping  $f: U_a \rightarrow U_b$  such that (i)  $f$  is one-to-one (bijective) and (ii)  $f$  preserves the system of relations in the following sense: The members  $a_1, \dots, a_n$  of  $U_a$  satisfy the relation  $r_a$  of  $R_a$  iff the corresponding members  $b_1 = f(a_1), \dots, b_n = f(a_n)$  of  $U_b$  satisfy the relation  $r_b$  of  $R_b$ , where  $r_b$  is the relation corresponding to  $r_a$  (for difficulties in how to cash out this notion of correspondence without reference to an interpretation function see Halvorson [3.135] and Glymour [3.138]).

Assume now that the target system  $T$  exhibits the structure  $S_T = \langle U_T, R_T \rangle$  and the model is the structure  $S_M = \langle U_M, R_M \rangle$ . Then the model represents the target iff it is isomorphic to the target:

#### **Definition 3.7 Structuralism 1**

A scientific model  $M$  represents its target  $T$  iff  $S_M$  is isomorphic to  $S_T$ .

This view is articulated explicitly by Ubbink, who posits that [3.139, p. 302]

“a model represents an object or matter of fact in virtue of this structure; so an object is a model [...] of matters of fact if, and only if, their structures are isomorphic.”

Views similar to Ubbink's seem operable in many versions of the semantic view. In fairness to proponents of the semantic view it ought to be pointed out, though, that for a long time representation was not the focus of attention in the view and the attribution of (something like) structuralism 1 (Definition 3.7) to the

semantic view is an extrapolation. Representation became a much-debated topic in the first decade of the 21st century, and many proponents of the semantic view then either moved away from structuralism 1 (Definition 3.7), or pointed out that they never held such a view. We turn to more advanced positions shortly, but to understand what motivates such positions it is helpful to understand why structuralism 1 (Definition 3.7) fails.

An immediate question concerns the target end structure  $S_T$ . At least *prima facie* target systems aren't structures; they are physical objects like planets, molecules, bacteria, tectonic plates, and populations of organisms. An early recognition that the relation between targets and structures is not straightforward can be found in Byerly, who emphasizes that structures are abstracted from objects [3.103, pp. 135–138]. The relation between structures and physical targets is indeed a serious question and we will return to it in Sect. 3.4.4. In this subsection we grant the structuralist the assumption that target systems are (or at least have) structures.

The first and most obvious problem is the same as with the similarity view: isomorphism is symmetrical, reflexive, and transitive, but epistemic representation isn't. This problem could be addressed by replacing isomorphism with an alternative mapping. Bartels [3.140], Lloyd [3.141], and Mundy [3.142] suggest homomorphism; van Fraassen [3.36, 101, 118] and Redhead isomorphic embeddings [3.119]; advocates of the partial structures approach prefer partial isomorphisms [3.102, 120, 121, 143–145]; and Swoyer [3.25] introduces what he calls  $\Delta/\Psi$ -morphisms. We refer to these collectively as *morphisms*.

This solves some, but not all problems. While many of these mappings are asymmetrical, they are all still reflexive, and at least some of them are also transitive. But even if these formal issues could be resolved in one way or another, a view based on structural mappings would still face other serious problems. For ease of presentation we discuss these problems in the context of the isomorphism view; *mutatis mutandis* other formal mappings suffer from the same difficulties (For detailed discussions of homomorphism and partial isomorphism see Suárez [3.23, pp. 239–241] and Pero and Suárez [3.146]; Mundy [3.142] discusses general constraints one may want to impose on morphisms.)

Like similarity, isomorphism is too inclusive: not all things that are isomorphic represent each other. In the case of similarity this case was brought home by Putnam's thought experiment with the ant crawling on the beach; in the case of isomorphism a look at the history of science will do the job. Many mathematical structures have been discovered and discussed long before they have been used in science. Non-Euclidean geometries were studied by mathematicians long before

Einstein used them in the context of spacetime theories, and Hilbert spaces were studied by mathematicians prior to their use in quantum theory. If representation was nothing over and above isomorphism, then we would have to conclude that Riemann discovered general relativity or that Hilbert invented quantum mechanics. This is obviously wrong. Isomorphism on its own does not establish representation [3.20, p. 10].

Isomorphism is more restrictive than similarity: not everything is isomorphic to everything else. But isomorphism is still too abundant to correctly identify the extension of a representation (i.e., the class of systems it represents), which gives rise to a version of the mistargeting problem. The root of the difficulties is that the same structures can be instantiated in different target systems. The  $1/r^2$  law of Newtonian gravity is also the *mathematical skeleton* of Coulomb's law of electrostatic attraction and the weakening of sound or light as a function of the distance to the source. The mathematical structure of the pendulum is also the structure of an electric circuit with condenser and solenoid (a detailed discussion of this case is provided by Kroes [3.147]). Linear equations are ubiquitous in physics, economics and psychology. Certain geometrical structures are instantiated by many different systems; just think about how many spherical things we find in the world. This shows that the same structure can be exhibited by more than one target system. Borrowing a term from the philosophy of mind, one can say that structures are *multiply realizable*. If representation is explicated solely in terms of isomorphism, then we have to conclude that, say, a model of a pendulum also represents an electric circuit. But this seems wrong. Hence isomorphism is too inclusive to correctly identify a representation's extension.

One might try to dismiss this point as an artifact of a misidentification of the target. Van Fraassen [3.101, p. 66], mentions a similar problem under the heading of "unintended realizations" and then expresses confidence that it will "disappear when we look at larger observable parts of the world". Even if there are multiply realizable structures to begin with, they vanish as science progresses and considers more complex systems because these systems are unlikely to have the same structure. Once we focus on a sufficiently large part of the world, no two phenomena will have the same structure. There is a problem with this counter, however. To appeal to *future science* to explain how models work *today* seems unconvincing. It is a matter of fact that we *currently* have models that represent electric circuits and sound waves, and we do not have to await future science providing us with more detailed accounts of a phenomenon to make our models represent what they actually already do represent.

As we have seen in the last section, a misrepresentation is one that portrays its target as having features it doesn't have. In the case of an isomorphism account of representation this presumably means that the model portrays the target as having structural properties that it doesn't have. However, isomorphism demands identity of structure: the structural properties of the model and the target must correspond to one another exactly. A misrepresentation won't be isomorphic to the target. By the lights of structuralism 1 (Definition 3.7) it is therefore is not a representation at all. Like simple similarity accounts, structuralism 1 (Definition 3.7) conflates misrepresentation with nonrepresentation.

Muller [3.148, p. 112] suggests that this problem can be overcome in a two-stage process: one first identifies a submodel of the model, which in fact is isomorphic to at least a part of the target. This *reduced* isomorphism establishes representation. One then constructs "a tailor-made morphism on a case by case basis" [3.148, p. 112] to account for accurate representation. Muller is explicit that this suggestion presupposes that there is "at least one resemblance" [3.148, p. 112] between model and target because "otherwise one would never be called a representation of the other" [3.148, p. 112]. While this may work in some cases, it is not a general solution. It is not clear whether all misrepresentations have isomorphic submodels. Models that are gross distortions of their targets (such as the liquid drop model of the nucleus or the logistic model of a population) may well not have such submodels. More generally, as Muller admits, his solution "precludes total misrepresentation" [3.148, p. 112]. So in effect it just limits the view that representation coincides with correct representation to a submodel. However, this is too restrictive a view of representation. Total misrepresentations may be useless, but they are representations nevertheless.

Another response refers to the partial structures approach and emphasizes that partial structures are in fact constructed to accommodate a mismatch between model and target and are therefore not open to this objection [3.53, p. 888]. It is true that the partial structures framework has a degree of flexibility that the standard view does not. However, we doubt that this flexibility stretches far enough. While the partial structure approach deals successfully with incomplete representations, it does not seem to deal well with distortive representations (we come back to this point in the next subsection). So the partial structures approach, while enjoying an advantage over the standard approach, is nevertheless not yet home and dry.

Like the similarity account, structuralism 1 (Definition 3.7) has a problem with nonexistent targets because no model can be isomorphic to something that doesn't

exist. If there is no ether, a model can't be isomorphic to it. Hence models without target cannot represent what they seem to represent.

Most of these problems can be resolved by making moves similar to the ones that lead to similarity 5 (Definition 3.6): introduce agents and hypothetical reasoning into the account of representation. Going through the motions one finds:

#### **Definition 3.8 Structuralism 2**

A scientific model  $M$  represents a target system  $T$  iff there is an agent  $A$  who uses  $M$  to represent a target system  $T$  by proposing a theoretical hypothesis  $H$  specifying an isomorphism between  $S_M$  and  $S_T$ .

Something similar to this was suggested by Adams [3.149, p. 259] who appeals to the idea that physical systems are the *intended* models of a theory in order to differentiate them from purely mathematical models of a theory. This suggestion is also in line with van Fraassen's recent pronouncements on representation. He offers the following as the *Hauptstatz* of a theory of representation: "there is no representation except in the sense that some things are used, made, or taken, to represent things as thus and so" [3.36, p. 23]. Likewise, Bueno submits that "representation is an *intentional* act relating two objects" [3.150, p. 94] (original emphasis), and Bueno and French point out that using one thing to represent another thing is not only a function of (partial) isomorphism but also depends on *pragmatic* factors "having to do with the use to which we put the relevant models" [3.53, p. 885]. This, of course, gives up on the idea of an account that reduces representation to intrinsic features of models and their targets. At least one extra element, the model user, also features in whatever relation is supposed to constitute the representational relationship between  $M$  and  $T$ . In a world with no agents, there would be no scientific representation.

This seems to be the right move. Like similarity 5 (Definition 3.6), structuralism 2 (Definition 3.8) accounts for the directionality of representation and has no problem with misrepresentation. But, again as in the case of similarity 5 (Definition 3.6), this is a Pyrrhic victory as the role of isomorphism has shifted. The crucial ingredient is the agent's intention and isomorphism has in fact become either a representational style or normative criterion for accurate representation. Let us now assess how well isomorphism fares as a response to these problems.

### **3.4.3 Accuracy, Style and Demarcation**

The problem of style is to identify representational styles and characterize them. Isomorphism offers an

obvious response to this challenge: one can represent a system by coming up with a model that is structurally isomorphic to it. We call this the isomorphism-style. This style also offers a clear-cut condition of accuracy: the representation is accurate if the hypothesized isomorphism holds; it is inaccurate if it doesn't.

This is a neat answer. The question is what status it has vis-à-vis the problem of style. Is the isomorphism-style merely one style among many other styles which are yet to be identified, or is it in some sense privileged? The former is uncontentious. However, the emphasis many structuralists place on isomorphism suggests that they do not regard isomorphism as merely one way among others to represent something. What they seem to have in mind is the stronger claim that a representation *must* be of that sort, or that the isomorphism-style is the only acceptable style.

This claim seems to conflict with scientific practice. Many representations are inaccurate in some way. As we have seen above, partial structures are well equipped to deal with incomplete representations. However, not all inaccuracies are due to something being left out. Some models distort, deform and twist properties of the target in ways that seem to undercut isomorphism. Some models in statistical mechanics have an infinite number of particles and the Newtonian model of the solar system represents the sun as perfect sphere where it in reality is fiery ball with no well-defined surface at all. It is at best unclear how isomorphism, partial or otherwise, can account for these kinds of idealizations. From an isomorphism perspective all one can say about such idealizations is that they are failed isomorphism representations (or isomorphism misrepresentations). This is rather uninformative. One might try to characterize these idealizations by looking at *how* they fail to be isomorphic to their targets, but we doubt that this is going very far. Understanding how distortive idealizations work requires a positive characterization of them, and we cannot see how such a characterization could be given within the isomorphism framework. So one has to recognize styles of representation other than isomorphism.

This raises that question of whether other mappings such as homomorphisms or embeddings would fit the bill. They would, we think, make valuable additions to the list of styles, but they would not fill all gaps. Like isomorphism, these mappings are not designed to accommodate distortive idealizations, and hence a list of styles that includes them still remains incomplete.

Structuralism's stand on the demarcation problem is by and large an open question. Unlike similarity, which has been widely discussed across different domains, isomorphism is tied closely to the formal framework of set theory, and it has been discussed only sparingly outside the context of the mathematized sciences. An ex-



ception is *French*, who discusses isomorphism accounts in the context of pictorial representation [3.35]. He discusses in detail *Budd's* [3.151] account of pictorial representation and points out that it is based on the notion of a structural isomorphism between the structure of the surface of the painting and the structure of the relevant visual field. Therefore representation is the perceived isomorphism of structure [3.35, pp. 1475–1476] (this point is reaffirmed by *Bueno* and *French* [3.53, pp. 864–865]; see *Downes* [3.80, pp. 423–425] for a critical discussion). In a similar vein, *Bueno* claims that the partial structures approach offers a framework in which different representations – among them “outputs of various instruments, micrographs, templates, diagrams, and a variety of other items” [3.150, p. 94] – can be accommodated. This would suggest that an isomorphism account of representation at least has a claim to being a universal account covering representations across different domains.

This approach faces a number of questions. First, neither a visual field nor a painting is a structure, and the notion of there being an isomorphism in the set theoretic sense between the two at the very least needs unpacking. The theory is committed to the claim that paintings and visual fields have structures, but, as we will see in the next subsection, this claim faces serious issues. Second, *Budd's* theory is only one among many theories of pictorial representation, and most alternatives do not invoke isomorphism. So there is question whether a universal claim can be built on *Budd's* theory. In fact, there is even a question about isomorphism's universality within scientific representation. Nonmathematized sciences work with models that aren't structures. *Godfrey-Smith* [3.152], for instance, argues that models in many parts of biology are imagined concrete objects. There is a question whether isomorphism can explain how models of that kind represent.

This points to a larger issue. The structuralist view is a rational reconstruction of scientific modeling, and as such it has some distance from the actual practice. Some philosophers have worried that this distance is too large and that the view is too far removed from the actual practice of science to be able to capture what matters to the practice of modeling (this is the thrust of many contributions to [3.11]; see also [3.73]). Although some models used by scientists may be best thought of as set theoretic structures, there are many where this seems to contradict how scientists actually talk about, and reason with, their models. Obvious examples include physical models like the San Francisco bay model [3.33], but also systems such as the idealized pendulum or imaginary populations of interbreeding animals. Such models have the strange property of being *concrete-if-real* and scientists talk about them as if they were real systems,

despite the fact that they are obviously not. *Thomson-Jones* [3.98] dubs this *face value practice*, and there is a question whether structuralism can account for that practice.

### 3.4.4 The Structure of Target Systems

Target systems are physical objects: atoms, planets, populations of rabbits, economic agents, etc. Isomorphism is a relation that holds between two structures and claiming that a set theoretic structure is isomorphic to a piece of the physical world is *prima facie* a category mistake. By definition, all of the mappings suggested – isomorphism, partial isomorphism, homomorphism, or isomorphic embedding – only hold between two structures. If we are to make sense of the claim that the model is isomorphic to its target we have to assume that the target somehow exhibits a certain structure  $S_T = \langle U_T, R_T \rangle$ . But what does it mean for a target system – a part of the physical world – to possess a structure, and where in the target system is the structure located?

The two prominent suggestions in the literature are that data models are the target end structures represented by models, and that structures are, in some sense, instantiated in target systems. The latter option comes in three versions. The first version is that a structure is ascribed to a system; the second version is that systems instantiate structural universals; and the third version claims that target systems simply are structures. We consider all suggestions in turn.

What are data models? Data are what we gather in experiments. When observing the motion of the moon, for instance, we choose a coordinate system and observe the position of the moon in this coordinate system at consecutive instants of time. We then write down these observations. The data thus gathered are called the *raw* data. The raw data then undergo a process of cleansing, rectification and regimentation: we throw away data points that are obviously faulty, take into consideration what the measurement errors are, take averages, and usually idealize the data, for instance by replacing discrete data points by a continuous function. Often, although not always, the result is a smooth curve through the data points that satisfies certain theoretical desiderata (*Harris* [3.153] and *van Fraassen* [3.36, pp. 166–168] elaborate on this process). These resulting data models can be treated as set theoretic structures. In many cases the data points are numeric and the data model is a smooth curve through these points. Such a curve is a relation over  $\mathbb{R}^n$  (for some  $n$ ), or subsets thereof, and hence it is structure in the requisite sense.

*Suppes* [3.122] was the first to suggest that data models are the targets of scientific models: models don't represent parts of the world; they represent data

structures. This approach has then been adopted by *van Fraassen*, when he declares that “[t]he whole point of having theoretical models is that they should fit the phenomena, that is, fit the models of data” [3.154, p. 667]. He has defended this position numerous times over the years ([3.77, p. 164], [3.101, p. 64], [3.118, p. 524], [3.155, p. 229] and [3.156, p. 271]) including in his most recent book on representation [3.36, pp. 246, 252]. So models don’t represent planets, atoms or populations; they represent data that are gathered when performing measurements on planets, atoms or populations.

This revisionary point of view has met with stiff resistance. *Muller* articulates the unease about this position as follows [3.148, p. 98]:

“the best one could say is that a data structure  $\mathcal{D}$  seems to act as *simulacrum* of the concrete actual being  $B$  [...] But this is not good enough. We don’t want simulacra. We want the real thing. Come on.”

*Muller*’s point is that science aims (or at least has to aim) to represent real systems in the world and not data structures. *Van Fraassen* calls this the “loss of reality objection” [3.36, p. 258] and accepts that the structuralist must ensure that models represent target systems, rather than finishing the story at the level of data. In his [3.36] he addresses this issue in detail and offers a solution. We discuss his solution below, but before doing so we want to articulate the objection in more detail. To this end we briefly revisit the discussion about phenomena and data which took place in the 1980s and 1990s.

*Bogen* and *Woodward* [3.157], *Woodward* [3.158], and more recently (and in a somewhat different guise) *Teller* [3.159], introduced the distinction between phenomena and data and argue that models represent phenomena, not data. The difference is best introduced with an example: the discovery of weak neutral currents [3.157, pp. 315–318]. What the model at stake consists of is particles: neutrinos, nucleons, and the  $Z^0$  particle, along with the reactions that take place between them. (The model we are talking about here is not the so-called standard model of elementary particles as a whole. Rather, what we have in mind is one specific model about the interaction of certain particles of the kind one would find in a theoretical paper on this experiment.) Nothing of that, however, shows in the relevant data. CERN (Conseil Européen pour la Recherche Nucléaire) in Geneva produced 290 000 bubble chamber photographs of which roughly 100 were considered to provide evidence for the existence of neutral currents. The notable point in this story is that there is no part of the model (provided by quantum field theory) that could be claimed to be isomorphic to these pho-

tographs. Weak neutral currents are the phenomenon under investigation; the photographs taken at CERN are the raw data, and any summary one might construct of the content of these photographs would be a data model. But it’s weak neutral currents that occur in the model; not any sort of data we gather in an experiment.

This is not to say that these data have nothing to do with the model. The model posits a certain number of particles and informs us about the way in which they interact both with each other and with their environment. Using this knowledge we can place them in a certain experimental context. The data we then gather in an experiment are the product of the elements of the model and of the way in which they operate in that context. Characteristically this context is one that we are able to control and about which we have reliable knowledge (knowledge about detectors, accelerators, photographic plates and so on). Using this and the model we can derive predictions about what the outcomes of an experiment will be. But, and this is the salient point, these predictions involve the entire experimental setup and not only the model and there is nothing in the model itself with which one could compare the data. Hence, data are highly contextual and there is a big gap between observable outcomes of experiments and anything one might call a substructure of a model of neutral currents.

To underwrite this claim *Bogen* and *Woodward* notice that parallel to the research at CERN, the National Accelerator Laboratory (NAL) in Chicago also performed an experiment to detect weak neutral currents, but the data obtained in that experiment were quite different. They consisted of records of patterns of discharge in electronic particle detectors. Though the experiments at CERN and at NAL were totally different and as a consequence the data gathered had nothing in common, they were meant to provide evidence for the same theoretical model. But the model, to reiterate the point, does not contain any of these contextual factors. It posits certain particles and their interaction with other particles, not how detectors work or what readings they show. That is, the model is not idiosyncratic to a special experimental context in the way the data are and therefore it is not surprising that they do not contain a substructure that is isomorphic to the data. For this reason, models represent phenomena, not data.

It is difficult to give a general characterization of phenomena because they do not belong to one of the traditional ontological categories [3.157, p. 321]. In fact, phenomena fall into many different established categories, including particular objects, features, events, processes, states, states of affairs, or they defy classification in these terms altogether. This, however, does not detract from the usefulness of the concept of a phe-

nomenon because specifying one particular ontological category to which all phenomena belong is inessential to the purpose of this section. What matters to the problem at hand is the distinctive role they play in connection with representation.

What then is the significance of data, if they are not the kind of things that models represent? The answer to this question is that data perform an evidential function. That is, data play the role of evidence for the presence of certain phenomena. The fact that we find a certain pattern in a bubble chamber photograph is evidence for the existence of neutral currents. Thus construed, we do not denigrate the importance of data in science, but we do not have to require that data have to be embeddable into the model at stake.

Those who want to establish data models as targets can reply to this in three ways. The first reply is an appeal to radical empiricism. By postulating phenomena over and above data we leave the firm ground of observable things and started engaging in transempirical speculation. But science has to restrict its claims to observables and remain silent (or at least agnostic) about the rest. Therefore, so the objection goes, phenomena are chimeras that cannot be part of any serious account of science. It is, however, doubtful that this helps the data model theorist. Firstly, note that it even rules out representing *observable phenomena*. To borrow *van Fraassen's* example on this story, a population model of deer reproduction would represent data, rather than deer [3.36, pp. 254–260]. Traditionally, empiricists would readily accept that deer, and the rates at which they reproduce, are observable phenomena. Denying that they are represented, by replacing them with data models, seems to be an implausible move. Secondly, irrespective of whether one understands phenomena realistically [3.157] or antirealistically [3.160], it is phenomena that models portray and not data. To deny the reality of phenomena just won't make a theoretical model *represent* data. Whether we regard neutral currents as real or not, it is neutral currents that are portrayed in a field-theoretical model, not bubble chamber photographs. Of course, one can suspend belief about the reality of these currents, but that is a different matter.

The second reply is to invoke a chain of representational relationships. *Brading* and *Landry* [3.137] point out that the connection between a model and the world can be broken down in two parts: the connection between a model and a data model, and the connection between a data model and the world [3.137, p. 575]. So the structuralist could claim that scientific models represent data models in virtue of an isomorphism between the two and additionally claim that data models in turn represent phenomena. But the key questions that

need to be addressed here are: (a) What establishes the representational relationship between data models and phenomena? and (b) Why if a scientific model represented some data model, which in turn represented some phenomenon, would that establish a representational relationship between the model and the phenomenon itself? With respect to the first question, *Brading* and *Landry* argue that it cannot be captured within the structuralist framework [3.137, p. 575]. The question has just been pushed back: rather than asking how a scientific model qua mathematical structure represents a phenomenon, we now ask how a data model qua mathematical structure represents a phenomenon. With respect to the second question, although representation is not intransitive, it is not transitive [3.20, pp. 11–12]. So more needs to be said regarding how a scientific model representing a data model, which in turn represents the phenomenon from which data are gathered, establishes a representational relationship between the first and last element in the representational chain.

The third reply is due to *van Fraassen* [3.36]. His *Wittgensteinian* solution is to diffuse the loss of reality objection. Once we pay sufficient attention to the pragmatic features of the contexts in which scientific and data models are used, *van Fraassen* claims, there actually is no difference between representing data and representing a target (or a phenomenon in Bogen and Woodward's sense) [3.36, p. 259]:

“in a context in which a given [data] model is *someone's* representation of a phenomenon, there is *for that person* no difference between the question *whether a theory* [theoretical model] *fits that representation* and the question *whether that theory fits the phenomenon.*”

*Van Fraassen's* argument for this claim is long and difficult and we cannot fully investigate it here; we restrict attention to one crucial ingredient and refer the reader to *Nguyen* [3.161] for a detailed discussion of the argument.

Moore's paradox is that we cannot assert sentences of the form *p and I don't believe that p*, where *p* is an arbitrary proposition. For instance, someone cannot assert that Napoleon was defeated in the battle of Waterloo and assert, at the same time, that she doesn't believe that Napoleon was defeated in the battle of Waterloo. *Van Fraassen's* treatment of Moore's paradox is that speakers cannot assert such sentences because the pragmatic commitments incurred by asserting the first conjunct include that the speaker believe that *p*. This commitment is then contradicted by the assertion of the second conjunct. So instances of Moore's paradox are pragmatic contradictions. *Van Fraassen* then draws an analogy between this paradox and the scientific representation. He

submits that a user simply cannot, on pain of pragmatic contradiction, assert that a data model of a target system be embeddable within a theoretical model without thereby accepting that the theoretical model represents the target.

However, *Nguyen* [3.161] argues that in the case of using a data model as a representation of a phenomenon, no such pragmatic commitment is incurred, and therefore no such contradiction follows when accompanied by doubt that the theoretical model also represents the phenomenon. To see why this is the case, consider a more mundane example of representation: a caricaturist can represent Margaret Thatcher as draconian without thereby committing himself to the belief that Margaret Thatcher really is draconian. Pragmatically speaking, acts of representation are weaker than acts of assertion: they do not incur the doxastic commitments required for van Fraassen's analogy to go through. So it seems van Fraassen doesn't succeed in dispelling the loss of reality objection. How target systems enter the picture in the structuralist account of scientific representation remains therefore a question that structuralists who invoke data models as providing the target-end structures must address. Without such an account the structuralist account of representation remains at the level of data, a position that seems implausible, and contrary to actual scientific practice.

We now turn to the second response: that a structure is instantiated in the system. As mentioned above, this response comes in three versions. The first is metaphysically more parsimonious and builds on the systems' constituents. Although target systems are not structures, they are composed of parts that instantiate physical properties and relations. The parts can be used to define the domain of individuals, and by considering the physical properties and relations purely extensionally, we arrive at a class of extensional relations defined over that domain (see for instance *Suppes'* discussion of the solar system [3.100, p. 22]). This supplies the required notion of structure. We might then say that physical systems instantiate a certain structure, and it is this structure that models are isomorphic to.

As an example consider the methane molecule. The molecule consists of a carbon atom and four hydrogen atoms grouped around it, forming a tetrahedron. Between each hydrogen atom and the carbon atom there is a covalent bond. One can then regard the atoms as objects and the bonds are relations. Denoting the carbon atom by  $a$ , and the four hydrogen atoms by  $b$ ,  $c$ ,  $d$ , and  $e$ , we obtain a structure  $S$  with the domain  $U = \{a, b, c, d, e\}$  and the relation  $r = \{\langle a, b \rangle, \langle b, a \rangle, \langle a, c \rangle, \langle c, a \rangle, \langle a, d \rangle, \langle d, a \rangle, \langle a, e \rangle, \langle e, a \rangle\}$ , which can be interpreted as *being connected by a covalent bond*.

The main problem facing this approach is the underdetermination of target-end structure. Underdetermination threatens in two distinct ways. Firstly, in order to identify the structure determined by a target system, a domain of objects is required. What counts as an object in a given target system is a substantial question [3.21]. One could just as well choose bonds as objects and consider the relation *sharing a node with another bond*. Denoting the bonds by  $a'$ ,  $b'$ ,  $c'$  and  $d'$ , we obtain a structure  $S'$  with the domain  $U' = \{a', b', c', d'\}$  and the relation  $r' = \{\langle a', b' \rangle, \langle b', a' \rangle, \langle a', c' \rangle, \langle c', a' \rangle, \langle a', d' \rangle, \langle d', a' \rangle, \langle b', c' \rangle, \langle c', b' \rangle, \langle b', d' \rangle, \langle d', b' \rangle, \langle c', d' \rangle, \langle d', c' \rangle\}$ . Obviously  $S$  and  $S'$  are not isomorphic. So which structure is picked out depends on how the system is described. Depending on which parts one regards as individuals and what relation one chooses, very different structures can emerge. And it takes little ingenuity to come up with further descriptions of the methane molecule, which lead to yet other structures.

There is nothing special about the methane molecule, and any target system can be presented under alternative descriptions, which ground different structures. So the lesson learned generalizes: there is no such thing as *the* structure of a target system. Systems only have a structure under a particular description, and there are many nonequivalent descriptions. This renders talk about a model being isomorphic to target system *simpliciter* meaningless. Structural claims do not *stand on their own* in that their truth rests on the truth of a more concrete description of the target system. As a consequence, descriptions are an integral part of an analysis of scientific representation.

In passing we note that *Frigg* [3.21, pp. 55–56] also provides another argument that pulls in the same direction: structural claims are abstract and are true only relative to a more concrete nonstructural description. For a critical discussion of this argument see *Frisch* [3.162, pp. 289–294] and *Portides*, Chap. 2.

How much of a problem this is depends on how austere one's conception of models is. The semantic view of theories was in many ways the result of an antilinguistic turn in the philosophy of science. Many proponents of the view aimed to exorcise language from an analysis of theories, and they emphasized that the model-world relationship ought to be understood as a *purely* structural relation. *Van Fraassen*, for instance, submits that “no concept which is essentially language dependent has any philosophical importance at all” [3.101, p. 56] and observes that “[t]he semantic view of theories makes language largely irrelevant” [3.155, p. 222]. And other proponents of the view, while less vocal about the irrelevance of language, have not assigned language a systematic place in their analysis of theories.

For someone of that provenance the above argument is bad news. However, a more attenuated position could integrate descriptions in the package of modeling, but this would involve abandoning the idea that representation can be cashed out solely in structural terms. *Bueno* and *French* have recently endorsed such a position. They accept the point that different descriptions lead to different structures and explain that such descriptions would involve “at the very least some minimal mathematics and certain physical assumptions” [3.53, p. 887]. Likewise, Munich structuralists explicitly acknowledge the need for a concrete description of the target system [3.163, pp. 37–38], and they consider these *informal descriptions* to be *internal* to the theory. This is a plausible move, but those endorsing this solution have to concede that there is more to epistemic representation than structures and morphisms.

The second way in which structural indeterminacy can surface is via Newman’s theorem. The theorem essentially says that any system instantiates any structure, the only constraint being cardinality (a practically identical conclusion is reached in Putnam’s so called model-theoretic argument; see *Demopoulos* [3.164] for a discussion). Hence, any structure of cardinality  $C$  is isomorphic to a target of cardinality  $C$  because the target instantiates any structure of cardinality  $C$  (see *Ketland* [3.165] and *Frigg* and *Votsis* [3.166] for discussions). This problem is not unsolvable, but all solutions require that among all structures formally instantiated by a target system one is singled out as being the true or natural structure of the system. How to do this in the structuralist tradition remains unclear (*Ainsworth* [3.167] provides a useful summary of the different solutions).

Newman’s theorem is both stronger and weaker than the argument from multiple descriptions. It’s stronger in that it provides more alternative structures than multiple descriptions. It’s weaker in that many of the structures it provides are *unphysical* because they are purely set theoretical combinations of elements. By contrast, descriptions pick out structures that a system can reasonably be seen as possessing.

The second version of the second response emerges from the literature on the applicability of mathematics. Structural platonists like *Resnik* [3.108] and *Shapiro* [3.41, 109, 168] take structures to be *ante rem* universals. In this view, structures exist independently of physical systems, yet they can be instantiated in physical systems. In this view systems instantiate structures and models are isomorphic to these instantiated structures.

This view raises all kind of metaphysical issues about the ontology of structures and the instantiation relation. Let us set aside these issues and assume that they

can be resolved in one way or another. This would still leave us with serious epistemic and semantic questions. How do we know a certain structure is instantiated in a system and how do we refer to it? Objects do not come with labels on their sleeves specifying which structures they instantiate, and proponents of structural universals face a serious problem in providing an account of *how we access* the structures instantiated by target systems. Even if – as a brute metaphysical fact – target systems only instantiate a small number of structures, and therefore there is a substantial question regarding whether or not scientific models represent them, this does not help us understand how we could ever come to know whether or not the isomorphism holds. It seems that individuating a domain of objects and identifying relations between them is the only way for us to access a structure. But then we are back to the first version of the response, and we are again faced with all the problems that it raises.

The third version of the second response is more radical. One might take target systems themselves to be structures. If this is the case then there is no problem with the idea that they can be isomorphic to a scientific model. One might expect ontic structural realists to take this position. If the world fundamentally is a structure, then there is nothing mysterious about the notion of an isomorphism between a model and the world. Surprisingly, some ontic structuralists have been hesitant to adopt such a view (see *French* and *Ladyman* [3.120, p. 113] and *French* [3.169, p. 195]). Others, however, seem to endorse it. *Tegmark* [3.170], for instance, offers an explicit defense of the idea that the world simply is a mathematical structure. He defines a seemingly moderate form of realism – what he calls the *external reality hypothesis* (ERH) – as the claim that “there exists an external physical reality completely independent of us humans” [3.170, p. 102] and argues that this entails that the world is a mathematical structure (his “mathematical universe hypothesis”) [3.170, p. 102]. His argument for this is based on the idea that a so-called *theory of everything* must be expressible in a form that is devoid of human-centric *baggage* (by the ERH), and the only theories that are devoid of such baggage are mathematical, which, strictly speaking, describe mathematical structures. Thus, since a complete theory of everything describes an external reality independent of humans, and since it describes a mathematical structure, the external reality itself *is* a mathematical structure.

This approach stands or falls on the strengths of its premise that a complete theory of everything will be formulated purely mathematically, without any human baggage, which in turn relies on a strict reductionist account of scientific knowledge [3.170, pp. 103–104]. Discussing this in any detail goes beyond our current

purposes. But it is worth noting that Tegmark’s discussion is focused on the claim that *fundamentally* the world is a mathematical structure. Even if this were the case, it seems irrelevant for many of our current scientific models, whose targets aren’t at this level. When modeling an airplane wing we don’t refer to the funda-

mental super-string structure of the bits of matter that make up the wing, and we don’t construct wing models that are isomorphic to such fundamental structures. So Tegmark’s account offers no answer to the question about where structures are to be found at the level of nonfundamental target systems.

## 3.5 The Inferential Conception

In this section we discuss accounts of scientific representation that analyze representation in terms of the inferential role of scientific models. On the previous accounts discussed, a model’s inferential capacity dropped out of whatever it was that was supposed to answer the ER-problem: proposed morphisms or similarity relations between models and their targets for example. The accounts discussed in this section build the notion of surrogative reasoning directly into the conditions on epistemic representation.

### 3.5.1 Deflationary Inferentialism

Suárez argues that we should adopt a “deflationary or minimalist attitude and strategy” [3.32, p. 770] when addressing the problem of epistemic representation. We will discuss deflationism in some detail below, but in order to formulate and discuss Suárez’s theory of representation we need at least a preliminary idea of what is meant by a deflationary attitude. In fact two different notions of deflationism are in operation in his account. The first is [3.32, p. 771]:

“abandoning the aim of a substantive theory to seek universal necessary and sufficient conditions that are met in each and every concrete real instance of scientific representation [...] necessary conditions will certainly be good enough.”

We call the view that a theory of representation should provide only necessary conditions *n*-deflationism (*n* for *necessary*). The second notion is that we should seek “no deeper features to representation other than its surface features” [3.32, p. 771] or “platitudes” [3.171, p. 40], and that we should deny that an analysis of a concept “is the kind of analysis that will shed explanatory light on our use of the concept” [3.172, p. 39]. We call this position *s*-deflationism (*s* for *surface feature*). As far as we can tell, Suárez intends his account of representation to be deflationary in both senses.

Suárez dubs the account that satisfies these criteria *inferentialism* [3.32, p. 773]:

#### Definition 3.9 Inferentialism 1

A scientific model *M* represents a target *T* only if (i) the representational force of *M* points towards *T*, and (ii) *M* allows competent and informed agents to draw specific inferences regarding *T*.

Notice that this condition is not an instantiation of the ER-scheme: in keeping with *n*-deflationism it features a material conditional rather than a biconditional and hence provides necessary (but not sufficient) conditions for *M* to represent *T*. We now discuss each condition in turn, trying to explicate in what way they satisfy *s*-deflationism.

The first condition is designed to make sure that *M* and *T* indeed enter into a representational relationship, and Suárez stresses that representational force is “necessary for any kind of representation” [3.32, p. 776]. But explaining representation in terms of representational force seems to shed little light on the matter as long as no analysis of representational force is offered. Suárez addresses this point by submitting that the first condition can be “satisfied by mere stipulation of a target for any source” [3.32, p. 771]. This might look like denotation as in Sect. 3.2. But Suárez stresses that this is not what he intends for two reasons. Firstly, he takes denotation to be a substantive relation between a model and its target, and the introduction of such a relation would violate the requirement of *s*-deflationism [3.172, p. 41]. Secondly, *M* can denote *T* only if *T* exists. Thus including denotation as a necessary condition on scientific representation “would rule out fictional representation, that is, representation of nonexistent entities” [3.32, p. 772], and “any adequate account of scientific representation must accommodate representations with fictional or imaginary targets” [3.172, p. 44].

The second issue is one that besets other accounts of representation too, in particular similarity and isomorphism accounts. The first reason, however, goes right to the heart of Suárez’s account: it makes good on the *s*-deflationary condition that nothing other than surface features can be included in an account of representation.

At a surface level one cannot explicate *representational force* at all and any attempt to specify what representational force consists in is a violation of *s*-deflationism.

The second necessary condition, that models allow competent and informed agents to draw specific inferences about their targets, is in fact just the *surrogative reasoning condition* we introduced in Sect. 3.1, now taken as a necessary condition on epistemic representation. The sorts of inferences that models allow are not constrained. Suárez points out that the condition “does not require that [*M*] allow deductive reasoning and inference; any type of reasoning inductive, analogical, abductive – is in principle allowed” [3.32, p. 773]. (The insistence on inference makes Suárez’s account an instance of what Chakravarty [3.173] calls a *functional conception* of representation.)

A problem for this approach is that we are left with no account of how these inferential rules are generated: what is it about models that allows them to license inferences about their targets, or what leads them to license some inferences and not others? Contessa makes this point most stridently when he argues that [3.29, p. 61]:

“On the inferential conception, the user’s ability to perform inferences from a vehicle [model] to a target seems to be a brute fact, which has no deeper explanation. This makes the connection between epistemic representation and valid surrogative reasoning needlessly obscure and the performance of valid surrogative inferences an activity as mysterious and unfathomable as soothsaying or divination.”

This seems correct, but Suárez can dismiss this complaint by appeal to *s*-deflationism. Since inferential capacity is supposed to be a surface-level feature of scientific representation, we are not supposed to ask for any elucidation about what makes an agent competent and well informed and how inferences are drawn.

For these reasons Suárez’s account is deflationary both in the sense of *n*-deflationism and of *s*-deflationism. His position provides us with a concept of epistemic representation that is cashed out in terms of an inexplicable notion of representational force and of an inexplicable capacity to ground inferences. This is very little indeed. It is the adoption of a deflationary attitude that allows him to block any attempt to further unpack these conditions and so the crucial question is: why should one adopt deflationism?

We turn to this question shortly. Before doing so we want to briefly outline how the above account fares with respect to the other problems introduced in Sect. 3.1. The account provides a neat explanation of the possibility of misrepresentation [3.32, p. 776]:

“part (ii) of this conception accounts for inaccuracy since it demands that we correctly draw inferences from the source about the target, but it does not demand that the conclusions of these inferences be all true, nor that all truths about the target may be inferred.”

Models represent their targets only if they license inferences about them. They represent them accurately to the extent that the conclusions of these inferences are true.

With respect to the representational demarcation problem, Suárez illustrates his account with a large range of representations, including diagrams, equations, scientific models, and nonscientific representations such as artistic portraits. He explicitly states that “if the inferential conception is right, scientific representation is in several respects very close to iconic modes of representation like painting” [3.32, p. 777] and he mentions the example of Velázquez’s portrait of Innocent X [3.32]. It is clear that the conditions of inferentialism 1 (Definition 3.9) are met by nonscientific as well as scientific epistemic representations. So, at least without sufficient conditions, there is no clear way of demarcating between the different kinds of epistemic representation.

Given the wide variety of types of representation that this account applies to, it’s unsurprising that Suárez has little to say about the ontological problem. The only constraint that inferentialism 1 (Definition 3.9) places on the ontology of models is that “[i]t requires [*M*] to have the internal structure that allows informed agents to correctly draw inferences about [*T*]” [3.32, p. 774]. And relatedly, since the account is supposed to apply to a wide variety of entities, including equations and mathematical structures, the account implies that mathematics is successfully applied in the sciences, but in keeping with the spirit of deflationism no explanation is offered about how this is possible.

Suárez does not directly address the problem of style, but a minimalist answer emerges from what he says about representation. On the one hand he explicitly acknowledges that many different kinds of inferences are allowed by the second condition in inferentialism 1 (Definition 3.9). In the passage quoted above he mentions inductive, analogical and abductive inferences. This could be interpreted as the beginning of classification of representational styles. On the other hand, Suárez remains silent about what these kinds are and about how they can be analyzed. This is unsurprising because spelling out what these inferences are, and what features of the model ground them, would amount to giving a substantial account, which is something Suárez wants to avoid.

Let us now return to the question about the motivation for deflationism. As we have seen, a commitment to deflationism about the concept is central to Suárez's approach to scientific representation. But deflationism comes in different guises, which Suárez illustrates by analogy with deflationism with respect to truth. Suárez [3.172] distinguishes between the *redundancy* theory (associated with Frank Ramsey and also referred to as the *no theory* view), *abstract minimalism* (associated with Crispin Wright) and the *use theory* (associated with Paul Horwich). What all three are claimed to have in common is that they accept the disquotational schema – i. e., instances of the form:  $P$  is true iff  $P$ . Moreover they [3.172, p. 37]

“either do not provide an analysis in terms of necessary and sufficient conditions, or if they do provide such conditions, they claim them to have no explanatory purchase.”

He claims that the redundancy theory of truth is characterized by the idea that [3.172, p. 39]:

“the terms *truth* and *falsity* do not admit a theoretical elucidation or analysis. But that, since they can be eliminated in principle – if not in practice – by disquotation, they do not in fact require such an analysis.”

So, as Suárez characterizes the position, the redundancy theory denies that any necessary and sufficient conditions for application of the truth predicate case be given. He argues that [3.172]:

“the generalization of this *no-theory theory* for any given putative concept  $X$  is the thought that  $X$  neither possesses nor requires necessary and sufficient conditions because it is not in fact a *genuine*, explanatory or substantive concept.”

This motivates  $n$ -deflationism (although one might ask why such a position would allow even necessary conditions. Suárez doesn't discuss this).

This approach faces a number of challenges. First, the argument is based on the premise that if deflationism is good for truth it must be good for representation. This premise is assumed tacitly. There is, however, a question whether the analogy between truth and representation is sufficiently robust to justify subjecting them to the same theoretical treatment. Surprisingly, Suárez offers little by way of explicit argument in favor of any sort of deflationary account of epistemic representation. In fact, the natural analogue of the linguistic notion of truth is accurate epistemic representation, rather than epistemic representation itself, which may be more appropriately compared with linguistic meaning. Second, the argument insinuates that deflationism is the cor-

rect analysis of truth. This, however, is far from an established fact. Different positions are available in the debate and whether deflationism (or any specific version of it) is superior to other proposals remains a matter of controversy (see, for instance, Kühne [3.174]). But as long as it's not clear that deflationism about truth is a superior position, it's hard to see how one can muster support for deflationism about representation by appealing to deflationism about truth.

Moreover, a position that allows only necessary conditions on epistemic representation faces a serious problem. While such an account allows us to *rule out* certain scenarios as instances of epistemic representation (for example a proper name doesn't allow for a competent and well informed language user to draw any specific inferences about its bearer and Callender and Cohen's salt shaker doesn't allow a user to draw any specific inferences about Madagascar), the lack of sufficient conditions doesn't allow us to *rule in* any scenario as an instance of epistemic representation. So on the basis of inferentialism 1 (Definition 3.9) we are never in position to assert that a particular model actually is a representation, which is an unsatisfactory situation.

The other two deflationary positions in the debate over truth are abstract minimalism and the use theory. Suárez characterizes the use theory as being based on the idea that “truth is nominally a property, although not a substantive or explanatory one, which is essentially defined by the platitudes of its use of the predicate in practice” [3.172, p. 40]. Abstract minimalism is presented as the view that while truth is [3.172, p. 40]:

“legitimately a property, which is abstractly characterized by the platitudes, it is a property that cannot explain anything, in particular it fails to explain the norms that govern its very use in practice.”

Both positions imply that necessary and sufficient conditions for truth *can* be given [3.172]. But on either account, such conditions only capture nonexplanatory surface features. This motivates  $s$ -deflationism.

Since  $s$ -deflationism explicitly allows for necessary and sufficient conditions, inferentialism 1 (Definition 3.9) can be extended to an instance of the ER-scheme, providing necessary and sufficient conditions (which also seems to be in line with Suárez and Solé [3.171, p. 41] who provide a formulation of inferentialism with a biconditional):

#### **Definition 3.10 Inferentialism 2**

A scientific model  $M$  represents a target  $T$  iff (i) the representational force of  $M$  points towards  $T$ , and (ii)  $M$  allows competent and informed agents to draw specific inferences regarding  $T$ .



If one takes conditions (i) and (ii) to refer to “features of activates within a normative practice, [that] do not stand for relations between sources and targets” [3.172, p. 46], then we arrive at a *use-based* account of epistemic representation. In order to understand a particular instance of a model  $M$  representing a target  $T$  we have to understand how scientists go about establishing that  $M$ 's representational force points towards  $T$ , and the inferential rules, and particular inferences from  $M$  to  $T$ , they use and make.

Plausibly, such a focus on practice amounts to looking at the inferential rules employed in each instance, or type of instance, of epistemic representation. This, however, raises a question about the status of any such analysis vis-à-vis the general theory of representation as given in inferentialism 2 (Definition 3.10). There seem to be two options. The first is to affirm inferentialism 2's (Definition 3.10) status as an exhaustive theory of representation. This, however, would imply that any analysis of the workings of a particular model would fall outside the scope of a theory of representation because any attempt to address Contessa's objection would push the investigation outside the territory delineated by *s*-deflationism. Such an approach seems to be overly purist. The second option is to understand inferentialism 2 (Definition 3.10) as providing abstract conditions that require concretization in each instance of epistemic representation (abstraction can here be understood, for instance, in *Cartwright's* [3.74] sense). Studying the concrete realizations of the abstract conditions is then an integral part of the theory. This approach seems plausible, but it renders deflationism obsolete. Thus understood, the view becomes indistinguishable from a theory that accepts the *surrogate reasoning condition* and the *requirement of directionality* as conditions of adequacy and analyzes them in pluralist spirit, that is, under the assumption that these conditions can have different concrete realizers in different contexts. But this program can be carried out without ever mentioning deflationism.

One might reply that the first option unfairly stacks the deck against inferentialism and point out that different inferential practices *can* be studied within the inferentialist framework. One way of making good on this idea would be to submit that the inferences from models to their targets should be taken as conceptually basic, denying that they need to be explained; in particular, denying that they need to be grounded by any (possibly varying) relation(s) that might hold between models and their targets. Such an approach is inspired by Brandom's inferentialism in the philosophy of language where the central idea is to reverse the order of explanation from representational notions – like truth and reference – to inferential notions – such as the va-

lidity of argument [3.175, 176]. Instead, we are urged to begin from the inferential role of sentences (or propositions, or concepts, and so on) – that is the role that they play in providing reasons for other sentences (or propositions etc.), and having such reasons provided for them – and from this reconstruct their representational aspects.

Such an approach is developed by *de Donato Rodríguez* and *Zamora Bonilla* [3.177] and seems like a fruitful route for future research, but for want of space we will not discuss it in detail here. There is no evidence that Suárez would endorse such an approach. And, more worrying for inferentialism 2 (Definition 3.10), it is not clear whether such an approach would satisfy *s*-deflationism. Each investigation into the inferential rules utilized in each instance, or type of instance of epistemic representation will likely be a substantial (possibly sociological or anthropological) project. Thus the *s*-deflationary credentials of the approach – at least if they are taken to require that nothing substantial can be said about scientific representation in each instance, as well as in general – are called into question.

Finally, if the conditions in inferentialism 2 (Definition 3.10) are taken to be abstract platitudes then we arrive at an abstract minimalism. Although inferentialism 2 (Definition 3.10) defines the concept of epistemic representation, the definition does not suffice to explain the use of any particular instance of epistemic representation for ([3.172, p. 48], cf. [3.171]):

“on the abstract minimalism here considered, to apply this notion to any given concrete case of representation requires that some additional relation obtains between  $[M]$  and  $[T]$ , or a property of  $[M]$  or  $[T]$ , or some other application condition.”

Hence, according to this approach representational force and inferential capacity are taken to be abstract platitudes that suffice to define the concept of scientific representation. However, because of their level of generality, they fail to explain any particular instance of it. To do this requires reference to additional features that vary from case to case. These other conditions can be “isomorphism or similarity” and they “would need to obtain in each concrete case of representation” ([3.171, p. 45], [3.32, p. 773], [3.172, p. 43]). These extra conditions are called the *means* of representation, the relations that scientists exploit in order to draw inferences about targets from their models, and are to be distinguished from conditions (i) and (ii), the *constituents* of representation, that define the concept ([3.23, p. 230], [3.171, p. 43], [3.172, p. 46], [3.178, pp. 93–94]). We are told that the means cannot be reduced to the constituents but that [3.171, p. 43]:

“all representational means (such as isomorphism and similarity) are concrete instantiations, or realizations, of one of the basic platitudes that constitute representation”

and that “there can be no application of representation without the simultaneous instantiation of a particular set of properties of  $[M]$  and  $[T]$ , and their relation” [3.171, p. 44].

Such an approach amounts to using conditions (i) and (ii) to answer the ER-problem, but again with the caveat that they are abstract conditions that require concretization in each instance of epistemic representation. In this sense it is immune to Contessa’s objection about the mysterious capacity that models have to license about their targets. They do so in virtue of more concrete relations that hold between models and their targets, albeit relations that vary from case to case. The key question facing this account is to fill in the details about what sort of relations concretize the abstract conditions. But we are now facing a similar problem as the above. Even if *s*-deflationism applies to epistemic representation in general, an investigation into each specific instance of will involve uncovering substantial relations that hold between models and their targets, which again conflicts with Suárez’s adherence to a deflationist approach.

### 3.5.2 Inflating Inferentialism: Interpretation

In response to difficulties like the above Contessa claims that “it is not clear why we should adopt a deflationary attitude *from the start*” [3.29, p. 50] and provides a “interpretational account” of scientific representation that is still, at least to some extent, inspired by Suárez’s account, but without being deflationary. Contessa claims [3.29, p. 48]:

“[t]he main difference between the interpretational conception [...] and Suárez’s inferential conception is that the interpretational account is a substantial account – interpretation is not just a ‘symptom’ of representation; it is what makes something an epistemic representation of a something else.”

To explain in virtue of what the inferences can be drawn, Contessa introduces the notion of an *interpretation* of a model, in terms of its target system as a necessary and sufficient condition on epistemic representation ([3.29, p. 57], [3.179, pp. 126–127]):

#### **Definition 3.11 Interpretation**

A scientific model  $M$  is an epistemic representation of a certain target  $T$  (for a certain user) if and only if the user adopts an interpretation of  $M$  in terms of  $T$ .

Contessa offers a detailed formal characterization of an interpretation, which we cannot repeat here for want of space (see [3.29, pp. 57–62] for details). The leading idea is that the model user first identifies a set of relevant objects in the model, and a set of properties and relations these objects instantiate, along with a set of relevant objects in the target and a set of properties and relations these objects instantiate. The user then:

1. Takes  $M$  to denote  $T$ .
2. Takes every identified object in the model to denote exactly one object in the target (and every relevant object in the target has to be so denoted and as a result there is a one-to-one correspondence between relevant objects in the model and relevant objects in the target).
3. Takes every property and relation in the model to denote a property or relation of the same arity in the target (and, again, and every property and relation in the target has to be so denoted and as a result there will be one-to-one correspondence between relevant properties and relations in the model and target).

A formal rendering of these conditions is what Contessa calls an *analytic interpretation* (he also includes an additional condition pertaining to functions in the model and target, which we suppress for brevity). The relationship between interpretations and the surrogative reasoning mentioned above is that it is in virtue of the user adopting an analytic interpretation that a model licenses inferences about its target.

At first sight Contessa’s interpretation may appear to be equivalent to setting up an isomorphism between model and target. This impression is correct in as far as an interpretation requires that there be a one-to-one correspondence between relevant elements and relations in the model and the target. However, unlike the isomorphism view, Contessa’s interpretations are not committed to models being structures, and relations can be interpreted as full-fledged relations rather than purely extensionally specified sets of tuples.

Interpretation (Definition 3.11) is a nondeflationary account of scientific representation: most (if not all) instances of scientific representation involve a model user adopting an analytic interpretation towards a target. The capacity for surrogative reasoning is then seen as a symptom of the more fundamental notion of a model user adopting an interpretation of a model in terms of its target. For this reason the adoption of an analytical interpretation is a substantial sufficient condition on establishing the representational relationship. Contessa focuses on the sufficiency of analytic interpretations rather than their necessity and adds that he does [3.29, p. 58]

“not mean to imply that all interpretation of vehicles [models] in terms of the target are necessarily analytic. Epistemic representations whose standard interpretations are not analytic are at least conceivable.”

Even with this in mind, it is clear that he intends that *some* interpretation is a necessary condition on epistemic representation.

Let’s now turn to how interpretation fares with respect to our questions for an account of epistemic representation as set out in Sect. 3.2. Modulo the caveat about nonanalytical interpretations, interpretation (Definition 3.11) provides necessary and sufficient conditions on epistemic representation and hence answers the ER-problem. Furthermore, it does so in a way that explains the directionality of representation: interpreting a model in terms of a target does not entail interpreting a target in terms of a model.

Contessa does not comment on the applicability of mathematics but since his account shares with the structuralist account an emphasis on relations and one-to-one model-target correspondence, Contessa can appeal to the same account of the applicability of mathematics as structuralist.

With respect to the demarcation problem, *Contessa* is explicit that “[p]ortraits, photographs, maps, graphs, and a large number of other representational devices” perform inferential functions [3.29, p. 54]. Since nothing in the notion of an interpretation seems restricted to scientific models, it is plausible to regard interpretation (Definition 3.11) as a universal theory of epistemic representation (a conclusion that is also supported by the fact that *Contessa* [3.29] uses the example of the London Underground map to motivate his account; see also [3.179]). As such, interpretation (Definition 3.11) seems to deny the existence of a substantial distinction between scientific and nonscientific epistemic representations (at least in terms of their representational properties). It remains unclear how interpretation (Definition 3.11) addresses the problem of style. As we have seen earlier, in particular visual representations fall into different categories. It is a question for future research how these can be classified within the interpretational framework.

With respect to the question of ontology, interpretation (Definition 3.11) itself places few constraints on what scientific models are, ontologically speaking. All it requires is that they consist of objects, properties, relations, and functions. For this reason our discussion in Sect. 3.3.3 above rears its head again here. As before, how to apply interpretation (Definition 3.11) to physical models can be understood relatively easily. But how to apply it to nonphysical models is less straightforward.

*Contessa* [3.180] distinguishes between mathematical models and fictional models, where fictional models are taken to be fictional objects. We briefly return to his ontological views in Sect. 3.6.

In order to deal with the possibility of misrepresentation, *Contessa* notes that “a user does not need to believe that every object in the model denotes some object in the system in order to interpret the model in terms of the system” [3.29, p. 59]. He illustrates this claim with an example of contemporary scientists using the Aristotelian model of the cosmos to represent the universe, pointing out that “in order to interpret the model in terms of the universe, we do not need to assume that the sphere of fixed stars itself [...] denotes anything in the universe” [3.29].

From this example it is clear that the relevant sets of objects, properties and functions isolated in the construction of the analytic interpretation do not need to exhaust the objects, properties, relations, and functions of either the model or the target. The model user can identify a relevant *proper* subset in each instance. This allows interpretation (Definition 3.11) to capture the common practice of abstraction in scientific models: a model need only represent some features of its target, and moreover, the model may have the sort of *surplus* features are not taken to represent anything in the target, i. e., that not all of a model’s features need to play a direct representational role.

This suggestion bears some resemblance to partial structures, and it suffers from the same problem too. In particular distortive idealisations are a source of problems for interpretation (Definition 3.11), as several commentators have observed (see *Shech* [3.181] and *Bolinska* [3.28]). *Contessa* is aware of this problem and illustrates it with the example of a massless string. His response to the problem is to appeal to a user’s corrective abilities [3.29, p. 60]:

“since models often misrepresent some aspect of the system or other, it is usually up to the user’s competence, judgment, and background knowledge to use the model successfully in spite of the fact that the model misrepresents certain aspects of the system.”

This is undoubtedly true, but it is unclear how such a view relates, or even derives from, interpretation (Definition 3.11). An appeal to the competence of users seems to be an ad hoc move that has no systematic grounding in the idea of an interpretation, and it is an open question how the notion of an interpretation could be amended to give distortive idealizations a systematic place.

*Ducheyne* [3.182] provides a variant of interpretation (Definition 3.11) that one might think could be used

to accommodate these distortive idealizations. The details of the account, which we won't state precisely here for want of space, can be found in [3.182, pp. 83–86]. The central idea is that each relevant relation specified in the interpretation holds precisely in the model, and corresponds to the same relation that holds only approximately (with respect to a given purpose) in the target. For example, the low mass of an actual pendulum's string approximates the masslessness of the string in the model. The one-to-one correspondence between (relevant) objects and relations in the model and target is retained, but the notion of a user taking relations in the model to denote relations in the target is replaced with the idea that the relations in the target are approximations of the ones they correspond to. Ducheyne calls this the *pragmatic limiting case* account of scientific representation (the pragmatic element comes from the fact that the level of approximation required is determined by the purpose of the model user).

However, if this account is to succeed in explaining how distortive idealizations are scientific representations, then more needs to be said about how a target relation can *approximate* a model relation. Ducheyne implicitly relies on the fact that relations are such that “we can determine *the extent to which* [they hold] empirically” [3.182, p. 83] (emphasis added). This suggests that he has quantifiable relations in mind, and that what it means for a relation  $r$  in the target to approximate a relation  $r'$  in the model is a matter of comparing numerical values, where a model user's purpose determines how close they must be if the former is to count as an approximation of the latter. But whether this exhausts the ways in which relations can be approximations remains unclear. Hendry [3.183], Laymon [3.184], Liu [3.185], Norton [3.186], and Ramsey [3.187], among others, offer discussions of different kinds of idealizations and approximations, and Ducheyne would have to make it plausible that all these can be accommodated in his account.

More importantly, Ducheyne's account has problems dealing with misrepresentations. Although it is designed to capture models that misrepresent by being approximations of their targets, it remains unclear how it deals with models that are outright mistaken. For example, it seems a stretch to say that Thomson's model of the atom (now derogatively referred to as the *plum pudding model*) is an approximation of what the quantum mechanical shell model tells us about atoms, and it seems unlikely that there is a useful sense in which the relations that hold between electrons in Thomson's model *approximate* those that hold in reality. But this does not mean that it is not a scientific representation of the atom; it's just an incorrect one. It does not seem to

be the case that all cases of scientific misrepresentation are instances where the model is an approximation of the target (or even conversely, it is not clear whether all instances of approximation need to be considered cases of *misrepresentation* in the sense that they license falsehoods about their targets).

### 3.5.3 The Denotation, Demonstration, and Interpretation Account

Our final account is *Hughes' denotation, demonstration, and interpretation* (DDI) account of scientific representation [3.188] and [3.189, Chap. 5]. This account has inspired both the inferential (see Suárez [3.32, p. 770] and [3.172]) and the interpretational account (see Contessa [3.179, p. 126]) discussed in this section.

Quoting directly from Goodman [3.64, p. 5], Hughes takes a model of a physical system to “be a symbol for it, stand for it, refer to it” [3.188, p. 330]. Presumably the idea is that a model denotes its target in the same way that a proper name denotes its bearer, or, stretching the notion of denotation slightly, a predicate denote elements in its extension. (Hughes [3.188, p. 330] notes that there is an additional complication when the model has multiple targets but this is not specific to the DDI account and is discussed in more detail in Sect. 3.8). This is the first *D* in *DDI*. What makes models epistemic representations and thereby distinguishes them from proper names, are the demonstration and interpretation conditions.

The demonstration condition, the second *D* in *DDI*, relies on a model being “a secondary subject that has, so to speak, a life of its own. In other words, [a] representation has an internal dynamic whose effects we can examine” [3.188, p. 331] (that models have an *internal dynamic* is all that Hughes has to say about the problem of ontology). The two examples offered by Hughes are both models of what happens when light is passed through two nearby slits. One model is mathematical where the internal dynamics are “supplied by the deductive, resources of the mathematics they employ” [3.188], the other is a physical ripple chamber where they are supplied by “the natural processes involved in the propagation of water waves” [3.188, p. 332].

Such demonstrations, on either mathematical models or physical models are still primarily about the models themselves. The final aspect of Hughes' account – the *I* in *DDI* – is interpretation of what has been demonstrated in the model in terms of the target system. This yields the predictions of the model [3.188, p. 333]. Unfortunately Hughes has little to say about what it means to interpret a result of a demonstration on a model in terms of its target system, and so one has

to retreat to an intuitive (and unanalyzed) notion of carrying over results from models to targets.

Now Hughes is explicit that he is not attempting to answer the ER-problem, and that he does not even offer denotation, demonstration and interpretation as individually necessary and jointly sufficient conditions for scientific representation. He prefers the more [3.188, p. 339]

“modest suggestion that, if we examine a theoretical model with these three activities in mind, we shall achieve some insight into the kind of representation that it provides.”

We are not sure how to interpret Hughes’ position in light of this. On one reading, he can be seen as describing how we *use* models. As such, *DDI* functions as a diachronic account of what a model user does when using a model in an attempt to learn about a target system. We first stipulate that the model stands for the target, then prove what we want to know, and finally *transfer* the results obtained in the model back to the target. Details aside, this picture seems by and large correct. The problem with the *DDI* account is that it does not explain why and how this is possible. Under what conditions is it true that the model denotes the target? What kinds of things are models that they allow for demonstrations? How does interpretation work; that is, how can results obtained in the model be transferred to the target? These are questions an account of epistemic representation has to address, but which are left unanswered by the *DDI* account thus interpreted. Accordingly, *DDI* provides an answer to a question distinct from the ER-problem. Although a valuable answer to the question of how models are used, it does not help us too much here, since it presupposes the very representational relationship we are interested in between models and their targets.

An alternative reading of Hughes’ account emerges when we consider the developments of the structuralist and similarity conceptions discussed previously, and the discussion of deflationism in Sect. 3.5.1: perhaps the very act of using a model, with all the user intentions and practices that brings with it, constitutes the epistemic representation relationship itself. And as such, perhaps the *DDI* conditions could be taken as an answer to the ER-problem:

### 3.6 The Fiction View of Models

In this section we discuss a number of recent attempts to analyze scientific modeling by drawing an analogy with literary fiction. We begin by introducing the leading ideas and differentiating between different strands

#### Definition 3.12 *DDI-ER*

A scientific model  $M$  represents a target  $T$  iff  $M$  denotes  $T$ , an agent (or collection of thereof)  $S$  exploits the internal dynamic of  $M$  to make demonstrations  $D$ , which in turn are interpreted by the agent (or collection of thereof) to be about  $T$ .

This account comes very close to interpretation (Definition 3.11) as discussed in Sect. 3.5.2. And as such it serves to answer the questions we set out in Sect. 3.1 above in the same way. But in this instance, the notion of what it means to *exploit an internal dynamic* and *interpret the results* of this to be about  $T$  need further explication. If the notion of an interpretation is cashed out in the same way as Contessa’s analytic interpretation, then the account will be vulnerable to the same issues as those discussed previously. In another place *Hughes* endorses Giere’s semantic view of theories, which he characterizes as connecting models to the target with a theoretical hypothesis [3.190, p. 121]. This suggests that an interpretation is a theoretical hypothesis in this sense. If so, then Hughes’s account collapses into a version of Giere’s.

Given that *Hughes* describes his account as “designedly skeletal [and in need] to be supplemented on a case-by-case basis” [3.188, p. 335], one option available is to take the demonstration and interpretation conditions to be abstract (in the sense of abstract minimalism discussed above), which require filling in each instance, or type of instance, of epistemic representation. As *Hughes* notes, his examples of the internal dynamics of mathematical and physical models are radically different with the demonstrations of the former utilizing mathematics, and the latter physical properties such as the propagation of water waves. Similar remarks apply to the interpretation of these demonstrations, as well as to denotation. But as with Suárez’s account, the definition sheds little light on the problem at hand as long as no concrete realizations of the abstract conditions are discussed. Despite *Hughes*’ claims to the contrary, such an account could prove a viable answer the ER-problem, and it seems to capture much of what is valuable about both the abstract minimalist version of inferentialism 2 (Definition 3.10) as well as interpretation (Definition 3.11) discussed above.

of argument. We then examine a number of accounts that analyze epistemic representation against the backdrop of literary fiction. We finally discuss criticisms of the fiction view.

### 3.6.1 Models and Fiction

Scientific discourse is rife with passages that appear to be descriptions of systems in a particular discipline, and the pages of textbooks and journals are filled with discussions of the properties and the behavior of those systems. Students of mechanics investigate at length the dynamical properties of a system consisting of two or three spinning spheres with homogeneous mass distributions gravitationally interacting only with each other. Population biologists study the evolution of one species that reproduces at a constant rate in an unchanging environment. And when studying the exchange of goods, economists consider a situation in which there are only two goods, two perfectly rational agents, no restrictions on available information, no transaction costs, no money, and dealings are done immediately. Their surface structure notwithstanding, no one would mistake descriptions of such systems as descriptions of an *actual* system: we know very well that there are no such systems (of course some models are actual systems – a scale model of a car in a wind tunnel for example – but in this section we focus on models that are not of this kind). Scientists sometimes express this fact by saying that they talk about *model land* (for instance [3.191, p.135]).

*Thomson-Jones* [3.98, p. 284] refers to such a description as a “description of a missing system”. These descriptions are embedded in what he calls the “face value practice” [3.98, p. 285]: the practice of talking and thinking about these systems as if they were real. We observe that the amplitude of an ideal pendulum remains constant over time in much the same way in which we say that the Moon’s mass is approximately  $7.34 \times 10^{22}$  kg. Yet the former statement is about a point mass suspended from a massless string – and there is no such thing in the world.

The face value practice raises a number of questions. What account should be given of these descriptions and what sort of objects, if any, do they describe? How should we analyze the face value practice? Are we putting forward truth-evaluable claims when putting forward descriptions of missing systems? An answer to these questions emerges from the following passage by *Peter Godfrey-Smith* [3.152, p. 735]:

“[...] I take at face value the fact that modelers often *take* themselves to be describing imaginary biological populations, imaginary neural networks, or imaginary economies. [...] Although these imagined entities are puzzling, I suggest that at least much of the time they might be treated as similar to something that we are all familiar with, the imagined objects of literary fiction. Here I have in

mind entities like Sherlock Holmes’ London, and Tolkein’s Middle Earth. [...] the model systems of science often work similarly to these familiar fictions.”

This is the core of the fiction view of models: models are akin to places and characters in literary fiction. When modeling the solar system as consisting of ten perfectly spherical spinning tops physicists describe (and *take themselves* to describe) an imaginary physical system; when considering an ecosystem with only one species biologists describe an imaginary population; and when investigating an economy without money and transaction costs economists describe an imaginary economy. These imaginary scenarios are tellingly like the places and characters in works of fiction like *Madame Bovary* and *Sherlock Holmes*.

Although hardly at the center of attention, the parallels between certain aspects of science and literary fiction have not gone unnoticed. Maxwell discussed in great detail the motion of “a purely imaginary fluid” in order to understand the electromagnetic field [3.192, pp. 159–160]. The parallel between science and fiction occupied center stage in *Vaihinger’s* [3.193] philosophy of the *as if*. More recently, the parallel has also been drawn specifically between models and fiction. *Cartwright* observes that “a model is a work of fiction” [3.194, p. 153] and later suggests an analysis of models as fables [3.73, Chap. 2]. *McCloskey* [3.195] emphasises the importance of narratives and stories in economics. *Fine* notes that modeling natural phenomena in every area of science involves fictions in *Vaihinger’s* sense [3.196, p. 16], and *Sklar* highlights that describing systems *as if* they were systems of some other kind is a royal route to success [3.197, p. 71]. *Elgin* [3.198, Chap. 6] argues that science shares important epistemic practices with artistic fiction. *Hartmann* [3.199] and *Morgan* [3.200] emphasize that stories and narratives play an important role in models, and *Morgan* [3.201] stresses the importance of imagination in model building. *Sugden* [3.202] points out that economic models describe “counterfactual worlds” constructed by the modeler. *Frigg* [3.30, 203] suggests that models are imaginary objects, and *Grüne-Yanoff* and *Schweitzer* [3.204] emphasize the importance of stories in the application of game theory. *Toon* [3.48, 205] has formulated an account of representation based on a theory of literary fiction. *Contessa* [3.180] provides a fictional ontology of models and *Levy* [3.43, 206] discusses models as fictions.

But simply likening modeling to fiction does not solve philosophical problems. Fictional discourse and fictional entities face well-known philosophical questions, and hence explaining models in terms of fictional

characters seems to amount to little more than to explain *obscurum per obscurius*. The challenge for proponents of the fiction view is to show that drawing an analogy between models and fiction has heuristic value.

A first step towards making the analogy productive is to get clear on what the problem is that the appeal to fiction is supposed to solve. This issue divides proponents of the fiction view into two groups. Authors belonging to the first camp see the analogy with fiction as providing an answer to the problem of ontology. Models, in that view, are *ontologically* on par with literary fiction while there is no productive parallel between models and fiction as far as the ER-problem (or indeed any other problem of representation) is concerned. Authors belonging to the second group hold the opposite view. They see the analogy with fiction first and foremost as providing an answer to the ER-problem (although, as we have seen, this may place restrictions on the ontology of models). Scientific representation, in this view, has to be understood along the lines of how literary fiction relates to reality. Positions on ontology vary. Some authors in this group also adopt a fiction view of ontology; some remain agnostic about the analogy's contribution to the matters of ontology; and some reject the problem of ontology altogether.

This being a review of models and representation, we refer the reader to *Gelfert's* contribution to this book for an in-depth discussion of the ontology of models, Chap. 1, and focus on the fiction view's contribution to semantics. Let us just note that those who see fiction as providing an ontology of models are spoiled for choice. In principle every option available in the extensive literature on fiction is a candidate for an ontology of models; for reviews of these options see *Friend* [3.207] and *Salis* [3.208]. Different authors have made different choices, with proposals being offered by *Contessa* [3.180], *Ducheyne* [3.72], *Frigg* [3.203], *Godfrey-Smith* [3.209], *Levy* [3.43], and *Sugden* [3.210]. *Cat* [3.211], *Liu* [3.212, 213], *Pincock* [3.214, Chap. 12], *Thomson-Jones* [3.98] and *Toon* [3.205] offer critical discussions of some of these approaches.

Even if these ontological problems were settled in a satisfactory manner, we would not be home and dry yet. *Vorms* [3.215, 216] argues that what's more important than the entity itself is the format in which the entity is presented. A fiction view that predominantly focuses on understanding the fictional entities themselves (and, once this task is out of the way, their relation to the real-world targets), misses an important aspect, namely how agents draw inferences from models. This, *Vorms* submits, crucially depends on the format under which they are presented to scientists, and

different formats allow scientists to draw different inferences. This ties in with *Knuuttila's* insistence that we ought to pay more attention to the “medium of representation” when studying models [3.9, 217].

One last point stands in need of clarification: the meaning of the term *fiction*. Setting aside subtleties that are irrelevant to the current discussion, the different uses of *fiction* fall into two groups: fiction as falsity and fiction as imagination [3.218]. Even though not mutually exclusive, the senses should be kept separate. The first use of *fiction* characterizes something as deviating from reality. We brand Peter's account of events a fiction if he does not report truthfully how things have happened. In the second use, *fiction* refers to a kind of literature, *literary fiction*. Rife prejudice notwithstanding, the defining feature of literary fiction is not falsity. Neither is everything that is said in, say, a novel untrue (novels like *War and Peace* contain correct historical information); nor does every text containing false reports qualify as fiction (a wrong news report or a faulty documentary do not by that token turn into fiction – they remain what they are, namely wrong factual statements). What makes a text fictional is the attitude that the reader is expected to adopt towards it. When reading a novel we are not meant to take the sentences we read as reports of fact; rather we are supposed to imagine the events described.

It is obvious from what has been said so far that the fiction view of models invokes the second sense of *fiction*. Authors in this tradition do not primarily intend to brand models as false; they aim to emphasize that models are presented as something to ponder. This is not to say the first sense of fiction is irrelevant in science. Traditionally fictions in that sense have been used as calculational devices for generating predictions, and recently *Bokulich* [3.14] emphasized the explanatory function of fictions. The first sense of fiction is also at work in philosophy where antirealist positions are described as fictionalism. For instance, someone is a fictionalist about numbers if she thinks that numbers don't exist (see *Kalderon* [3.219] for a discussion of several fictionalisms of this kind). Scientific antirealists are fictionalists about many aspects of scientific theories, and hence *Fine* characterizes fictionalism as an “antirealist position in the debate over scientific realism” [3.196, 220, 221], a position echoed in *Winsberg* [3.222] and *Suárez* [3.223]. *Morrison* [3.224] and *Purves* [3.225] and offer critical discussions of this approach, which the latter calls fiction as “truth conducive falsehood” [3.225, p. 236]; *Woods* [3.226] offers a critical assessment of fictionalism in general. Although there are interesting discussions to be had about the role that this kind of fictions play in the philosophy of science, it is not our interest here.

### 3.6.2 Direct Representation

In this subsection and the next we discuss proposals that have used the analogy between models and fiction to elucidate representation.

Most theories of representation we have encountered so far posit that there are model systems and construe epistemic representation as a relation between two entities, the model system and the target system. *Toon* calls this the *indirect* view of representation [3.205, p. 43]; *Levy*, speaking specifically about the fiction view of models, refers to it as the *whole-cloth fiction* view [3.206, p. 741]. Indeed, *Weisberg* views this indirectness as the defining feature of modeling [3.227]. This view faces the problem of ontology because it has to say what kind of things model systems are. This view contrasts with what *Toon* [3.205, p. 43] and *Levy* [3.43, p. 790] call a *direct* view of representation (*Levy* [3.206, p. 741] earlier also referred to it as the *worldly fiction* view). This view does not recognize model systems and aims instead to explain epistemic representation as a form of direct description. Model descriptions (like the description of an ideal pendulum) provide an “imaginative description of real things” [3.206, p. 741] such as actual pendula, and there is no such thing as a model system of which the pendulum description is literally true [3.205, pp. 43–44]. In what follows we use *Toon*’s terminology and refer to this approach as *direct representation*.

*Toon* and *Levy* both reject the indirect approach because of metaphysical worries about fictional entities, and they both argue that the direct view has the considerable advantage that it does not have to deal with the vexed problem of the ontology of model systems and their comparison with real things at all. *Levy* [3.43, p. 790] sees his approach as “largely complimentary to *Toon*’s”. So we first discuss *Toon*’s approach and then turn to *Levy*’s.

*Toon* [3.48, 205, 228] takes as his point of departure *Walton*’s [3.229] theory of representation in the arts. At the heart of this theory is the notion of a game of make believe. The simplest examples of these games are children’s plays [3.229, p. 11]. In one such play we imagine that stumps are bears and if we spot a stump we imagine that we spot a bear. In *Walton*’s terminology the stumps are *props*, and the rule that we imagine a bear when we see a stump is a *principle of generation*. Together a prop and a principle of generation prescribe what is to be imagined. If a proposition is so prescribed to be imagined, then the proposition is *fictional* in the relevant game. The term *fictional* has nothing to do with falsity; on the contrary, it indicates that the proposition is *true in the game*. The set of propositions actually imagined by someone need not coincide with the set

of all fictional propositions in game. It could be the case that there is a stump somewhere that no one has seen and hence no one imagines that it’s a bear. Yet the proposition that the unseen stump is a bear is fictional in the game.

*Walton* considers a vast variety of different props. In the current context two kinds of props are particularly important. The first are objects like statues. Consider a statue showing Napoleon on horseback [3.205, p. 37]. The statue is the prop, and the games of make believe for it are governed by certain principles of generation that apply to statues of this kind. So when seeing the statue we are mandated to imagine, for instance, that Napoleon has a certain physiognomy and certain facial expressions. We are not mandated to imagine that Napoleon was made of bronze, or that he hasn’t moved for more than 100 years.

The second important kind of props are works of literary fiction. In this case the text is the prop, which together with principles of generation appropriate for literary fictions of a certain kind, generates fictional truths by prescribing readers to imagine certain things. For instance, when reading *The War of the Worlds* [3.205, p. 39] we are prescribed to imagine that the dome of St Paul’s Cathedral has been attacked by aliens and now has a gaping hole on its western side.

In *Walton*’s theory something is a *representation* if it has the social function of serving as a prop in a game of make believe, and something is an *object of a representation* if the representation prescribes us to imagine something about the object [3.229, pp. 35, 39]. In the above examples the statue and the written text are the props, and Napoleon and St Paul’s Cathedral, respectively, are the objects of the representations.

The crucial move now is to say that models are props in games of make believe. Specifically, material models – such as an architectural model of the Forth Road Bridge – are like the statue of Napoleon [3.205, p. 37]: the model is the prop and the bridge is the object of the representation. The same observation applies to theoretical models, such as a mechanical model of a bob bouncing on a spring. The model portrays the bob as a point mass and the spring as perfectly elastic. The model description represents the real ball and spring system in the same way in which a literary text represents its objects [3.205, pp. 39–40]: the model description prescribes imaginings about the real system – we are supposed to imagine the real spring as perfectly elastic and the bob as a point mass.

We now see why *Toon*’s account is a direct view of modeling. Theoretical model descriptions represent actual concrete objects: the Forth Road Bridge and the bob on a spring. There is no intermediary en-



tivity of which model descriptions are literally true and which are doing the representing. Models prescribe imaginings about a real world target, and that is what representation consists in.

This is an elegant account of representation, but it is not without problems. The first issue is that it does not offer an answer to the ER-problem. Imagining that the target has a certain feature does not tell us how the imagined feature relates to the properties the target actually has, and so there is no mechanism to transfer model results to the target. Imagining the pendulum bob to be a point mass tells us nothing about which, if any, claims about point masses are also true of the real bob. *Toon* mentions this problem briefly. His response is that [3.205, pp. 68–69]:

“Principles of generation often link properties of models to properties of the system they represent in a rather direct way. If the model has a certain property then we are to imagine that system does too. If the model is accurate, then the model and the system will be similar in this respect. [...] [But] not all principles of generation are so straightforward. [...] In some cases similarity seems to play no role at all.”

In as far as the transfer mechanism is similarity, the view moves close to the similarity view, which brings with it both some of the benefits and the problems we have discussed in Sect. 3.3. The cases in which similarity plays no role are left unresolved and it remains unclear how surrogative reasoning with such models is supposed to happen.

The next issue is that not all models have a target system, which is a serious problem for a view that analyzes representation in terms of imagining something *about* a target. *Toon* is well aware of this issue and calls them *models without objects* [3.205, p. 76]. Some of these are models of discredited entities like the ether and phlogiston, which were initially thought to have a target but then turned out not to have one [3.205, p. 76]. But not all models without objects are errors: architectural plans of buildings that are never built or models of experiments that are never carried out fall into the same category [3.205, p. 76].

*Toon* addresses this problem by drawing another analogy with fiction. He points out that not all novels are like *The War of the Worlds*, which has an object. Passages from *Dracula*, for instance, “do not represent any actual, concrete object but are instead about fictional characters” [3.205, p. 54]. Models without a target are like passages from *Dracula*. So the solution to the problem is to separate the two cases neatly. When a model has target then it represents that target

by prescribing imaginings about the target; if a model has no target it prescribes imaginings about a fictional character [3.205, p. 54].

*Toon* immediately admits that models without targets “give rise to all the usual problems with fictional characters” [3.205, p. 54]. However, he seems to think that this is a problem we can live with because the more important case is the one where models do have a target, and his account offers a neat solution there. He offers the following summative statement of his account [3.205, p. 62]:

#### **Definition 3.13 Direct Representation**

A scientific model  $M$  represents a target system  $T$  iff  $M$  functions as prop in game of make believe.

This definition takes it to be understood that the imaginings prescribed are about the target  $T$  if there is a target, and about a fictional character if there isn’t because there need not be any object that the model prescribes imaginings about [3.205, p. 81].

This bifurcation of imaginative activities raises questions. The first is whether the bifurcation squares with the face value practice. *Toon*’s presentation would suggest that the imaginative practices involved in models with targets are very different from the ones involved in models without them. Moreover, they require a different analysis because imagining something about an existing object is different from imagining something about a fictional entity. This, however, does not seem to sit well with scientific practice. In some cases we are mistaken: we think that the target exists but then find out that it doesn’t (as in the case of phlogiston). But does that make a difference to the imaginative engagement with a phlogiston model of combustion? Even today we can understand and use such models in much the same way as its original protagonists did, and knowing that there is no target seems to make little, if any, difference to our imaginative engagement with the model. Of course the presence or absence of a target matters to many other issues, most notably surrogative reasoning (there is nothing to reason about if there is no target!), but it seems to have little importance for how we imaginatively engage with the scenario presented to us in a model.

In other cases it is simply left open whether there is target when the model is developed. In elementary particle physics, for instance, a scenario is often proposed simply as a suggestion worth considering and only later, when all the details are worked out, the question is asked whether this scenario bears an interesting relation to what happens in nature, and if so what the relation is. So, again, the question of whether there is or isn’t a target seems to have little, if any, influence

on the imaginative engagement of physicists with scenarios in the research process. This does not preclude different philosophical analyses being given of modeling with and without a target, but any such analysis will have to make clear the commonalities between the two.

Let us now turn to a few other aspects of direct representation (Definition 3.13). The view successfully solves the problem of asymmetry. Even if it uses similarity in response to the ER-problem, the imaginative process is clearly directed towards the target. An appeal to imagination also solves the problem of misrepresentation because there is no expectation that our imaginations are correct when interpreted as statements about the target. Given its roots in a theory of representation in art, it's natural to renounce any attempts to demarcate scientific representation from other kinds of representation [3.205, p. 62]. The problem of ontology is dispelled for representations with an object, but it remains unresolved for representations without one. However, direct representation (Definition 3.13) offers at best a partial answer to the ER-problem, and nothing is said about either the problem of style and/or standards of accuracy. Similarly, Toon remains silent about the applicability of mathematics.

Levy also rejects an indirect view primarily because of the unwieldiness of its ontology and endorses a direct view of representation ([3.43, pp. 780–790], [3.206, pp. 744–747]). Like Toon, he develops his version of the direct view by appeal to Walton's notion of prop-oriented make believe. When, for instance, we're asked where in Italy the town of Crotona lies, we can be told that it's in the arch of the Italian boot. In doing so we are asked to imagine something about the shape of Italy and this imagination is used to convey geographical information. Levy then submits that "we treat models as games of prop-oriented make believe" [3.206, p. 791]. Hence modeling consists in imagining something directly about the target.

Levy pays careful attention to the ER-problem. In his [3.206, p. 744] he proposed that the problem be conceptualized in analogy with metaphors, but immediately added that this was only a beginning which requires substantial elaboration. In his [3.43, pp. 792–796] he takes a different route and appeals to Yablo's [3.230] theory of partial truth. The core idea of this view is that a statement is partially true "if it is true when evaluated only relative to a subset of the circumstances that make up its subject matter – the subset corresponding to the relevant content-part" [3.43, p. 792]. Levy submits that this will also work for a number of cases of modeling, but immediately adds that there are other sorts of cases that don't fit the mold [3.43, p. 794]. Such cases often are ones in which

distortive idealizations are crucial and cannot be set aside. These require a different treatment and it's an open question what this treatment would be.

Levy offers a radical solution to the problem of models without targets: there aren't any! He first broadens the notion of a target system, allowing for models that are only loosely connected to targets [3.43, pp. 796–797]. To this end he appeals to Godfrey-Smith's notion of *hub-and-spoke* cases: families of models where only some have a target (which makes them the hub models) and the others are connected to them via conceptual links (spokes) but don't have a specific target. Levy points out that such cases should be understood as having a *generalized target*. If something that looks like a model doesn't meet the requirement of having even a generalized target, then it's not a model at all. Levy mentions structures like the game of life and observes that they are "bits of mathematics" rather than models [3.43, p. 797]. This eliminates the need for fictional characters in the case of targetless models.

This is a heroic act of liberation, but questions about it remain. The direct view renders fictional entities otiose by positing that a model is nothing but an act of imagining something about a concrete actual thing. But generalized targets are not concrete actual things, and often not even classes of such things. There is a serious question whether one can still reap the (alleged) benefits of a view that analyzes modeling as imaginings about concrete things, if the things about which we imagine something are no longer concrete. Population growth or complex behavior are not concrete things like rabbits and stumps, and this would seem to pull the rug from underneath a direct approach to representation. Likewise, the claim that models without target are just mathematics stands in need of further elucidation. Looking back at Toon's examples of such models, a view that considers them just mathematics does not come out looking very natural.

### 3.6.3 Parables and Fables

Cartwright [3.231] focuses on highly idealized models such as Schelling's model of social segregation [3.232] and Pissarides' model of the labor market [3.233]. The problem with these models is that the objects and situations we find in such models are not at all like the things in the world that we are interested in. Cities aren't organized as checkerboards and people don't move according to simple algorithmic rules (as they do in Schelling's model), and there are no laborers who are solely interested in leisure and income (as is the case in Pissarides' model). Yet we are supposed to learn something about the real world from these models. The question is how.

*Cartwright* submits that an answer to this question emerges from a comparison of models with narratives, in particular fables and parables. An example of a fable is the following: “A marten eats the grouse; a fox throttles the marten; the tooth of the wolf, the fox. Moral: the weaker are always prey to the stronger” [3.231, p. 20]. The characters in the fable are highly idiosyncratic, and typically we aren’t interested in them per se – we don’t read fables to learn about foxes and martens. What we are interested in is the fable’s general and more abstract conclusion, in the above example that the weaker are always prey to the stronger. In the case of the fable the moral is typically built in the story and explicitly stated [3.231].

*Cartwright* then invites us to consider the parable of the laborers in the vineyard told in the Gospel of Matthew [3.231]. A man goes to the market to hire day laborers. He hires the first group early in the morning, and then returns several times during the day to hire more laborers, and he hires the last group shortly before dusk. Some worked all day, while some hardly started when the day ended. Yet he pays the same amount to all of them. Like in a fable, when engaging with a parable the reader takes no intrinsic interest in the actors and instead tries to extract a more general moral. But unlike in fables, in parables no moral appears as part of the parable itself [3.231, p. 29]. Hence parables need interpretation, and alternative interpretations are possible. The above fable is often interpreted as being about the entry to God’s kingdom, but, as *Cartwright* observes, it can just as well be interpreted as making the market-based capitalist point that you get what you contract for, and should not appeal to higher forms of justice [3.231, p. 21].

These are features models share with fables and parables: “like the characters in the fable, the objects in the model are highly special and do not in general resemble the ones we want to learn about” [3.231, p. 20] and the “lesson of the model is, properly, more abstract than what is seen to happen in the model” [3.231, p. 28]. This leaves the question whether models are fables or parables. Some models are like fables in that they have the conclusion explicitly stated in them. But most models are like parables [3.231, p. 29]: their lesson is not written in the models themselves [3.231, p. 21], and worse: “a variety of morals can be attributed to the models” [3.231, p. 21]. A model, just like a parable, is interpreted against a rich background of theory and observation, and the conclusion we draw depends to a large extent on the background [3.231, p. 30].

So far the focus was on deriving a conclusion about the model *itself*. *Cartwright* is clear that one more step is needed: “In many cases we want to use the results of these models to inform our conclusions about

a range of actually occurring (so-called *target*) situations” [3.231, p. 22] (original emphasis). In fact, making this transfer of model results to the real world is the ER-problem. Unfortunately she does not offer much by way of explaining this step and merely observes that “a description of what happens in the model that does not fit the target gets recast as one that can” [3.231, p. 20]. This gestures in the right direction, but more would have to be said about how exactly a model description is recast to allow for transfer of model results to target systems. In earlier work *Cartwright* observed that what underlies the relationship between models and their targets is a “loose notion of resemblance” [3.73, pp. 192–193] and [3.74, pp. 261–262]. This could be read as suggesting that she would endorse some kind of similarity view of representation. Such a view, however, is independent of an appeal to fables and parables.

In passing we would like to mention that the same kind of models is also discussed in *Sugden* [3.202, 210]. However, his interest is in induction rather than representation, and if reframed in representational terms then his account becomes a similarity account like *Giere*’s. See *Grüne-Yanoff* [3.234] and *Knuuttila* [3.235] for a discussion.

### 3.6.4 Against Fiction

The criticisms we have encountered above were intrinsic criticisms of particular versions of the fiction view, and as such they presuppose a constructive engagement with the view’s point of departure. Some critics think that any such engagement is misplaced because the view got started on the wrong foot entirely. There are five different lines of attack. The first criticism is driven by philosophical worries about fiction. Fictions, so the argument goes, are intrinsically dubious and are beset with so many serious problems that one should steer away from them whenever possible. So it could be claimed that assigning them a central role in science is a manifestation of philosophical masochism. This, however, overstates the problems with fictions. Sure enough, there is controversy about fictions. But the problems pertaining to fictions aren’t more devastating than those surrounding other items on the philosophical curriculum, and these problems surely don’t render fictions off limits.

The second criticism, offered for example by *Giere* [3.97, p. 257], is that the fiction view – involuntarily – plays into the hands of irrationalists. Creationists and other science skeptics will find great comfort, if not powerful rhetorical ammunition, in the fact that philosophers of science say that scientists produce fiction. This, so the argument goes, will be seen

as a justification of the view that religious dogma is on par with, or even superior to, scientific knowledge. Hence the fiction view of models undermines the authority of science and fosters the cause of those who wish to replace science with religious or other unscientific worldviews.

Needless to say, we share Giere's concerns about creationism. In order not to misidentify the problem it is important to point out that Giere's claim is not that the view itself – or its proponents – support creationism; his worry is that the view is a dangerous tool when it falls into the wrong hands. What follows from this, however, is not that the fiction view itself should be abandoned; but rather that some care is needed when dealing with the press office. As long as the fiction view of models is discussed in informed circles, and, when popularized, is presented carefully and with the necessary qualifications, it is no more dangerous than other ideas, which, when taken out of context, can be put to uses that would (probably) send shivers down the spines of their progenitors (think, for instance, of the use of Darwinism to justify eugenics).

The third objection, also due to *Giere*, has it that the fiction view misidentifies the aims of models. Giere agrees that from an *ontological* point of view scientific models and works of fictions are on par, but emphasizes that “[i]t is their differing function in practice that makes it inappropriate to regard scientific models as works of fiction” [3.97, p. 249]. *Giere* identifies three functional differences [3.97, pp. 249–252]. First, while fictions are the product of a single author's individual endeavors, scientific models are the result of a public effort because scientists discuss their creations with their colleagues and subject them to public scrutiny. Second, there is a clear distinction between fiction and nonfiction books, and even when a book classified as nonfiction is found to contain false claims, it is not reclassified as fiction. Third, unlike works of fiction, whose prime purpose is to entertain (although some works can also give insight into certain aspects of human life), scientific models are representations of certain aspects of the world.

These observations, although correct in themselves, have no force against the fiction view of models. First, whether a fiction is the product of an individual or a collective effort has no impact on its status as a fiction; a collectively produced fiction is just a different kind of fiction. Even if *War and Peace* (to take Giere's example) had been written in a collective effort by all established Russian writers of Tolstoy's time, it would still be a fiction. Vice versa, even if Newton had never discussed his model of the solar system with anybody before publishing it, it would still be science. The history of production is immaterial to the fictional status

of a work. Second, as we have seen in Sect. 3.6.1, falsity is not a defining feature of fiction. We agree with Giere that there is a clear distinction between texts of fiction and nonfiction, but we deny that this distinction is defined by truth or falsity; it is the attitude that we are supposed to adopt towards the text's content that makes the difference. Once this is realized, the problem fades away. Third, many proponents of the fiction view (those belonging to the first group mentioned in Sect. 3.6.1) are clear that problems of ontology should be kept separate from function and agree that it is one of the prime function of models to represent. This point has been stressed by *Godfrey-Smith* [3.209, pp. 108–111] and it is explicit in other views such as *Frigg's* [3.203].

The fourth objection is due to *Magnani*, who dismisses the fiction view for misconstruing the role of models in the process of scientific discovery. The fundamental role played by models, he emphasizes [3.236, p. 3]:

“is the one we find in the core conceptual discovery processes, and that these kinds of models cannot be indicated as fictional at all, because they are constitutive of new scientific frameworks and new empirical domains.”

This criticism seems to be based on an understanding of fiction as falsity because falsities can't play a constitutive role in the constitution of new empirical domains. We reiterate that the fiction view is not committed to the *fiction as falsity* account and hence is not open to this objection.

The fifth objection is that fictions are superfluous and hence should not be regarded as forming part of (let alone *being*) scientific models because we can give a systematic account of how scientific models work without invoking fictions. This point has been made in different ways by *Pincock* [3.214, Chap. 12] and *Weisberg* [3.33, Chap. 4] (for a discussion of *Weisberg's* arguments see *Odenbaugh* [3.237]). We cannot do justice to the details of their sophisticated arguments here, and will concern ourselves only with their main conclusion. They argue that scientific models are mathematical objects and that they relate to the world due to the fact that there is a relationship between the mathematical properties of the model and the properties found in the target system (in *Weisberg's* version similarity relations to a parametrized version of the target). In other words, models are mathematical structures and they represent due to there being certain mathematical relations between these structures and a mathematical rendering of the target system. (*Weisberg* includes fictions as convenient *folk ontology* that may serve as a crutch when thinking about the model, but takes them to be ultimately dispensable when it comes to explain-

ing how models relate to the world.) This, however, brings us back to a structuralist theory of representation, and this theory, as we have seen in Sect. 3.4, is

far from unproblematic. So it is at best an open question whether getting rid of fiction provides an obvious advantage.

## 3.7 Representation-as

In this section we discuss approaches that depart from *Goodman's* notion of *representation-as* [3.64]. In his account of aesthetic representation the idea is that a work of art does not just denote its subject, but moreover it represents it as being thus or so. *Elgin* [3.34] further developed this account and, crucially, suggested that it also applies to scientific representations. This is a vital insight and it provides the entry point to what we think of as the most promising account of epistemic representation.

In this section we present Goodman and Elgin's notion of *representation-as*, and outline how it is a complex type of reference involving a mixture of denotation and what they call exemplification. We introduce the term of art *representation-as* to indicate that we are talking about the specific concept that emerges from Goodman's and Elgin's writings. We then discuss how the account needs to be developed in the context of scientific representation. And finally we present our own answer to the ER-problem, and demonstrate how it answers the questions laid out in Sect. 3.1.

### 3.7.1 Exemplification and Representation-as

Many instances of epistemic representation are instances of representation-as. Caricatures are paradigmatic examples: Churchill is represented as a bulldog, Thatcher is represented as a boxer, and the Olympic Stadium is represented as a UFO. Using these caricatures we can attempt to learn about their targets: attempt to learn about a politician's personality or a building's appearance. The notion applies beyond caricatures. Holbein's *Portrait of Henry VIII* represents Henry as imposing and powerful and Stoddart's statue of David Hume represents him as thoughtful and wise. The leading idea is that scientific representation works in much the same way. A model of the solar system represents the sun as perfect sphere; the logistic model of growth represents the population as reproducing at fixed intervals of time; and so on. In each instance, models can be used to attempt to learn about their targets by determining what the former represent the latter as being. So representation-as relates, in a way to be made more specific below, to the surrogate reasoning condition discussed in Sect. 3.1.

The locution of representation-as functions in the following way: An object  $X$  (e.g., a picture, statue, or model) represents a subject  $Y$  (e.g., a person or target system) as being thus or so ( $Z$ ). The question then is what establishes this sort of representational relationship? The answer requires presenting some of the tools Goodman and Elgin use to develop their account of representation-as.

One of the central posits of *Goodman's* account is that denotation is "the core of representation" [3.64, p. 5]. Stoddart's statue of David Hume denotes Hume and a model of the solar system denotes the solar system. In that sense the statue and the model are representations of their respective targets. To distinguish representation of something from other notions of representation we introduce the technical term *representation-of*. Denotation is what establishes representation-of. (For a number of qualifications and caveats about denotation see our [3.238, Sect. 2]).

Not all representations are a representation-of. A picture showing a unicorn is not a representation-of a unicorn because things that don't exist can't be denoted. Yet there is a clear sense in which such a picture is a representation. *Goodman* and *Elgin's* solution to this is to distinguish between being a representation-of something and being a something-representation ([3.34, pp. 1–2], [3.64, pp. 21–26]). What makes a picture a something-representation (despite the fact it may fail to denote anything) is that it is the sort of symbol that denotes. *Elgin* argues [3.34, pp. 1–2]:

"A picture that portrays a griffin, a map that maps the route to Mordor [...] are all representations, although they do not represent anything. To be a representation, a symbol need not itself denote, but it needs to be the sort of symbol that denotes. Griffin pictures are representations then because they are animal pictures, and some animal pictures denote animals. Middle Earth maps are representations because they are maps and some maps denote real locations. [...] So whether a symbol is a representation is a question of what kind of symbol it is."

These representations can be classified into genres, in a way that does not depend on what they are

representations-of (since some may fail to denote), but instead on what they portray. In the case of pictures, this is fairly intuitive (how this is to be developed in the case of scientific models is discussed below). If a picture portrays a man, it is a man-representation, if it portrays a griffin it is a griffin-representation and so on. In general, a picture  $X$  is  $Z$ -representation if it portrays  $Z$ . The crucial point is that this does not presuppose that  $X$  be a representation-of  $Z$ ; indeed  $X$  can be  $Z$ -representation without denoting anything. A picture must denote a man to be a representation-of a man. But it need not denote anything to be a man-representation.

The next notion we need to introduce is *exemplification*. An item exemplifies a property if it at once instantiates the property and refers to it [3.64, p. 53]:

“Exemplification is possession plus reference. To have without symbolizing is merely to possess, while to symbolize without having is to refer in some other way than by exemplifying.”

Exemplification is a mode of reference that holds between items and properties. In the current context properties are to be understood in the widest possible sense. An item can exemplify one-place properties, multi-place properties (i.e., relations), higher order properties, structural properties, etc. Paradigmatic examples of exemplification are samples. A chip of paint on a manufacturer’s sample card both instantiates a certain color, and at the same time refers to that color [3.239, p. 71].

But although exemplification requires instantiation, not every property instantiated by an object is exemplified by it. The chip of paint does not, for example, exemplify its shape or its location on the card. In order to exemplify a property, an object must both instantiate the property and the property itself must be made epistemically salient. How saliency is established will be determined on a case-by-case basis, and we say more about this below.

We can now turn to the conditions under which  $X$  represents  $Y$  as  $Z$ . A first stab would be to say that  $X$  represents  $Y$  as  $Z$  if  $X$  is a  $Z$ -representation and denotes  $Y$ . This however, is not yet good enough. It is important that properties of  $Z$  are *transferred* to  $Y$ . *Elgin* makes this explicit [3.34, p. 10]:

“[ $X$ ] does not merely denote [ $Y$ ] and happen to be a [ $Z$ ]-representation. Rather in being a [ $Z$ ]-representation, [ $X$ ] exemplifies certain properties and imputes those properties or related ones to [ $Y$ ]. [...] The properties exemplified in the [ $Z$ ]-representation thus serve as a bridge that connects [ $X$ ] to [ $Y$ ].”

This gives a name to the crucial step: imputation. This step can be analyzed in terms of stipulation by a user of a representation. When someone uses  $X$  as a representation-as, she has to stipulate that certain properties that are exemplified in  $X$  be imputed to  $Y$ . We emphasize that imputation does not imply truth:  $Y$  may or may not have the properties imputed to it by  $X$ . So the representation can be seen as generating a claim about  $Y$  that can be true or false; it should not be understood as producing truisms.

Applied to scientific models, the account of epistemic representation that emerges from Goodman and Elgin’s discussion of representation can then be summarized as follows:

#### Definition 3.14 Representation-As

A scientific model  $M$  represents a target system  $T$  iff:

1.  $M$  denotes  $T$
2.  $M$  is a  $Z$ -representation exemplifying properties  $P_1, \dots, P_n$
3.  $P_1, \dots, P_n$ , or related properties, are imputed to  $T$ .

It should be added that the first condition can easily be extended to include part-part denotation. In a family portrait the entire portrait denotes the family; at the same time a part of the portrait can denote the mother and another part the father. This is obvious and unproblematic.

We think that this account is on the right track, but all three conditions need to be further developed to furnish a full-fledged account of epistemic representation (at least as applied to scientific models). The developments needed are of different kinds, though. The first condition needs more specificity. How is denotation characterized? What different ways of establishing denotation are there? And how is denotation established in particular cases? These are but some of the questions that a complete account of epistemic representation will have to answer. In many cases epistemic representation seems to borrow denotation from linguistic descriptions in which they are embellished and denotation is in effect borrowed from language. So the philosopher of science can turn to the philosophy of language to get a deeper understanding of denotation. This is an interesting project, but it is not one we can pursue here.

In contrast with denotation the other two conditions need to be reformulated because an account molded on visual representations is only an imperfect match for scientific representations. This is the task for the next section.

### 3.7.2 From Pictures to Models: The Denotation, Exemplification, Keying-up and Imputation Account

According to Goodman and Elgin, for a picture to be a *Z*-representation it has to be the kind of symbol that denotes. On the face of it, there is a mismatch between pictures and scientific models in this regard. The Schelling model represents social segregation with a checkerboard; billiard balls are used to represent molecules; the Phillips–Newlyn model uses a system of pipes and reservoirs to represent the flow of money through an economy; and the worm *Caenorhabditis elegans* is used as a model of other organisms. But neither checkerboards, billiard balls, pipes, or worms seem to belong to classes of objects that typically denote. The same observation applies to scientific fictions (frictionless planes, utility maximizing agents, and so on) and the mathematical objects used in science. In fact, matrices, curvilinear geometries, Hilbert spaces etc. were all studied as mathematical objects before they became important in the empirical sciences.

Rather than relying on the idea that scientific models belong to classes of objects that typically denote we propose directly introducing an agent and ground representation in this agent's actions. Specific checkerboards, systems of pipes, frictionless places and mathematical structures, are epistemic representations because they are used by an agent to represent a system. When an agent uses an object as a representation, we call it a *base*.

What allows us to classify bases into *Z*-representations is also less clear in the case of scientific representation. We approach this issue in two steps. The first is to recognize the importance of the intrinsic constitution of the base. Pictures are typically canvases covered with paint. They are classified as *Z*-representations because under appropriate circumstances the canvas is recognized as portraying a *Z*. Much can be said about the canvas' material constitution (the thickness or chemical constitution of the paint, etc.), but these are generally of little interest to understanding what the picture portrays. By contrast, the properties of a scientific model – qua material object – do matter. How water flows through the pipes in the Phillips–Newlyn model is crucial to how it represents the movement of money in an economy. That *Caenorhabditis elegans* is a biological organism is of vital importance for how it is used representationally. In fact, models are frequently classified according to what their material base is. We talk about a pipe model of the economy or worm model of cell division because their bases are pipes and worms. Here we introduce a term of art to recognize that scientific models are generally categorized according to their

material constitution. An *O*-object specifies the kind of object something is, qua physical object.

*O*-objects become representations when they are used as such. But how are they classified as *Z*-representations? How does the Phillips–Newlyn machine become an economy-representation, or how does a collection of billiard balls become a gas-representation? (Again, recall that this is not because they denote economies or gases.) We suggest, and this is the second step, that this requires an act of *interpretation* (notice that we do not use *interpretation* in the same sense as Contessa). In the case of pictures, the nature of this interpretation has been the center of attention for a good while: how one sees a canvas covered with paint as showing a cathedral is regarded by many as one of the important problems of aesthetics. Schier [3.240, p. 1] dubbed it the “enigma of depiction”, and an entire body of literature is concerned with it (*Kulvicki* [3.241] provides a useful review). In the case of scientific models we don't think a simple and universal account of how models are interpreted as *Z*-representations can be given. Interpreting an *O*-object as a *Z*-representation requires attributing properties of *Zs* to the object. How this is done will depend on disciplinary traditions, research interests, background theory and much more. In fact, *interpretation* is a blank to be filled, and it will be filled differently in different cases.

Some examples should help elucidate what we mean by this. In the case of scale models the interpretation is *close* to the *O*-object in that it interprets the object in its *own* terms. The small car is interpreted as a car-representation and the small ship is interpreted as a ship-representation. Likewise, in the case of the Army Corps' model of the San Francisco bay [3.33], parts of the model bay are interpreted in terms of the real bay. In cases like these, the same predicates that apply to the base (qua *O*-object) are applied to the object in order to make it into a *Z*-representation (here  $O = Z$ ). But this is not always the case. For example, the Phillips–Newlyn machine is a system of pipes and reservoirs, but it becomes an economy-representation only when the quantity and flow of water throughout the system are interpreted as the quantity and flow of money throughout an economy. The system is interpreted in terms of predicates that do not apply to the object (qua *O*-object), but turn it into a *Z*-representation (here *O* and *Z* come apart). In sum, an *O*-object that has been chosen as the base of a representation becomes a *Z*-representation if *O* is interpreted in terms of *Z*.

Next in line is exemplification. Much can be said about exemplification in general, but the points by and large carry over from the general discussion to the case of models without much ado. There is one difference,

though, in cases like the Phillips–Newlyn machine. Recall that exemplification was defined as the instantiation of a property  $P$  by an object in such a way that the object thereby refers to  $P$ . How can the Phillips–Newlyn machine exemplify economic properties when it does not, strictly speaking, instantiate them? The crucial point is that nothing in the current account depends on instantiation being literal instantiation. On this point we are in agreement with Goodman and Elgin, whose account relies on nonliteral instantiation. The portrait of Henry cannot, strictly speaking, instantiate the property of being male, even if it represents him as such. Goodman and Elgin call this metaphorical instantiation ([3.64, pp. 50–51], [3.239, p. 81]).

What matters is that properties are epistemically accessible and salient, and this can be achieved with what we call *instantiation-under-an-interpretation*  $I$ , *I-instantiation* for short. An economic interpretation of the Phillips–Newlyn machine interprets amounts of water as amounts of money. It does so by introducing a clearly circumscribed rule of proportionality:  $x$  liters of water correspond to  $y$  millions of the model-economy’s currency. This rule is applied without exception when the machine is interpreted as an economy-representation. So we say that under the economic interpretation  $I_e$  the machine  $I_e$ -instantiates money properties. With the notion of *I-instantiation* at hand, exemplification poses no problem.

The final issue to clear is the imputation of the model’s exemplified properties to the target system. In particular, which properties are so imputed? Elgin describes this as the imputation of the properties exemplified by  $M$  or *related ones*. The observation that the properties exemplified by a scientific model and the properties imputed to its target system need not be identical is correct. In fact, few, if any, models in science portray their targets as exhibiting exactly the same features as the model itself. The problem with invoking *related* properties is not its correctness, but its lack of specificity. Any property can be related to any other property in some way or other, and as long as no specific relation is specified it remains unclear which properties are imputed onto the system.

In the context of science, the relation between the properties exemplified and the ones ascribed to the system is sometimes described as one of simplification [3.198, p. 184], idealization [3.198, p. 184] and approximation [3.34, p. 11]. This could suggest that *related ones* means *idealized*, at least in the context of science (we are not attributing this claim to Elgin; we are merely considering the option), perhaps similar to the way in which Ducheyne’s account discussed above took target properties to be approximations of model properties. But shifting from *related* to *idealized* or

*approximated* (or any of their cognates) makes things worse, not better. For one, *idealization* can mean very different things in different contexts and hence describing the relation between two properties as *idealization* adds little specificity (see Jones [3.242] for a discussion of different kinds of idealization). For another, while the relationship between some representation-target properties may be characterized in terms of idealization, many cannot. A map of the world exemplifies a distance of 29 cm between the two points labeled *Paris* and *New York*; the distance between the two cities is 5800 km; but 29 cm is not an idealization of 5800 km. A scale model of a ship being towed through water is not an idealization of an actual ship, at least not in any obvious way. Or in standard representations of Mandelbrot sets the color of a point indicates the speed of divergence of an iterative function for certain parameter value associated with that point, but color is not an idealization of divergence speed.

For this reason it is preferable, in our view, to build a specification of the relationship between model properties and target properties directly into an account of epistemic representation. Let  $P_1, \dots, P_n$  be the properties exemplified by  $M$ , and let  $Q_1, \dots, Q_m$  be the *related* properties that  $M$  imputes to  $T$  (where  $n$  and  $m$  are positive natural numbers that can but need not be equal). Then the representation  $M$  must come with a key  $K$  that specifies how exactly  $P_1, \dots, P_n$  are converted into  $Q_1, \dots, Q_m$  [3.50]. Borrowing notation from algebra (somewhat loosely) we can write  $K((P_1, \dots, P_n)) = \langle Q_1, \dots, Q_m \rangle$ .  $K$  can, but need not be, the identity function; any rule that associates a unique set  $Q_1, \dots, Q_m$  with  $P_1, \dots, P_n$  is admissible. The relevant clause in the definition of representation-as then becomes:  $M$  exemplifies  $P_1, \dots, P_n$  and the representation imputes properties  $Q_1, \dots, Q_m$  to  $T$  where the two sets of properties are connected to each other by a key  $K$ .

The above examples help illustrate what we have in mind. Let us begin with the example of the map (in fact the idea of a key is motivated by a study of maps; for a discussion of maps see Galton [3.243] and Sismondo and Chrisman [3.244]).  $P$  is a measured distance on the map between the point labeled *New York* and the point labeled *Paris*;  $Q$  is the distance between New York and Paris in the world; and  $K$  is the scale of the map (in the above case,  $1 : 20000000$ ). So the key allows us to translate a property of the map (the 29 cm distance) into a property of the world (that New York and Paris are 5800 km apart). But the key involved in the scale model of the ship is more complicated. One of the  $P$ s in this instance is the resistance the model ship faces when moved through the water in a tank. But this doesn’t translate into the resistance faced by the actual ship in the same way in which distances in a map trans-



late into distances in reality. In fact, the relation between the resistance of the model and the resistance of the real ship stand in a complicated nonlinear relationship because smaller models encounter disproportionate effects due to the viscosity of the fluid. The exact form of the key is often highly nontrivial and emerges as the result of a thoroughgoing study of the situation; see Sterrett [3.245] for a discussion of fluid mechanics. In the representation of the Madelbrod set in [3.246, p. 660] a key is used that translates color into divergence speed [3.246, p. 695]. The square shown is a segment of the complex plane and each point represents a complex number. This number is used as parameter value for an iterative function. If the function converges for number  $c$ , then the point in the plane representing  $c$  is colored black. If the function diverges, then a shading from yellow over green to blue is used to indicate the speed of divergence, where yellow is slow, green is in the middle and blue is fast.

Neither of these keys is obvious or trivial. Determining how to move from properties exemplified by models to properties of their target systems can be a significant task, and should not go unrecognized in an account of scientific representation. In general  $K$  is a blank to be filled, and it depends on a number of factors: the scientific discipline, the context, the aims and purposes for which  $M$  is used, the theoretical backdrop against which  $M$  operates, etc. Building  $K$  into the definition of representation-as does not prejudice the nature of  $K$ , much less single out a particular key as the correct one. The requirement merely is that there must be *some* key for  $M$  to qualify as a representation-as.

With these modifications in place we can now formulate our own account of representation [3.238, 247]. Consider an agent who chooses an  $O$ -object as the base of representation and turns it into  $Z$ -representation by adopting an interpretation  $I$ . Let  $M$  refer to the package of the  $O$ -object together with the interpretation  $I$  that turns it into a  $Z$ -representation. Then:

#### Definition 3.15 DEKI

A scientific model  $M$  represents a target  $T$  iff:

1.  $M$  denotes  $T$  (and, possibly, parts of  $M$  denote parts of  $T$ )
2.  $M$  is a  $Z$ -representation exemplifying properties  $P_1, \dots, P_n$
3.  $M$  comes with a key,  $K$ , specifying how  $P_1, \dots, P_n$  are translated into a set of features  $Q_1, \dots, Q_m$ :  $K((P_1, \dots, P_n)) = (Q_1, \dots, Q_m)$
4. The model imputes at least one of the properties  $Q_1, \dots, Q_m$  onto  $T$ .

We call this the DEKI account of representation to highlight its key features: denotation, exemplification, keying-up and imputation.

Before highlighting some issues with this account, let us clarify how the account answers the questions we laid out in Sect. 3.1. Firstly, as an answer to the ER-problem, DEKI (Definition 3.15) provides an abstract framework in which to think about epistemic representation. In general, what concretizes each of the conditions needs to be investigated on a case-by-case basis. But far from being a defect, this degree of abstractness is an advantage. *Epistemic representation*, and even the narrower *model-representation*, are umbrella terms covering a vast array of different activities in different fields, and a view that sees representations in fields as diverse as elementary particle physics, evolutionary biology, hydrology and rational choice theory work in exactly the same way is either mistaken or too coarse to make important features visible. DEKI (Definition 3.15) occupies the right middle ground: it is general enough to cover a large array of cases and yet it highlights what all instances of scientific representation have in common. At the same time the account offers an elegant solution to the problem of models without targets: a model that apparently represents  $Z$  while there is no  $Z$  is a  $Z$ -representation but not representation of a  $Z$ .

It should be clear how we can use models to perform surrogative reasoning about their targets according to DEKI (Definition 3.15). The account requires that we investigate the properties that are exhibited by the model. These are then translated into a set of properties that are imputed onto the target. This act of imputation supplies a hypothesis about the target system: does it, or does it not, have those properties? This hypothesis does not have to be true, and as such DEKI (Definition 3.15) allows for the possibility of misrepresentation in a straightforward manner.

DEKI's (Definition 3.15) abstract character also allows us to talk about different styles of representation. Style, on the DEKI (Definition 3.15) account, is not a monolithic concept; instead it has several dimensions. Firstly, different  $O$ -objects can be chosen. In this way we may speak, say, of the *checkerboard style* and of the *cellular automaton style*. In each case a specific kind of object has been chosen for various modeling purposes. Secondly, the notion of an interpretation allows us to talk about how closely connected the properties of the model are to those that the object  $I$ -instantiates. Thirdly, different types of keys could be used to characterize different styles. In some instances the key might be the identity key, which would amount to a style of modeling that aims to construct replicas of target systems; in other cases the key incorporates different kinds of ideal-

izations or abstractions, which gives rise to idealization and abstraction keys. But different keys may be associated with entirely different representational styles.

Similarly, DEKI (Definition 3.15) suggests that there is no significant difference between scientific representations and other kinds of epistemic representation, at least at the general level. However, this is not to say that the two cannot be demarcated whatsoever. The sorts of interpretations under which pictures portray *Zs* seem to be different to the sorts of interpretations that are adopted in the scientific framework. Whether or not this can be cashed of more specifically is an interesting question that we cannot investigate here.

Many details in DEKI (Definition 3.15) still need to be spelled out. But the most significant difficulty, perhaps, arises in connection with the problem of ontology. It is not by accident that we have illustrated the account with a physical model, the Phillips–Newlyn machine. Exemplification requires instantiation, which is easily understood for material models, but is highly problematic in the context of nonconcrete models. One option is to view models as fictional entities as discussed in Sect. 3.6. But whether, and if so how, fictional entities instantiate properties is controversially discussed and more philosophical work is needed to make sense of such a notion. It is therefore an open question how this account works for nonconcrete models; for a discussion and a proposal see *Frigg and Nguyen* [3.248].

Finally, the account provides us with resources with which to think about the applicability of mathematics.

### 3.8 Envoi

We reviewed theories of epistemic representation. That each approach faces a number of challenges and that there is no consensus on the matter will not have come as a surprise to anybody. We hope, however, that we managed to map the lay of the land and to uncover the fault lines, and thereby aid future discussions.

Like the problem of style, various options are available. Firstly, mathematical structures themselves can be taken to be *O*-objects and feature as bases of representation. They can be interpreted on their own terms and therefore exemplify strictly mathematical properties. If one were of a structuralist bent, then the appropriate mathematical properties could be *structural*, which could then be imputed onto the target system (although notice that this approach faces a similar problem to the question of target-end structure discussed in Sect. 3.4.4). Alternatively, the key could provide a translation of these mathematical properties into ones more readily applicable to physical systems. A third alternative would be to take scientific models to be fictional objects, and then adopt an interpretation towards them under which they exemplify mathematical properties. Again, these could be imputed directly onto the target system, or translated into an alternative set of properties. Finally, these fictional models could themselves exemplify physical properties, but in doing so exemplify structural ones as well. Whenever a physical property is exemplified, this provides an extensional relation defined over the objects that instantiate it. The pros and cons of each of these approaches demands further research, but for the purposes of this chapter we simply note that DEKI (Definition 3.15) puts all of these options on the table. Using the framework of *O*-objects, interpretations, exemplification, keys, and imputation provides a novel way in which to think about the applicability of mathematics.

**Acknowledgments.** The authors are listed alphabetically; the chapter is fully collaborative. We would like to thank Demetris Portides and Fiora Salis for helpful comments on an earlier draft.

### References

- |     |   |     |   |
|-----|---|-----|---|
| 3.1 | G. Boniolo: <i>On Scientific Representations: From Kant to a New Philosophy of Science</i> (Palgrave Macmillan, Hampshire, New York 2007)   | 3.5 | J. Elkins: <i>The Domain of Images</i> (Cornell Univ. Press, Ithaca, London 1999)   |
| 3.2 | L. Perini: The truth in pictures, <i>Philos. Sci.</i> <b>72</b> , 262–285 (2005)  | 3.6 | K. Warmbröd: Primitive representation and misrepresentation, <i>Topoi</i> <b>11</b> , 89–101 (1992)   |
| 3.3 | L. Perini: Visual representation and confirmation, <i>Philos. Sci.</i> <b>72</b> , 913–926 (2005)   | 3.7 | C. Peirce: Principles of philosophy and elements of logic. In: <i>Collected Papers of Charles Sanders Peirce, Volumes I and II: Principles of Philosophy and Elements of Logic</i> , ed. by C. Hartshorne, P. Weiss (Harvard Univ. Press, Cambridge 1932) |
| 3.4 | L. Perini: Scientific representation and the semiotics of pictures. In: <i>New Waves in the Philosophy of Science</i> , ed. by P.D. Magnus, J. Busch (Macmillan, New York 2010) pp. 131–154 | 3.8 | E. Tal: Measurement in science. In: <i>Stanford Encyclopedia of Philosophy</i> , ed. by E.N. Zalta, <a href="http://">http://</a>   |

- [plato.stanford.edu/archives/sum2015/entries/measurement-science/](http://plato.stanford.edu/archives/sum2015/entries/measurement-science/) (Summer 2015 Edition)
- 3.9 T. Knuuttila: Models as Epistemic Artefacts: Toward a Non-Representationalist Account of Scientific Representation, Ph.D. Thesis (Univ. Helsinki, Helsinki 2005)
- 3.10 T. Knuuttila: Modelling and representing: An artefactual approach to model-based representation, *Stud. Hist. Philos. Sci.* **42**, 262–271 (2011)
- 3.11 M. Morgan, M. Morrison (Eds.): *Models as Mediators: Perspectives on Natural and Social Science* (Cambridge Univ. Press, Cambridge 1999)
- 3.12 S. Hartmann: Models as a tool for theory construction: Some strategies of preliminary physics. In: *Theories and Models in Scientific Processes*, Vol. 44, ed. by W.E. Herfel, W. Krajewski, I. Niiniluoto, R. Wojcicki (Rodopi, Amsterdam, Atlanta 1995) pp. 49–67
- 3.13 I. Peschard: Making sense of modeling: Beyond representation, *Eur. J. Philos. Sci.* **1**, 335–352 (2011)
- 3.14 A. Bokulich: Explanatory fictions. In: *Fictions in Science. Philosophical Essays on Modelling and Idealization*, ed. by M. Suárez (Routledge, London, New York 2009) pp. 91–109
- 3.15 A.G. Kennedy: A non representationalist view of model explanation, *Stud. Hist. Philos. Sci.* **43**, 326–332 (2012)
- 3.16 A.I. Woody: More telltale signs: What attention to representation reveals about scientific explanation, *Philos. Sci.* **71**, 780–793 (2004)
- 3.17 J. Reiss: The explanation paradox, *J. Econ. Methodol.* **19**, 43–62 (2012)
- 3.18 M. Lynch, S. Woolgar: *Representation in Scientific Practice* (MIT, Cambridge 1990)
- 3.19 R.N. Giere: No representation without representation, *Biol. Philos.* **9**, 113–120 (1994)
- 3.20 R. Frigg: *Models and Representation: Why Structures Are Not Enough*, Measurement in Physics and Economics Project Discussion Paper, Vol. DP MEAS 25/02 (London School of Economics, London 2002)
- 3.21 R. Frigg: Scientific representation and the semantic view of theories, *Theoria* **55**, 49–65 (2006)
- 3.22 M. Morrison: Models as representational structures. In: *Nancy Cartwright's Philosophy of Science*, ed. by S. Hartmann, C. Hoefer, L. Bovens (Routledge, New York 2008) pp. 67–90
- 3.23 M. Suárez: Scientific representation: Against similarity and isomorphism, *Int. Stud. Philos. Sci.* **17**, 225–244 (2003)
- 3.24 S. Laurence, E. Margolis: Concepts and cognitive science. In: *Concepts: Core Readings*, ed. by S. Laurence, E. Margolis (MIT, Cambridge 1999) pp. 3–81
- 3.25 C. Swoyer: Structural representation and surrogative reasoning, *Synthese* **87**, 449–508 (1991)
- 3.26 C. Callender, J. Cohen: There is no special problem about scientific representation, *Theoria* **55**, 7–25 (2006)
- 3.27 D.M. Bailer-Jones: When scientific models represent, *Int. Stud. Philos. Sci.* **17**, 59–74 (2003)
- 3.28 A. Bolinska: Epistemic representation, informativeness and the aim of faithful representation, *Synthese* **190**, 219–234 (2013)
- 3.29 G. Contessa: Scientific representation, interpretation, and surrogative reasoning, *Philos. Sci.* **74**, 48–68 (2007)
- 3.30 R. Frigg: Re-Presenting Scientific Representation, Ph.D. Thesis (London School of Economics and Political Science, London 2003)
- 3.31 C. Liu: Deflationism on scientific representation. In: *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, ed. by V. Karakostas, D. Dieks (Springer, Dordrecht 2013) pp. 93–102
- 3.32 M. Suárez: An inferential conception of scientific representation, *Philos. Sci.* **71**, 767–779 (2004)
- 3.33 M. Weisberg: *Simulation and Similarity: Using Models to Understand the World* (Oxford Univ. Press, Oxford 2013)
- 3.34 C.Z. Elgin: Telling instances. In: *Beyond Mimesis and Convention: Representation in Art and Science*, ed. by R. Frigg, M.C. Hunter (Springer, Berlin, New York 2010) pp. 1–18
- 3.35 S. French: A model-theoretic account of representation (or, I don't know much about art ... but I know it involves isomorphism), *Philos. Sci.* **70**, 1472–1483 (2003)
- 3.36 B.C. van Fraassen: *Scientific Representation: Paradoxes of Perspective* (Oxford Univ. Press, Oxford 2008)
- 3.37 A.I. Woody: Putting quantum mechanics to work in chemistry: The power of diagrammatic representation, *Philos. Sci.* **67**, S612–S627 (2000)
- 3.38 S. Stich, T. Warfield (Eds.): *Mental Representation: A Reader* (Blackwell, Oxford 1994)
- 3.39 K. Sterelny, P.E. Griffiths: *Sex and Death: An Introduction to Philosophy of Biology* (Univ. Chicago Press, London, Chicago 1999)
- 3.40 E. Wigner: The unreasonable effectiveness of mathematics in the natural sciences, *Commun. Pure Appl. Math.* **13**, 1–14 (1960)
- 3.41 S. Shapiro: *Philosophy of Mathematics: Structure and Ontology* (Oxford Univ. Press, Oxford 1997)
- 3.42 O. Bueno, M. Colyvan: An inferential conception of the application of mathematics, *Nous* **45**, 345–374 (2011)
- 3.43 A. Levy: Modeling without models, *Philos. Stud.* **152**, 781–798 (2015)
- 3.44 I. Hacking: *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science* (Cambridge Univ. Press, Cambridge 1983)
- 3.45 A. Rosenblueth, N. Wiener: The role of models in science, *Philos. Sci.* **12**, 316–321 (1945)
- 3.46 R.A. Ankeny, S. Leonelli: What's so special about model organisms?, *Stud. Hist. Philos. Sci.* **42**, 313–323 (2011)
- 3.47 U. Klein (Ed.): *Tools and Modes of Representation in the Laboratory Sciences* (Kluwer, London, Dordrecht 2001)
- 3.48 A. Toon: Models as make-believe. In: *Beyond Mimesis and Convention: Representation in Art and Science*, ed. by R. Frigg, M. Hunter (Springer, Berlin 2010) pp. 71–96
- 3.49 A. Toon: Similarity and scientific representation, *Int. Stud. Philos. Sci.* **26**, 241–257 (2012)

- 3.50 R. Frigg: Fiction and scientific representation. In: *Beyond Mimesis and Convention: Representation in Art and Science*, ed. by R. Frigg, M. Hunter (Springer, Berlin, New York 2010) pp. 97–138
- 3.51 R.N. Giere: An agent-based conception of models and scientific representation, *Synthese* **172**, 269–281 (2010)
- 3.52 P. Teller: Twilight of the perfect model model, *Erkenntnis* **55**, 393–415 (2001)
- 3.53 O. Bueno, S. French: How theories represent, *Br. J. Philos. Sci.* **62**, 857–894 (2011)
- 3.54 A.F. MacKay: Mr. Donnellan and Humpty Dumpty on referring, *Philos. Rev.* **77**, 197–202 (1968)
- 3.55 K.S. Donnellan: Putting Humpty Dumpty together again, *Philos. Rev.* **77**, 203–215 (1968)
- 3.56 E. Michaelson: This and That: A Theory of Reference for Names, Demonstratives, and Things in Between, Ph.D. Thesis (Univ. California, Los Angeles 2013)
- 3.57 M. Reimer, E. Michaelson: Reference. In: *Stanford Encyclopedia of Philosophy*, ed. by E.N. Zalta, <http://plato.stanford.edu/archives/win2014/entries/reference/> (Winter Edition 2014)
- 3.58 C. Abell: Canny resemblance, *Philos. Rev.* **118**, 183–223 (2009)
- 3.59 D. Lopes: *Understanding Pictures* (Oxford Univ. Press, Oxford 2004)
- 3.60 R.N. Giere: How models are used to represent reality, *Philos. Sci.* **71**, 742–752 (2004)
- 3.61 R.N. Giere: Visual models and scientific judgement. In: *Picturing Knowledge: Historical and Philosophical Problems Concerning the Use of Art in Science*, ed. by B.S. Baigrie (Univ. Toronto Press, Toronto 1996) pp. 269–302
- 3.62 B. Kralemann, C. Lattmann: Models as icons: Modeling models in the semiotic framework of Peirce's theory of signs, *Synthese* **190**, 3397–3420 (2013)
- 3.63 R. Frigg, S. Bradley, H. Du, L.A. Smith: Laplace's demon and the adventures of his apprentices, *Philos. Sci.* **81**, 31–59 (2014)
- 3.64 N. Goodman: *Languages of Art* (Hackett, Indianapolis, Cambridge 1976)
- 3.65 A. Yaghmaie: Reflexive, symmetric and transitive scientific representations, <http://philsci-archival.pitt.edu/9454> (2012)
- 3.66 A. Tversky, I. Gati: Studies of similarity. In: *Cognition and Categorization*, ed. by E. Rosch, B. Lloyd (Lawrence Erlbaum Associates, Hillsdale New Jersey 1978) pp. 79–98
- 3.67 M. Poznic: Representation and similarity: Suárez on necessary and sufficient conditions of scientific representation, *J. Gen. Philos. Sci.* (2015), doi:10.1007/s10838-015-9307-7
- 3.68 H. Putnam: *Reason, Truth, and History* (Cambridge Univ. Press, Cambridge 1981)
- 3.69 M. Black: How do pictures represent? In: *Art, Perception, and Reality*, ed. by E. Gombrich, J. Hochberg, M. Black (Johns Hopkins Univ. Press, London, Baltimore 1973) pp. 95–130
- 3.70 J.L. Aronson, R. Harré, E. Cornell Way: *Realism Rescued: How Scientific Progress is Possible* (Open Court, Chicago 1995)
- 3.71 R.N. Giere: *Explaining Science: A Cognitive Approach* (Chicago Univ. Press, Chicago 1988)
- 3.72 S. Ducheyne: Towards an ontology of scientific models, *Metaphysica* **9**, 119–127 (2008)
- 3.73 N. Cartwright: *The Dappled World: A Study of the Boundaries of Science* (Cambridge Univ. Press, Cambridge 1999)
- 3.74 N. Cartwright: Models and the limits of theory: Quantum hamiltonians and the BCS models of superconductivity. In: *Models as Mediators: Perspectives on Natural and Social Science*, ed. by M. Morgan, M. Morrison (Cambridge Univ. Press, Cambridge 1999) pp. 241–281
- 3.75 L. Apostel: Towards the formal study of models in the non-formal sciences. In: *The Concept and the Role of the Model in Mathematics and Natural and Social Sciences*, ed. by H. Freudenthal (Reidel, Dordrecht 1961) pp. 1–37
- 3.76 A.-M. Rusanen, O. Lappi: An information semantic account of scientific models. In: *EPSA Philosophy of Science: Amsterdam 2009*, ed. by H.W. de Regt, S. Hartmann, S. Okasha (Springer, Dordrecht 2012) pp. 315–328
- 3.77 B.C. van Fraassen: *The Empirical Stance* (Yale Univ. Press, New Haven, London 2002)
- 3.78 H. Putnam: *The Collapse of the Fact-Value Distinction* (Harvard Univ. Press, Cambridge 2002)
- 3.79 U. Mäki: Models and the locus of their truth, *Synthese* **180**, 47–63 (2011)
- 3.80 S.M. Downes: Models, pictures, and unified accounts of representation: Lessons from aesthetics for philosophy of science, *Perspect. Sci.* **17**, 417–428 (2009)
- 3.81 M. Morreau: It simply does not add up: The trouble with overall similarity, *J. Philos.* **107**, 469–490 (2010)
- 3.82 W.V.O. Quine: *Ontological Relativity and Other Essays* (Columbia Univ. Press, New York 1969)
- 3.83 N. Goodman: Seven strictures on similarity. In: *Problems and Projects*, ed. by N. Goodman (Bobbs-Merrill, Indianapolis, New York 1972) pp. 437–446
- 3.84 L. Decock, I. Douven: Similarity after Goodman, *Rev. Philos. Psychol.* **2**, 61–75 (2011)
- 3.85 R.N. Shepard: Multidimensional scaling, tree-fitting, and clustering, *Science* **210**, 390–398 (1980)
- 3.86 A. Tversky: Features of similarity, *Psychol. Rev.* **84**, 327–352 (1977)
- 3.87 M. Weisberg: Getting serious about similarity, *Philos. Sci.* **79**, 785–794 (2012)
- 3.88 M. Hesse: *Models and Analogies in Science* (Sheed Ward, London 1963)
- 3.89 W. Parker: Getting (even more) serious about similarity, *Biol. Philos.* **30**, 267–276 (2015)
- 3.90 I. Niiniluoto: Analogy and similarity in scientific reasoning. In: *In Analogical Reasoning: Perspectives of Artificial Intelligence, Cognitive Science, and Philosophy*, ed. by D.H. Helman (Kluwer, Dordrecht 1988) pp. 271–298
- 3.91 M. Weisberg: Biology and philosophy symposium on simulation and similarity: Using models to understand the world: Response to critics, *Biol. Philos.* **30**, 299–310 (2015)

- 3.92 A. Toon: Playing with molecules, *Stud. Hist. Philos. Sci.* **42**, 580–589 (2011)
- 3.93 M. Morgan, T. Knuuttila: Models and modelling in economics. In: *Philosophy of Economics*, ed. by U. Mäki (Elsevier, Amsterdam 2012) pp. 49–87
- 3.94 M. Thomson-Jones: Modeling without mathematics, *Philos. Sci.* **79**, 761–772 (2012)
- 3.95 G. Rosen: Abstract objects. In: *The Stanford Encyclopedia of Philosophy*, ed. by E.N. Zalta, <http://plato.stanford.edu/archives/fall2014/entries/abstract-objects/> (Fall 2014 Edition)
- 3.96 S. Hale: Spacetime and the abstract-concrete distinction, *Philos. Stud.* **53**, 85–102 (1988)
- 3.97 R.N. Giere: Why scientific models should not be regarded as works of fiction. In: *Fictions in Science. Philosophical Essays on Modelling and Idealization*, ed. by M. Suárez (Routledge, London 2009) pp. 248–258
- 3.98 M. Thomson-Jones: Missing systems and face value practise, *Synthese* **172**, 283–299 (2010)
- 3.99 D.M. Armstrong: *Universals: An Opinionated Introduction* (Westview, London 1989)
- 3.100 P. Suppes: *Representation and Invariance of Scientific Structures* (CSLI Publications, Stanford 2002)
- 3.101 B.C. van Fraassen: *The Scientific Image* (Oxford Univ. Press, Oxford 1980)
- 3.102 N.C.A. Da Costa, S. French: *Science and Partial Truth: A Unitary Approach to Models and Scientific Reasoning* (Oxford Univ. Press, Oxford 2003)
- 3.103 H. Byerly: Model-structures and model-objects, *Br. J. Philos. Sci.* **20**, 135–144 (1969)
- 3.104 A. Chakravartty: The semantic or model-theoretic view of theories and scientific realism, *Synthese* **127**, 325–345 (2001)
- 3.105 C. Klein: Multiple realizability and the semantic view of theories, *Philos. Stud.* **163**, 683–695 (2013)
- 3.106 D. Portides: Scientific models and the semantic view of theories, *Philos. Sci.* **72**, 1287–1289 (2005)
- 3.107 D. Portides: Why the model-theoretic view of theories does not adequately depict the methodology of theory application. In: *EPSA Epistemology and Methodology of Science*, ed. by M. Suárez, M. Dorato, M. Rédei (Springer, Dordrecht 2010) pp. 211–220
- 3.108 M.D. Resnik: *Mathematics as a Science of Patterns* (Oxford Univ. Press, Oxford 1997)
- 3.109 S. Shapiro: *Thinking About Mathematics* (Oxford Univ. Press, Oxford 2000)
- 3.110 M. Thomson-Jones: Structuralism about scientific representation. In: *Scientific Structuralism*, ed. by A. Bokulich, P. Bokulich (Springer, Dordrecht 2011) pp. 119–141
- 3.111 M. Machover: *Set Theory, Logic and Their Limitations* (Cambridge Univ. Press, Cambridge 1996)
- 3.112 W. Hodges: *A Shorter Model Theory* (Cambridge Univ. Press, Cambridge 1997)
- 3.113 C.E. Rickart: *Structuralism and Structure: A Mathematical Perspective* (World Scientific Publishing, Singapore 1995)
- 3.114 G.S. Boolos, R.C. Jeffrey: *Computability and Logic* (Cambridge Univ. Press, Cambridge 1989)
- 3.115 B. Russell: *Introduction to Mathematical Philosophy* (Routledge, London, New York 1993)
- 3.116 H.B. Enderton: *A Mathematical Introduction to Logic* (Harcourt, San Diego, New York 2001)
- 3.117 P. Suppes: A comparison of the meaning and uses of models in mathematics and the empirical sciences. In: *Studies in the Methodology and Foundations of Science: Selected Papers from 1951 to 1969*, ed. by P. Suppes (Reidel, Dordrecht 1969) pp. 10–23, 1960
- 3.118 B.C. van Fraassen: Structure and perspective: Philosophical perplexity and paradox. In: *Logic and Scientific Methods*, ed. by M.L. Dalla Chiara (Kluwer, Dordrecht 1997) pp. 511–530
- 3.119 M. Redhead: The intelligibility of the universe. In: *Philosophy at the New Millennium*, ed. by A. O'Hear (Cambridge Univ. Press, Cambridge 2001)
- 3.120 S. French, J. Ladyman: Reinflating the semantic approach, *Int. Stud. Philos. Sci.* **13**, 103–121 (1999)
- 3.121 N.C.A. Da Costa, S. French: The model-theoretic approach to the philosophy of science, *Philos. Sci.* **57**, 248–265 (1990)
- 3.122 P. Suppes: Models of data. In: *Studies in the Methodology and Foundations of Science: Selected Papers from 1951 to 1969*, ed. by P. Suppes (Reidel, Dordrecht 1969) pp. 24–35, 1962
- 3.123 P. Suppes: *Set-Theoretical Structures in Science* (Stanford Univ., Stanford 1970), lecture notes
- 3.124 B.C. van Fraassen: *Quantum Mechanics: An Empiricist View* (Oxford Univ. Press, Oxford 1991)
- 3.125 B.C. van Fraassen: A philosophical approach to foundations of science, *Found. Sci.* **1**, 5–9 (1995)
- 3.126 N.C.A. Da Costa, S. French: Models, theories, and structures: Thirty years on, *Philos. Sci.* **67**, 116–127 (2000)
- 3.127 M. Dummett: *Frege: Philosophy of Mathematics* (Duckworth, London 1991)
- 3.128 G. Hellman: *Mathematics Without Numbers: Towards a Modal-Structural Interpretation* (Oxford Univ. Press, Oxford 1989)
- 3.129 G. Hellman: Structuralism without structures, *Philos. Math.* **4**, 100–123 (1996)
- 3.130 O. Bueno, S. French, J. Ladyman: On representing the relationship between the mathematical and the empirical, *Philos. Sci.* **69**, 452–473 (2002)
- 3.131 S. French: Keeping quiet on the ontology of models, *Synthese* **172**, 231–249 (2010)
- 3.132 S. French, J. Saatsi: Realism about structure: The semantic view and nonlinguistic representations, *Philos. Sci.* **73**, 548–559 (2006)
- 3.133 S. French, P. Vickers: Are there no things that are scientific theories?, *Br. J. Philos. Sci.* **62**, 771–804 (2011)
- 3.134 E. Landry: Shared structure need not be shared set-structure, *Synthese* **158**, 1–17 (2007)
- 3.135 H. Halvorson: What scientific theories could not be, *Philos. Sci.* **79**, 183–206 (2012)
- 3.136 H. Halvorson: Scientific theories. In: *The Oxford Handbook of Philosophy of Science*, ed. by P. Humphreys (Oxford Univ. Press, Oxford 2016)
- 3.137 K. Brading, E. Landry: Scientific structuralism: Presentation and representation, *Philos. Sci.* **73**, 571–581 (2006)

- 3.138 C. Glymour: Theoretical equivalence and the semantic view of theories, *Philos. Sci.* **80**, 286–297 (2013)
- 3.139 J.B. Ubbink: Model, description and knowledge, *Synthese* **12**, 302–319 (1960)
- 3.140 A. Bartels: Defending the structural concept of representation, *Theoria* **21**, 7–19 (2006)
- 3.141 E. Lloyd: A semantic approach to the structure of population genetics, *Philos. Sci.* **51**, 242–264 (1984)
- 3.142 B. Mundy: On the general theory of meaningful representation, *Synthese* **67**, 391–437 (1986)
- 3.143 S. French: The reasonable effectiveness of mathematics: Partial structures and the application of group theory to physics, *Synthese* **125**, 103–120 (2000)
- 3.144 O. Bueno: Empirical adequacy: A partial structure approach, *Stud. Hist. Philos. Sci.* **28**, 585–610 (1997)
- 3.145 O. Bueno: What is structural empiricism? Scientific change in an empiricist setting, *Erkenntnis* **50**, 59–85 (1999)
- 3.146 F. Pero, M. Suárez: Varieties of misrepresentation and homomorphism, *Eur. J. Philos. Sci.* **6**(1), 71–90 (2016)
- 3.147 P. Kroes: Structural analogies between physical systems, *Br. J. Philos. Sci.* **40**, 145–154 (1989)
- 3.148 F.A. Muller: Reflections on the revolution at Stanford, *Synthese* **183**, 87–114 (2011)
- 3.149 E.W. Adams: The foundations of rigid body mechanics and the derivation of its laws from those of particle mechanics. In: *The Axiomatic Method: With Special Reference to Geometry and Physics*, ed. by L. Henkin, P. Suppes, A. Tarski (North-Holland, Amsterdam 1959) pp. 250–265
- 3.150 O. Bueno: Models and scientific representations. In: *New Waves in Philosophy of Science*, ed. by P.D. Magnus, J. Busch (Pelgrave MacMillan, Hampshire 2010) pp. 94–111
- 3.151 M. Budd: How pictures look. In: *Virtue and Taste*, ed. by D. Knowles, J. Skorupski (Blackwell, Oxford 1993) pp. 154–175
- 3.152 P. Godfrey-Smith: The strategy of model-based science, *Biol. Philos.* **21**, 725–740 (2006)
- 3.153 T. Harris: Data models and the acquisition and manipulation of data, *Philos. Sci.* **70**, 1508–1517 (2003)
- 3.154 B.C. van Fraassen: Theory construction and experiment: An empiricist view, *Proc. Philos. Sci.* **2**, 663–677 (1981)
- 3.155 B.C. van Fraassen: *Laws and Symmetry* (Clarendon, Oxford 1989)
- 3.156 B.C. van Fraassen: Empiricism in the philosophy of science. In: *Images of Science: Essays on Realism and Empiricism with a Reply from Bas C. van Fraassen*, ed. by P.M. Churchland, C.A. Hooker (Univ. Chicago Press, London, Chicago 1985) pp. 245–308
- 3.157 J. Bogen, J. Woodward: Saving the phenomena, *Philos. Rev.* **97**, 303–352 (1988)
- 3.158 J. Woodward: Data and phenomena, *Synthese* **79**, 393–472 (1989)
- 3.159 P. Teller: Whither constructive empiricism, *Philos. Stud.* **106**, 123–150 (2001)
- 3.160 J.W. McAllister: Phenomena and patterns in data sets, *Erkenntnis* **47**, 217–228 (1997)
- 3.161 J. Nguyen: On the pragmatic equivalence between representing data and phenomena, *Philos. Sci.* **83**, 171–191 (2016)
- 3.162 M. Frisch: Users, structures, and representation, *Br. J. Philos. Sci.* **66**, 285–306 (2015)
- 3.163 W. Balzer, C.U. Moulines, J.D. Sneed: *An Architectonic for Science the Structuralist Program* (D. Reidel, Dordrecht 1987)
- 3.164 W. Demopoulos: On the rational reconstruction of our theoretical knowledge, *Br. J. Philos. Sci.* **54**, 371–403 (2003)
- 3.165 J. Ketland: Empirical adequacy and ramsification, *Br. J. Philos. Sci.* **55**, 287–300 (2004)
- 3.166 R. Frigg, I. Votsis: Everything you always wanted to know about structural realism but were afraid to ask, *Eur. J. Philos. Sci.* **1**, 227–276 (2011)
- 3.167 P. Ainsworth: Newman's objection, *Br. J. Philos. Sci.* **60**, 135–171 (2009)
- 3.168 S. Shapiro: Mathematics and reality, *Philos. Sci.* **50**, 523–548 (1983)
- 3.169 S. French: *The Structure of the World. Metaphysics and Representation* (Oxford Univ. Press, Oxford 2014)
- 3.170 M. Tegmark: The mathematical universe, *Found. Phys.* **38**, 101–150 (2008)
- 3.171 M. Suárez, A. Solé: On the analogy between cognitive representation and truth, *Theoria* **55**, 39–48 (2006)
- 3.172 M. Suárez: Deflationary representation, inference, and practice, *Stud. Hist. Philos. Sci.* **49**, 36–47 (2015)
- 3.173 A. Chakravartty: Informational versus functional theories of scientific representation, *Synthese* **172**, 197–213 (2010)
- 3.174 W. Künne: *Conceptions of Truth* (Clarendon, Oxford 2003)
- 3.175 R.B. Brandom: *Making it Explicit: Reasoning, Representing and Discursive Commitment* (Harvard Univ. Press, Cambridge 1994)
- 3.176 R.B. Brandom: *Articulating Reasons: An Introduction to Inferentialism* (Harvard Univ. Press, Cambridge 2000)
- 3.177 X. de Donato Rodriguez, J. Zamora Bonilla: Credibility, idealisation, and model building: An inferential approach, *Erkenntnis* **70**, 101–118 (2009)
- 3.178 M. Suárez: Scientific Representation, *Philos. Compass* **5**, 91–101 (2010)
- 3.179 G. Contessa: Scientific models and representation. In: *The Continuum Companion to the Philosophy of Science*, ed. by S. French, J. Saatsi (Continuum Press, London 2011) pp. 120–137
- 3.180 G. Contessa: Scientific models and fictional objects, *Synthese* **172**, 215–229 (2010)
- 3.181 E. Shech: Scientific misrepresentation and guides to ontology: The need for representational code and contents, *Synthese* **192**(11), 3463–3485 (2015)
- 3.182 S. Ducheyne: Scientific representations as limiting cases, *Erkenntnis* **76**, 73–89 (2012)
- 3.183 R.F. Hendry: Models and approximations in quantum chemistry. In: *Idealization IX: Idealization in Contemporary Physics*, ed. by N. Shanks (Rodopi,

- Amsterdam 1998) pp. 123–142
- 3.184 R. Laymon: Computer simulations, idealizations and approximations, *Proc. Bienn. Meet. Philos. Sci. Assoc.*, Vol. 2 (1990) pp. 519–534
- 3.185 C. Liu: Explaining the emergence of cooperative phenomena, *Philos. Sci.* **66**, S92–S106 (1999)
- 3.186 J. Norton: Approximation and idealization: Why the difference matters, *Philos. Sci.* **79**, 207–232 (2012)
- 3.187 J.L. Ramsey: Approximation. In: *The Philosophy of Science: An Encyclopedia*, ed. by S. Sarkar, J. Pfeifer (Routledge, New York 2006) pp. 24–27
- 3.188 R.I.G. Hughes: Models and representation, *Philos. Sci.* **64**, S325–S336 (1997)
- 3.189 R.I.G. Hughes: *The Theoretical Practises of Physics: Philosophical Essays* (Oxford Univ. Press, Oxford 2010)
- 3.190 R.I.G. Hughes: Laws of nature, laws of physics, and the representational account of theories, *ProtoSociology* **12**, 113–143 (1998)
- 3.191 L.A. Smith: *Chaos: A Very Short Introduction* (Oxford Univ. Press, Oxford 2007)
- 3.192 W.D. Niven (Ed.): *The Scientific Papers of James Clerk Maxwell* (Dover Publications, New York 1965)
- 3.193 H. Vaihinger: *The Philosophy of as if: A System of the Theoretical, Practical, and Religious Fictions of Mankind* (Kegan Paul, London 1911) p. 1924, English translation
- 3.194 N. Cartwright: *How the Laws of Physics Lie* (Oxford Univ. Press, Oxford 1983)
- 3.195 D.N. McCloskey: Storytelling in economics. In: *Narrative in Culture. The uses of Storytelling in the Sciences, Philosophy, and Literature*, ed. by C. Nash (Routledge, London 1990) pp. 5–22
- 3.196 A. Fine: Fictionalism, *Midwest Stud. Philos.* **18**, 1–18 (1993)
- 3.197 L. Sklar: *Theory and Truth. Philosophical Critique Within Foundational Science* (Oxford Univ. Press, Oxford 2000)
- 3.198 C.Z. Elgin: *Considered Judgement* (Princeton Univ. Press, Princeton 1996)
- 3.199 S. Hartmann: Models and stories in hadron physics. In: *Models as Mediators. Perspectives on Natural and Social Science*, ed. by M. Morgan, M. Morrison (Cambridge Univ. Press, Cambridge 1999) pp. 326–346
- 3.200 M. Morgan: Models, stories and the economic world, *J. Econ. Methodol.* **8**, 361–384 (2001)
- 3.201 M. Morgan: Imagination and imaging in model building, *Philos. Sci.* **71**, 753–766 (2004)
- 3.202 R. Sugden: Credible worlds: The status of theoretical models in economics, *J. Econ. Methodol.* **7**, 1–31 (2000)
- 3.203 R. Frigg: Models and fiction, *Synthese* **172**, 251–268 (2010)
- 3.204 T. Grüne-Yanoff, P. Schweinzer: The roles of stories in applying game theory, *J. Econ. Methodol.* **15**, 131–146 (2008)
- 3.205 A. Toon: *Models as Make-Believe. Imagination, Fiction and Scientific Representation* (Palgrave Macmillan, Basingstoke 2012)
- 3.206 A. Levy: Models, fictions, and realism: Two packages, *Philos. Sci.* **79**, 738–748 (2012)
- 3.207 S. Friend: Fictional characters, *Philos. Compass* **2**, 141–156 (2007)
- 3.208 F. Salis: Fictional entities. In: *Online Companion to Problems in Analytical Philosophy*, ed. by J. Branquinho, R. Santos, doi:10.13140/2.1.1931.9040 (2014)
- 3.209 P. Godfrey-Smith: Models and fictions in science, *Philos. Stud.* **143**, 101–116 (2009)
- 3.210 R. Sugden: Credible worlds, capacities and mechanisms, *Erkenntnis* **70**, 3–27 (2009)
- 3.211 J. Cat: Who's afraid of scientific fictions?: Mauricio Suárez (Ed.): *Fictions in Science. Philosophical Essays on Modeling and Idealization*, *J. Gen. Philos. Sci.* **43**, 187–194 (2012), book review
- 3.212 C. Liu: A Study of model and representation based on a Duhemian thesis. In: *Philosophy and Cognitive Science: Western and Eastern studies*, ed. by L. Magnani, P. Li (Springer, Berlin, Heidelberg 2012) pp. 115–141
- 3.213 C. Liu: Symbolic versus modelistic elements in scientific modeling, *Theoria* **30**, 287–300 (2015)
- 3.214 C. Pincock: *Mathematics and Scientific Representation* (Oxford Univ. Press, Oxford 2012)
- 3.215 M. Vorms: Representing with imaginary models: Formats matter, *Stud. Hist. Philos. Sci.* **42**, 287–295 (2011)
- 3.216 M. Vorms: Formats of representation in scientific theorising. In: *Models, Simulations, and Representations*, ed. by P. Humphreys, C. Imbert (Routledge, New York 2012) pp. 250–274
- 3.217 T. Knuuttila, M. Boon: How do models give us knowledge? The case of Carnot's ideal heat engine, *Eur. J. Philos. Sci.* **1**, 309–334 (2011)
- 3.218 R. Frigg: Fiction in science. In: *Fictions and Models: New Essays*, ed. by J. Woods (Philosophia, Munich 2010) pp. 247–287
- 3.219 M.E. Kalderon (Ed.): *Fictionalism in Metaphysics* (Oxford Univ. Press, Oxford 2005)
- 3.220 A. Fine: Fictionalism. In: *Routledge Encyclopedia of Philosophy*, ed. by E. Craig (Routledge, London 1998)
- 3.221 A. Fine: Science fictions: Comment on Godfrey-Smith, *Philos. Stud.* **143**, 117–125 (2009)
- 3.222 E. Winsberg: A function for fictions: Expanding the scope of science. In: *Fictions in Science: Philosophical Essays in on Modeling and Idealization*, ed. by M. Suárez (Routledge, New York 2009) pp. 179–191
- 3.223 M. Suárez: Scientific fictions as rules of inference. In: *Fictions in Science: Philosophical Essays in on Modeling and Idealization*, ed. by M. Suárez (Routledge, New York 2009) pp. 158–178
- 3.224 M. Morrison: Fictions, representations, and reality. In: *Fictions in Science: Philosophical Essays on Modeling and Idealization*, ed. by M. Suárez (Routledge, New York 2009) pp. 110–135
- 3.225 G.M. Purves: Finding truth in fictions: Identifying non-fictions in imaginary cracks, *Synthese* **190**, 235–251 (2013)
- 3.226 J. Woods: Against fictionalism. In: *Model-Based Reasoning in Science and Technology: Theoretical and Cognitive Issues*, ed. by L. Magnani (Springer, Berlin, Heidelberg 2014) pp. 9–42

- 3.227 M. Weisberg: Who is a modeler?, *Br. J. Philos. Sci.* **58**, 207–233 (2007)
- 3.228 A. Toon: The ontology of theoretical modelling: Models as make-believe, *Synthese* **172**, 301–315 (2010)
- 3.229 K.L. Walton: *Mimesis as Make-Believe: On the Foundations of the Representational Arts* (Harvard Univ. Press, Cambridge 1990)
- 3.230 S. Yablo: *Aboutness* (Princeton Univ. Press, Princeton 2014)
- 3.231 N. Cartwright: Models: Parables v fables. In: *Beyond Mimesis and Convention. Representation in Art and Science*, ed. by R. Frigg, M.C. Hunter (Springer, Berlin, New York 2010) pp. 19–32
- 3.232 T. Schelling: *Micromotives and Macrobehavior* (Norton, New York 1978)
- 3.233 C.A. Pissarides: Loss of skill during unemployment and the persistence of unemployment shocks, *Q. J. Econ.* **107**, 1371–1391 (1992)
- 3.234 T. Grüne-Yanoff: Learning from minimal economic models, *Erkenntnis* **70**, 81–99 (2009)
- 3.235 T. Knuuttila: Isolating representations versus credible constructions? *Economic modelling in theory and practice*, *Erkenntnis* **70**, 59–80 (2009)
- 3.236 L. Magnani: Scientific models are not fictions: Model-based science as epistemic warfare. In: *Philosophy and Cognitive Science: Western and Eastern Studies*, ed. by L. Magnani, P. Li (Springer, Berlin, Heidelberg 2012) pp. 1–38
- 3.237 J. Odenbaugh: Semblance or similarity?, *Reflections on simulation and similarity*, *Biol. Philos.* **30**, 277–291 (2015)
- 3.238 R. Frigg, J. Nguyen: Scientific representation is representation as. In: *Philosophy of Science in Practice: Nancy Cartwright and the Nature of Scientific Reasoning*, ed. by H.-K. Chao, R. Julian, C. Szu-Ting (Springer, New York 2017), in press
- 3.239 C.Z. Elgin: *With Reference to Reference* (Hackett, Indianapolis 1983)
- 3.240 F. Schier: *Deeper in Pictures: An Essay on Pictorial Representation* (Cambridge Univ. Press, Cambridge 1986)
- 3.241 J. Kulvicki: Pictorial representation, *Philos. Compass* **1**, 535–546 (2006)
- 3.242 M. Jones: Idealization and abstraction: A framework. In: *Idealization XII: Correcting the Model-Idealization and Abstraction in the Sciences*, ed. by M. Jones, N. Cartwright (Rodopi, Amsterdam 2005) pp. 173–218
- 3.243 A. Galton: Space, time, and the representation of geographical reality, *Topoi* **20**, 173–187 (2001)
- 3.244 S. Sismondo, N. Chrisman: Deflationary metaphysics and the nature of maps, *Proc. Philos. Sci.* **68**, 38–49 (2001)
- 3.245 S.G. Sterrett: Models of machines and models of phenomena, *Int. Stud. Philos. Sci.* **20**, 69–80 (2006)
- 3.246 J.H. Argyris, G. Faust, M. Haase: *Die Erforschung des Chaos: Eine Einführung für Naturwissenschaftler und Ingenieure* (Vieweg Teubner, Braunschweig 1994)
- 3.247 R. Frigg, J. Nguyen: *The Turn of the Valve: Representing with Material Models*, Unpublished Manuscript
- 3.248 R. Frigg, J. Nguyen: The fiction view of models reloaded, forthcoming in *The Monist*, July 2016



# Models and Explanation

## 4. Models and Explanation

Alisa Bokulich

Detailed examinations of scientific practice have revealed that the use of idealized models in the sciences is pervasive. These models play a central role in not only the investigation and prediction of phenomena, but also in their received scientific explanations. This has led philosophers of science to begin revising the traditional philosophical accounts of scientific explanation in order to make sense of this practice. These new model-based accounts of scientific explanation, however, raise a number of key questions: Can the fictions and falsehoods inherent in the modeling practice do real explanatory work? Do some highly abstract and mathematical models exhibit a non-causal form of scientific explanation? How can one distinguish an exploratory *how-possibly* model explanation from a genuine *how-actually* model explanation? Do modelers face tradeoffs such that a model that is optimized for yielding explanatory insight, for example, might fail to be the most predictively accurate, and vice versa? This chapter explores the various answers that have been given to these questions.

4.1	<b>The Explanatory Function of Models</b> .....	104
4.2	<b>Explanatory Fictions: Can Falsehoods Explain?</b> .....	108
4.3	<b>Explanatory Models and Noncausal Explanations</b> .....	112
4.4	<b>How-Possibly versus How-Actually Model Explanations</b> .....	114
4.5	<b>Tradeoffs in Modeling: Explanation versus Other Functions for Models</b> .....	115
4.6	<b>Conclusion</b> .....	116
	<b>References</b> .....	117

Explanation is one of the central aims of science, and the attempt to understand the nature of scientific explanation is at the heart of the philosophy of science. An explanation can be analyzed as consisting of two parts, a phenomenon or event to be explained, known as the *explanandum*, and that which does the job of explaining, the *explanans*. On the traditional approach, to explain a phenomenon is either to deduce the explanandum phenomenon from the relevant laws of nature and initial conditions, such as on the deductive-nomological (DN) account [4.1], or to trace the detailed causal chain leading up to that event, such as on the causal-mechanical account [4.2]. Underlying this traditional approach are the assumptions that, in order to genuinely explain, the explanans must be entirely true, and that the more complete and detailed the explanans is, the better the scientific explanation.

As philosophers of science have turned to more careful examinations of actual scientific practice, however, there have been three key observations that have challenged this traditional approach: first, many of the phenomena scientists seek to explain are incredibly complex; second, the laws of nature supposedly needed for explanation are either few and far between or entirely absent in many of the sciences; and third, a detailed causal description of the chain of events and interactions leading up to a phenomenon are often either beyond our grasp or not in fact what is most important for a scientific understanding of the phenomenon.

More generally, there has been a growing recognition that much of science is a model-based activity. (For an overview of many different types of models in science, and some of the philosophical issues regarding the nature and use of such models, refer to [4.3]).

Models are by definition incomplete and idealized descriptions of the systems they describe. This practice raises all sorts of epistemological questions, such as how can it be that false models lead to true insights?

And most relevant to our discussion here, how might the extensive use of models in science lead us to revise our philosophical account of scientific explanation?

## 4.1 The Explanatory Function of Models

Model-based explanations (or model explanations, for short) are explanations in which the explanans appeal to certain properties or behaviors observed in an idealized model or computer simulation as part of an explanation for why the (typically real-world) explanandum phenomenon exhibits the features that it does. For example, one might explain why sparrows of a certain species vary in their feather coloration from pale to dark by appealing to a particular game theory model: although coloration is unrelated to fitness, such a polymorphism can be a badge of status that allows the sparrows to avoid unnecessary conflicts over resources; dark birds are dominant and displace the pale birds from food sources. The model demonstrates that such a strategy is stable and successful, and hence can be used as part of the explanation for why we find this polymorphism among sparrows (see [4.4, 5] for further discussion).

There are, of course, many perils in assuming that just because we see a phenomenon or pattern exhibited in a model that it therefore explains why we see it in the real world: the same pattern or phenomenon could be produced in multiple, very different ways, and hence it might be only a phenomenological model at best, useful for prediction, but not a genuine explanation. Explanation and the concomitant notion of understanding are what we call success terms: if the purported explanation is not, in fact, right (right in some sense that will need to be spelled out) and the understanding is only illusory, then it is not, in fact, a genuine explanation. Determining what the success conditions are for a genuine explanation is the central philosophical problem in scientific explanation.

Those who have defended the explanatory power of models have typically argued that further conditions must be met in order for a model's exhibiting of a salient pattern or phenomenon to count as part of a genuine explanation of its real-world counterpart. Not all models are explanatory, and an adequate account of model explanation must provide grounds for making such discriminations. As we will see, however, different approaches have filled in these further requirements in different ways.

One of the earliest defenses of the view that models can explain is McMullin's [4.6] *hypothetico-structural* HS account of model explanations. In an HS explanation, one explains a complex phenomenon by pos-

tulating an underlying structural model whose features are causally responsible for the phenomenon to be explained. McMullin notes that such models are often tentative or metaphorical, but that a good model explanation will lay out a research program for the further refinement of the model. On his account, the justification of the model as genuinely explanatory involves a process known as de-idealization, where features that were left out are added back or a more realistic representation of those processes is given. More specifically he requires that one be able to give a theoretical justification for this de-idealization process, so that it is not merely an ad hoc fitting of the model to the data. He writes [4.7, p. 261]:

“If techniques for which no theoretical justification can be given have to be utilized to correct a formal idealization, this is taken to count against the explanatory propriety of that idealization. The model itself in such a case is suspect, no matter how good the predictive results it may produce.”

He further notes that a theoretical justification for the de-idealization process will only succeed if the original model has successfully captured the real structure of the phenomenon of interest.

As an example, McMullin [4.8] describes the fertility of the continental drift model in explaining why the continents seem to fit together like pieces of a puzzle and why similar fossils are found at distant locations. The continental drift model involved all sorts of idealizations and gaps: most notably, the chief proponent of this approach, Alfred Wegener, could offer no account of the forces or mechanisms by which the massive continents could move. Strictly speaking, we now know that the continental drift model is false, and has been supplanted by plate tectonics. But as McMullin notes, the continental drift model nonetheless captures key features of the real structure of the phenomenon of interest, and, hence, succeeds in giving genuine explanatory insight.

While McMullin's account of HS model explanations fits in many cases, there are other examples of model explanations in the sciences that do not seem to fit his account. First, there seem to be examples of model explanations where the idealizations are ineliminable, and, hence, they cannot be justified through

anything like the de-idealization analysis that McMullin describes [4.9]. Second, not all models are related to their target phenomena via an idealization: some models represent through a fictionalization [4.10]. Third, insofar as McMullin's HS model explanations are a subspecies of causal explanations, they do not account for noncausal model explanations. These sort of cases will be discussed more fully in subsequent sections.

Another early account of the explanatory power of models is Cartwright's [4.11] *simulacrum* account of explanation, which she introduces as an alternative to the DN account of explanation and elaborates in her book *How the Laws of Physics Lie*. Drawing on Duhem's [4.12] theory of explanation, she argues [4.11, p. 152]:

“To explain a phenomenon is to find a model that fits it into the basic framework of the theory and that thus allows us to derive analogues for the messy and complicated phenomenological laws which are true of it.”

According to Cartwright, the laws of physics do not describe our real messy world, only the idealized world we construct in our models. She gives the example of the harmonic oscillator model, which is used in quantum mechanics to describe a wide variety of systems. One describes a real-world helium-neon laser as if it were a van der Pol oscillator; this is how the phenomenon becomes tractable and we are able to make use of the mathematical framework of our theory. The laws of quantum mechanics are true in this model, but this model is just a simulacrum of the real-world phenomenon. By *model*, Cartwright means “an especially prepared, usually fictional description of the system under study” [4.11, p. 158]. She notes that while some of the properties ascribed to the objects in the models are idealizations, there are other properties that are pure fictions; hence, one should not think of models in terms of idealizations alone.

Although Cartwright's simulacrum account is highly suggestive, it leaves unanswered many key questions, such as when a model should or should not be counted as explanatory. Elgin and Sober [4.13] offer a possible emendation to Cartwright's account that they argue discriminates which sorts of idealized causal models can explain. The key, according to their approach, is to determine whether or not the idealizations in the model are what they call *harmless*. A harmless idealization is one that if corrected “wouldn't make much difference in the predicted value of the effect variable” [4.13, p. 448]. They illustrate this approach using the example of optimality models in evolutionary biology. Optimality models are models that determine what value of a trait maximizes fitness (is optimal) for

an organism given certain constraints (e.g., the optimal length of a bear's fur, given the benefits of longer fur and the costs of growing it, or the optimal height at which crows should drop walnuts in order to crack open the shells, given the costs of flying higher, etc.). If organisms are indeed fitter the closer a trait is to the optimal value, and if natural selection is the only force operating, then the optimal value for that trait will evolve in the population. Thus, optimality models are used to explain why organisms have trait values at or near the optimal value (e.g., why crows drop walnuts from an average of 3 m high [4.14]).

As Elgin and Sober note, optimality models contain all sorts of idealizations: “they describe evolutionary trajectories of populations that are infinitely large in which reproduction is asexual with offspring always resembling their parents, etc.” [4.13, p. 447]. Nonetheless, they argue that these models are genuinely explanatory when it can be shown that the value described in the explanandum is close to the value predicted by the idealized model; when this happens we can conclude that the idealizations in the model are harmless [4.13, p. 448]. Apart from this concession about *harmless* idealizations, Elgin and Sober's account of explanation remains close to the traditional DN account in that they further require:

1. The explanans must cite the cause of the explanandum
2. The explanans must cite a law
3. All of the explanans propositions must be true [4.13, p. 446]

though their condition 3 might better be stated as all the explanans propositions are *either true or harmlessly false*.

As a general account of model explanations, however, one might argue that the approaches of Cartwright, Elgin, and Sober are too restrictive. As noted before, this approach still depends on there being laws of nature from which the phenomenon is to be derived, and such laws just might not be available. Moreover, it is not clear that explanatory models will contain only harmless idealizations. There may very well be cases in which the idealizations make a difference (are not harmless) and yet are essential to the explanation (e.g., [4.15, 16]).

While the simulacrum approach of Cartwright, especially as further developed by Elgin and Sober, largely draws its inspiration from the traditional DN approach to explanation, there are other approaches to model explanation that are tied more closely to the traditional causal-mechanical approach to explanation. Craver [4.17], for example, has argued that models are explanatory when they describe mechanisms. He writes “[...] the distinction between explanatory and nonex-

planatory models is that the [former], and not the [latter] describe mechanisms” [4.17, p. 367]. The central notion of mechanism, here, can be understood as consisting of the various components or parts of the phenomenon of interest, the activities of those components, and how they are organized in relation to each other.

Craver imposes rather strict conditions on when such mechanistic models can be counted as explanatory; he writes, “To characterize the phenomenon correctly and completely is the first restrictive step in turning a model into an acceptable mechanistic explanation” [4.17, p. 369]. (Some have argued that if one has a complete and accurate description of the system or phenomenon of interest, then it is not clear that one has a model [4.18]). Craver analyzes the example of the Hodgkin–Huxley mathematical model of the action potential in an axon (nerve fiber). Despite the fact that this model allowed Hodgkin and Huxley to derive many electrical features of neurons, and the fact that it was based on a number of fundamental laws of physics and chemistry, Craver argues that it was not in fact an explanatory model. He describes it instead as merely a phenomenological model because it failed to accurately describe the details of the underlying mechanism.

A similar mechanistic approach to model explanation has been developed by Kaplan [4.19], who introduces what he calls the mechanism–model–mapping (or 3M) constraint. He defines the 3M constraint as follows [4.19, p. 347]:

“A model of a target phenomenon explains that phenomenon to the extent that (a) the variables in the model correspond to identifiable components, activities, and organizational features of the target mechanism that produces, maintains, or underlies the phenomenon, and (b) the (perhaps mathematical) variables in the model correspond to causal relations among the components of the target mechanism.”

Kaplan takes this 3M constraint to provide a demarcation line between explanatory and nonexplanatory models. He further notes that [4.19, p. 347]

“3M aligns with the highly plausible assumption that the more accurate and detailed the model is for a target system or phenomenon the better it explains that phenomenon.”

Models that do not comply with 3M are rejected as nonexplanatory, being at best phenomenological models, useful for prediction, but giving no explanatory insight. In requiring that, explanatory models describe the *real* components and activities in the mechanism

that are *in fact* responsible for producing the phenomenon ([4.17, p. 361], [4.19, p. 353]). Craver and Kaplan rule out the possibility that fictional, metaphorical, or strongly idealized models can be explanatory.

One of the most comprehensive defenses of the explanatory power of models is given by Bokulich [4.18, 20–22], who argues that model explanations such as the three discussed previously (McMullin, Cartwright–Elgin–Sober, and Craver–Kaplan), can be seen as special cases of a more general account of the explanatory power of models. Bokulich’s approach draws on Woodward’s counterfactual account of explanation, in which [4.23, p. 11]

“the explanation must enable us to see what sort of difference it would have made for the explanandum if the factors cited in the explanans had been different in various possible ways.”

She argues that model explanations typically share the following three features: first, the explanans makes essential reference to a scientific model, which, as is the case with all models, will be an idealized, abstracted, or fictionalized representation of the target system. Second, the model explains the explanandum by showing how the elements of the model correctly capture the patterns of counterfactual dependence in the target system, enabling one to answer a wide range of what Woodward calls *what-if-things-had-been-different* questions. Finally, there must be what Bokulich calls a *justificatory step*, specifying the domain of applicability of the model and showing where and to what extent the model can be trusted as an adequate representation of the target for the purpose(s) in question [4.18, p. 39]; see also [4.22, p. 730]. She notes that this justificatory step can proceed bottom-up through something like a de-idealization analysis (as McMullin, Elgin, and Sober describe), top-down through an overarching theory (such as in the semiclassical mechanics examples Bokulich [4.20, 21] discusses), or through some combination.

Arguably one of the advantages of Bokulich’s approach is that it is not tied to one particular conception of scientific explanation, such as the DN or mechanistic accounts. By relaxing Woodward’s manipulationist construal of the counterfactual condition, Bokulich’s approach can even be extended to highly abstract, structural, or mathematical model explanations. She argues that the various *subspecies* of model explanation can be distinguished by noting what she calls the *origin* or ground of the counterfactual dependence. She explains, it could be either [4.18, p. 40]

“the elements represented in the model *causally producing* the explanandum (in the case of causal

model explanations), the elements of the model *being the mechanistic parts which make up the explanandum-system whole* (in the case of mechanistic model explanations), or the explanandum being a consequence of the laws cited in the model (in the case of covering law model explanations)."

She goes on to identify a fourth type of model explanation, which she calls structural model explanation, in which the counterfactual dependence is grounded in the typically mathematical structure of the theory, which limits the sorts of objects, properties, states, or behaviors that are admissible within the framework of that theory [4.18, p. 40]. Bokulich's approach can be thought of as one way to flesh out Morrison's suggestive, but unelaborated, remark that "the reason models are explanatory is that in representing these systems, they exhibit certain kinds of structural dependencies" [4.24, p. 63].

More recently, Rice [4.25] has drawn on Bokulich's account to develop a similar approach to the explanatory power of models that likewise uses Woodward's counterfactual approach without the manipulation condition. He writes [4.25, p. 20]:

"The requirement that these counterfactuals must enable one to, in principle, *intervene* in the system restricts Woodward's account to specifically causal explanations. However, I think it is a mistake to require that all scientific explanations must be causal. Indeed, if one looks at many of the explanations offered by scientific modelers, causes are not mentioned."

Compare this to Bokulich's statement [4.18, p. 39]:

"I think it is a mistake to construe all scientific explanation as a species of causal explanation, and more to the point here, it is certainly not the case that all model explanations should be understood as causal explanations. Thus while I shall adopt Woodward's account of explanation as the exhibiting of a pattern of counterfactual dependence, I will not construe this dependence narrowly in terms of the possible causal manipulations of the system"

Rice rightly notes that the question of causation is conceptually distinct from the question of what explains. He further requires on this approach that model explanations provide two kinds of counterfactual information, namely both what the phenomenon depends on and what sorts of changes are irrelevant to that phenomenon. Following Batterman [4.9, 15, 26], he notes that for explanations of phenomena that exhibit a kind of universality, an important part of the explanation is understanding that the particular causal details or processes are irrelevant – the same phenomenon would have been reproduced even if the causal details had been different in certain ways.

As an illustration, Rice discusses the case of optimality modeling in biology. He notes that optimality models are not only highly idealized, but also can be understood as a type of equilibrium explanation, where "most of the explanatory work in these models is done by *synchronic mathematical representations of structural features of the system*" [4.25, p. 8]. He connects this to the counterfactual account of model explanation as follows [4.25, p. 17]:

"Optimality models primarily focus on noncausal counterfactual relations between structural features and the system's equilibrium point. Moreover, these features can sometimes explain the target phenomenon without requiring any additional causal claims about the relationships represented in the model."

These causal details are irrelevant because the structural features cited in the model are multiply realizable; indeed, this is what allows optimality models to be used in explaining a wide variety of features across a diversity of biological systems.

In the approaches to model explanations discussed here, two controversial issues have arisen that merit closer scrutiny: first, whether the fictions or falsehoods in models can themselves do real explanatory work (i. e., even when they are neither harmless, deidealizable, nor eliminable), and second, whether many model explanations illustrate an important, but often overlooked, noncausal form of explanation. These issues will be taken up in turn in the next two sections.

## 4.2 Explanatory Fictions: Can Falsehoods Explain?

Models contain all sorts of falsehoods, from omissions, abstractions, and idealizations to outright fictions. One of the most controversial issues in model explanations is whether these falsehoods, which are inherent in the modeling practice, are compatible with the explanatory aims of science. *Reiss* in the context of explanatory models in economics has called this tension the *explanation paradox*: he writes [4.27, p. 43]:

“Three mutually inconsistent hypotheses concerning models and explanation are widely held: (1) economic models are false; (2) economic models are nevertheless explanatory; and (3) only true accounts explain. Commentators have typically resolved the paradox by rejecting either one of these hypotheses. I will argue that none of the proposed resolutions work and conclude that therefore the paradox is genuine and likely to stay.”

(This paradox, and some criticisms to *Reiss*’s approach (such as [4.28] are explored in a special issue of the *Journal of Economic Methodology* (volume 20, issue 3).)

The field has largely split into two camps on this issue: those who think it is only the true parts of models that do explanatory work and those who think the falsehoods play an essential role in the model explanation. Those in the former camp rely on things like de-idealization and harmless analyses to show that the falsehoods do not get in the way of the true parts of the model that do the real explanatory work. Those in the latter camp have the challenging task of showing that some idealizations are essential and some fictions yield true insights.

The *received view* is that the false parts of models only concern those things that are explanatorily irrelevant. Defenders of the received view include *Strevens*, who in his book detailing his kairetic account of scientific explanation (*Strevens* takes the term kairetic from the ancient Greek word *kairos*, meaning crucial moment [4.29, p. 477].), writes, “No causal account of explanation – certainly not the kairetic account – allows nonveridical models to explain” [4.29, p. 297]. He spells out more carefully how such a view is to be reconciled with the widespread use of idealized models to explain phenomena in nature, by drawing the following distinction [4.29, p. 318]:

“The content of an idealized model, then, can be divided into two parts. The first part contains the difference-makers for the explanatory target. [...] The second part is all idealization; its overt claims are false but its role is to point to parts of the actual

world that do not make a difference to the explanatory target.”

In other words, it is only the true parts of the model that do any explanatory work. The false parts are harmless, and hence should be able to be de-idealized away without affecting the explanation.

On the other side, a number of scholars have argued for the counterintuitive conclusion that sometimes it is in part *because* of their falsehoods – not despite them – that models explain. *Batterman* [4.9, 15, 26], for example, has argued that some idealizations are explanatorily ineliminable, that is, the idealizations or falsehoods themselves do real explanatory work. *Batterman* considers continuum model explanations of phenomena such as shocks (e.g., compressions traveling through a gas in a tube) and breaking drops (e.g., the shape of water as it drips from a faucet). In order to explain such phenomena, scientists make the idealization that the gas or fluid is a continuum (rather than describing it veridically as a collection of discrete gas or water molecules). These false continuum assumptions are essential for obtaining the desired explanation. In the breaking drop case, it turns out that different fluids of different viscosities dripping from faucets of different widths will all exhibit the same shape upon breakup. The explanation depends on a singularity that exists only in the (false) continuum model; such an explanation does not exist on the de-idealized molecular dynamics approach [4.15, pp. 442–443]). Hence, he concludes [4.15, p. 427],

“continuum idealizations are explanatorily ineliminable and [...] a full understanding of certain physical phenomena cannot be obtained through completely detailed, nonidealized representations.”

If such analyses are right, then they show that not all idealizations can be de-idealized, and, moreover, those falsehoods can play an essential role in the explanation.

*Bokulich* [4.10, 20–22] has similarly defended the view that it is not just the true parts of models that can do explanatory work, arguing that in some cases even fictions can be explanatory. She writes, “some fictions can give us genuine insight into the way the world is, and hence be genuinely explanatory and yield real understanding” [4.10, p. 94]. She argues that some fictions are able to do this by capturing in their fictional representation real patterns of structural dependencies in the world. As an example, she discusses semiclassical models whereby fictional electron orbits are used to explain peculiar features of quantum spectra. Although, according to quantum mechanics, electrons do not follow definite trajectories or orbits (i. e., such orbits are

fictions), physicists recognized that puzzling peaks in the recurrence spectrum of atoms in strong magnetic fields have a one-to-one correspondence with particular closed classical orbits [4.30, pp. 2789–2790] (quoted in [4.10, p. 99]):

“The resonances [...] form a series of strikingly simple and regular organization, not previously anticipated or predicted. [...] The regular type resonances can be physically rationalized and explained by classical periodic orbits of the electron on closed trajectories starting at and returning to the proton as origin.”

As she explains, at no point are these physicists challenging the status of quantum mechanics as the true, fundamental ontological theory; rather, they are deploying the fiction with the express recognition that it is indeed a literally false representation (interestingly this was one of the *Vaihinger*’s criteria for a *scientific* fiction [4.31, p. 98]). Nonetheless, it is a representation that is able to yield true physical insight and understanding by carefully capturing in its fictional representation the appropriate patterns of counterfactual dependence of the target phenomenon.

*Bokulich* [4.10, 20–22] offers several such examples of explanatory fictional models from semiclassical mechanics, where the received explanation of quantum phenomena appeals to classical structures, such as the Lyapunov (stability) exponents of classical trajectories, that have no clear quantum counterpart. Moreover, she notes that these semiclassical models with their fictional assumption of classical trajectories are valued not primarily as calculation tools (often they require calculations that are just as complicated), but rather are valued as models that provide an unparalleled level of physical insight into the structure of the quantum phenomena. Bokulich is careful to note that not just any fiction can do this kind of explanatory work; indeed, most fictions cannot. She shows more specifically how these semiclassical examples meet the three criteria of her account of model-based explanation, discussed earlier (e.g., [4.10, p. 106]).

A more pedestrian example of an explanatory fiction, and one that brings out some of the objections to such claims, is the case of light rays postulated by the ray (or geometrical) theory of optics. Strictly speaking, light rays are a fiction. The currently accepted fundamental theory of wave optics denies that they exist. Yet, light rays seem to play a central role in the scientific explanation of lots of phenomena, such as shadows and rainbows. The physicists *Kleppner* and *Delos*, for example, note [4.32, p. 610]:

“When one sees the sharp shadows of buildings in a city, it seems difficult to insist that light-rays are merely calculational tools that provide approximations to the full solution of the wave equation.”

Similarly, *Batterman*, argues [4.33, pp. 154–155]:

“One cannot explain various features of the rainbow (in particular, the universal patterns of intensities and fringe spacings) without ultimately having to appeal to the structural stability of ray theoretic structures called caustics – focal properties of families of rays.”

*Batterman* is quite explicit that he does not think that an explanatory appeal to these ray-theoretic structures requires reifying the rays; they are indeed fictions.

Some, such as *Belot*, want to dismiss ray-optics models as nothing but a mathematical device devoid of any physical content outside of the fundamental (wave) theory. He writes [4.34, p. 151]:

“The mathematics of the less fundamental theory is definable in terms of that of the more fundamental theory; so the requisite mathematical results can be proved by someone whose repertoire of interpreted physical theories included only the latter.”

The point is roughly this: it looks like in *Batterman*’s examples that one is making an explanatory appeal to fictional entities from a *less fundamental* theory that has been superseded (e.g., ray optics or classical mechanics). However, all one needs from that superseded theory is the mathematics – one does not need to give those bits of mathematics a physical interpretation in terms of the fictional entities or structures. Moreover, that mathematics appears to be definable in terms of the mathematics of the true *fundamental* theory. Hence, those fictional entities are not, in fact, playing an explanatory role.

*Batterman* has responded to these objections, arguing that in order to have an explanation, one does, in fact, need the fictional physical interpretation of that mathematics, and hence the explanatory resources of the nonfundamental theory. He explains [4.33, p. 159]:

“Without the physical interpretation to begin with, we would not know *what* boundary conditions to join to the differential equation. Neither, would we know *how* to join those boundary conditions to the equation. Put another way, we must examine the physical details of the *boundaries* (the shape, reflective and refractive details of the drops, etc.) in order to set up the *boundary conditions* required for the mathematical solution to the equation.”

In other words, without appealing to the fictional rays, we would not have the relevant information we need to appropriately set up and solve the mathematical model that is needed for the explanation.

In a paper with *Jansson*, Belot has raised similar objections against Bokulich's arguments that classical structures can play a role in explaining quantum phenomena. They write [4.35, p. 82]:

“Bokulich and others see explanations that draw on semiclassical considerations as involving elements of classical physics as well as of quantum physics. [...] But there is an alternative way of thinking of semiclassical mechanics: [...] starting with the formalism of quantum mechanics one proves theorems about approximate solutions – theorems that happen to involve some of the mathematical apparatus of classical mechanics. But this need not tempt us to think that there is [classical] physics in our explanations.”

Once again, we see the objection that it is just the bare mathematics, not the mathematics with its physical interpretation that is involved in the explanation. On Bokulich's view, however, it is precisely by connecting that *mathematical apparatus* to its physical interpretation in terms of classical mechanics, that one gains a deeper physical insight into the system one is studying. On her view, explanation is importantly about advancing understanding, and for this the physical interpretation is important. (*Potochnik* [4.5, Chap. 5] has also argued for a tight connection between explanation and understanding, responding to some of the traditional objections against this association. More broadly, she emphasizes the communicative function of explanation over the ontological approach to explanation, which makes more room for nonveridical model explanations than the traditional approach.) Even though classical mechanics is not the true fundamental theory, there are important respects in which it gets things right, and hence reasoning with fictional classical structures within the well-established confines of semiclassical mechanics, can yield explanatory insight and deepen our understanding.

As we have seen, these claims that fictions can explain (in special cases such as ray optics and classical structures) remain controversial and involve subtle issues. These debates are not entirely new, however, and they have some interesting historical antecedents, for example, in the works of Niels Bohr and James Clerk Maxwell. More specifically, when Bohr is articulating his widely misunderstood *correspondence principle*, (for an accessible discussion see [4.36]) he argues that one can explain why only certain quantum transitions between stationary states in atoms are allowed by ap-

pealing to which harmonic components appear in the Fourier decomposition of the electron's classical orbit (see [4.20, Sect. 4.2] and references therein). He does this even long after he has conceded to the new quantum theory that classical electron trajectories in the atom are impossible (i. e., they are a fiction). Although *Heisenberg* used this formulation of the correspondence principle to construct his matrix mechanics, he argued that “it must be emphasized that this correspondence is a purely formal result” [4.37, p. 83], and should not be thought of as involving any physical content from the other theory. Bohr, by contrast, was dissatisfied with this interpretation of the correspondence principle as pure mathematics, arguing instead that it revealed a deep *physical* connection between classical and quantum mechanics. Even earlier, we can see some of these issues arising in the work of *Maxwell*, who, in exploiting the utility of fictional models and physical analogies between disparate fields, argued ([4.38, p. 187]; for a discussion, see [4.39]):

“My aim has been to present the mathematical ideas to the mind in an embodied form [...] not as mere symbols, which convey neither the same ideas, nor readily adapt themselves to the phenomena to be explained.”

Three other challenges have been raised against the explanatory power of fictional models. First, there is a kind of slippery-slope worry that, once we admit some fictional models as explanatory, we will not have any grounds on which to dismiss other fictional models as nonexplanatory. *Bokulich* [4.22] introduces a framework for addressing this problem. Second, *Schindler* [4.40] has raised what he sees as a tension in Bokulich's account. He claims that on one hand she says semiclassical explanations of quantum phenomena are autonomous in the sense that they provide more insight than the quantum mechanical ones. Yet, on the other hand, she notes that semiclassical models are justified through semiclassical theory, which connects these representations as a kind of approximation to the full quantum mechanics. Hence, they cannot be autonomous. This objection seems to trade on an equivocation of the term *autonomous*: in the first case, *autonomous* is used to mean “a representation of the phenomenon that yields more physical insight” and in the second case *autonomous* is used to mean “cannot be mathematical justified through various approximation methods”. These seem to be two entirely different concepts, and, hence, not really in tension with each other. Moreover, Bokulich never uses the term *autonomous* to describe either, so this seems to be a misleading reading of her view.



Schindler also rehearses the objection, raised by *Belot* and *Jansson* [4.35], that by eliminating the interventionist condition in Woodward's counterfactual approach to explanation she loses what he calls "the asymmetry-individuating function", by which he means her account seems susceptible to the traditional problem of asymmetry that plagued the DN account of explanation (e.g., that falling barometers could be used to explain impending storms or shadows could be used to explain the height of flag poles, to recall Sylvain Bromberger's well-known examples). This problem was taken to be solved by the causal approach to explanation, whereby one secures the explanatory asymmetry simply by appealing to the asymmetry of causation. It is important to note, however, that this is not an objection specifically to Bokulich's account of structural model explanation, but rather is a challenge for any noncausal account of explanation (*Bokulich* outlines a solution to the problem of asymmetry for her account in [4.22]). Since many examples of explanatory models purport to be noncausal explanations, we will examine this topic more fully in the next section.

Another context in which this issue about the explanatory power of fictional models arises is in connection with cognitive models in psychology and cognitive neuroscience. *Weiskopf*, for example, discusses how psychological capacities are often understood in terms of cognitive models that functionally abstract from the underlying real system. More specifically, he notes [4.41, p. 328]:

"In attempting to understand the high level dynamics of complex systems like brains, modelers have recourse to many techniques for constructing such indirect accounts [...] *reification, functional abstraction, and fictionalization.*"

By reification, he means "positing something with the characteristics of a more or less stable and enduring object, where in fact no such thing exists" [4.41, p. 328]. He gives as an example the positing of symbolic representations in classical computational systems, even though he notes that nothing in the brain seems to *stand still* or be manipulable in the way symbols do. Functional abstraction, he argues occurs when we [4.41, p. 329]

"decompose a modeled system into subsystems and other components on the basis of what they do, rather than their correspondence with organizations and groupings in the target system."

He notes that this occurs when there are cross-cutting functional groupings that do not map onto the structural or anatomical divisions of the brain. He notes that this strategy emphasizes *networks, not locations* in

relating cognition to neural structures. Finally, there is also fictionalization, which, as he describes [4.41, p. 331],

"involves putting components into a model that are known not to correspond to any element of the modeled system, but which serve an essential role in getting the models to operate correctly."

He gives as an example of a fiction in cognitive modeling what are called *fast enabling links* (FELs), which are independent of the channels by which cells actually communicate and are assumed to have functionally infinite propagation speeds, allowing two cells to fire in synchrony [4.41, p. 331]. Despite being false in these ways, some modelers take these fictions to be essential to the operation of the model and not likely to be eliminated in future versions.

*Weiskopf* concludes that models involving reifications, functional abstractions, and fictions, can nonetheless in some cases succeed in "meeting the general normative constraints on explanatory models perfectly well" [4.41, p. 332], and hence such models can be counted as genuinely explanatory. Although *Weiskopf* recognizes the many great successes of mechanistic explanations in biological and neural systems, he wants to resist an *imperialism* that attempts to reduce all cases of model explanations in these fields to mechanistic model explanations.

More recently, *Buckner* [4.42] has criticized *Weiskopf's* arguments that functionalist models involving fictions, abstractions, and reification can be explanatory and defended the mechanist's maxim (e.g., as articulated by Craver and Kaplan) that only mechanistic models can genuinely explain. *Buckner* employs two strategies in arguing against *Weiskopf*: first, in cases where the models do explain, he argues that they are really just mechanism sketches, and where they cannot be reconstructed mechanistically, he dismisses them as impoverished explanations. He writes [4.42, p. 3]:

"Concerning fictionalization and reification, I concede that models featuring such components cannot be interpreted as mechanism sketches, but argue that interpreting their nonlocalizable components as natural kinds comes with clear costs in terms of those models' counterfactual power. [...] Functional abstraction, on the other hand, can be considered a legitimate source of kinds, but only on the condition that the functionally abstract models be interpreted as sketches that could be elaborated into a more complete mechanistic model."

An essential feature of mechanistic models seems to be that their components are localizable. *Weiskopf* argues, however, that his functional kinds are multi-

ply realizable, that is, they apply to many different kinds of underlying mechanisms, and that in some cases, they are distributed in the sense that they ascribe to a given model component capacities that are distributed amongst distinct parts of the physical system. Hence, without localization, such models cannot be reconstructed as mechanistic models.

What of Buckner's claim that fictional models will be impoverished with regard to their counterfactual power? Consider again Weiskopf's example of the fictional FELs, which are posited in the model to allow the cells to achieve synchrony. Buckner argues explanations involving models with FELs are impoverished in that if one had a true account of synchrony, that model explanation would support *more* counterfactual knowledge. It is not clear, however, that this objection undermines the explanatory power of models involving FELs per se; rather it seems only to suggest that if we knew more and had the true account of syn-

chrony we might have a *deeper* explanation (at least on the assumption that this true account of synchrony would allow us to answer a wider range of what-if-things-had-been-different questions) (For an account of explanatory depth, see [4.43]). However, the explanation involving the fiction might still be perfectly adequate for the purpose for which it is being deployed, and hence it need not even be counted as impoverished. For example, there might be some explananda (ones other than the explanandum of *how do cells achieve synchrony*) for which it simply does not matter *how* cells achieve synchrony; the fact that they *do* achieve synchrony might be all that is required for some purposes.

Weiskopf is not alone in trying to make room for nonmechanistic model explanations; Irvine [4.44] and Ross [4.45] have also recently defended nonmechanistic model explanations in cognitive science and biology. Their approaches argue for noncausal forms of model explanation, which we will turn to next.

### 4.3 Explanatory Models and Noncausal Explanations

Recently, there has been a growing interest in noncausal forms of explanation. Similar to Bokulich's [4.20, 21] approach, many of these seek to understand noncausal explanations within the context of Woodward's [4.23] counterfactual approach to explanation without the interventionist criterion that restricts his account specifically to causal explanation [4.25, 46]. Noncausal explanations are usually defined negatively as explaining by some means *other than* citing causes, though this is presumably a heterogeneous group. We have already seen one type of noncausal model-based explanation: [4.20, 21] structural model explanations in physics. More recently, examples have been given in fields ranging from biology to cognitive science. Highly mathematical model explanations are another type of noncausal explanation, though not all mathematical models are noncausal. A few recent examples are considered here.

In the context of biology and cognitive science, Irvine [4.44] has argued for the need to go beyond the causal-mechanical account of model explanation and defends what she calls a noncausal structural form of model explanation. She focuses specifically on reinforcement learning (RL) models in cognitive science and optimality models in biology. She notes that although RL and optimality models can be construed as providing causal explanations in some contexts, there are other contexts in which causal explanations miss the mark. She writes [4.44, p. 11]:

“In the account developed here, it is not the presence of idealisation or abstraction in models that

is important, nor the lack of description of causal dynamics or use of robustness analyses to test the models. Instead, it is the bare fact that some models and target systems have equilibrium points [that] are highly O-robust with respect to initial conditions and perturbations. [...] This alone can drive a claim about noncausal structural explanations.”

By O-robustness, Irvine means a robust convergence to an optimal state across a range of interventions, whether it be an optimization of fitness or an optimization of decision-making strategies. Her argument is that since interventions (in the sense of Woodward) do not make a difference to the convergence on the optimal state, that convergence cannot be explained causally, and is instead due to structural features of the model and target system it explains.

Another recent approach to noncausal model explanation is Batterman and Rice's [4.47] minimal model explanations. Minimal models are models that explain patterns of macroscopic behavior for systems that are heterogeneous at smaller scales. Batterman and Rice discuss two examples of minimal models in depth: the Lattice Gas Automaton model, which is used to explain large-scale patterns in fluid flow, and Fisher's Sex Ratio model, which is used to explain why one typically finds a 1 : 1 ratio of males to females, across diverse populations of species. In both cases, they argue [4.47, p. 373]:

“these minimal models are explanatory because there is a detailed story about why the myriad details that distinguish a class of systems are irrelevant

to their large-scale behavior. This story demonstrates, rather than assumes, a kind of stability or robustness of the large-scale behavior we want to explain under drastic changes in the various details of the system.”

They make two further claims about these minimal model explanations. First, they argue that these explanations are “distinct from various causal, mechanical, difference making, and so on, strategies prominent in the philosophical literature” [4.47, p. 349]. Second, they argue that the explanatory power of minimal models cannot be accounted for by any kind of mirroring or mapping between the model and target system (what they call the *common features* account). Instead, these noncausal explanations work by showing that the minimal model and diverse real-world systems fall into the same universality class. This latter claim has been challenged by *Lange* [4.48] who, though sympathetic to their claim that minimal models are a noncausal form of model explanation, argues that their explanatory power does in fact derive from the model sharing features in common with the diverse systems it describes (i. e., the *common features* account Batterman and Rice reject).

*Ross* [4.45] has applied the minimal models account to dynamical model explanations in the neurosciences. More specifically, she considers as an explanandum phenomenon the fact that a diverse set of neural systems (e.g., rat hippocampal neurons, crustacean motor neurons, and human cortical neurons *Ross* [4.45, p. 48]), which are quite different at the molecular level, nonetheless all exhibit the same *type I* excitability behavior. She shows that the explanation for this involves applying mathematical abstraction techniques to the various detailed models of each particular type of neural system and then showing that all these diverse systems converge on one and the same canonical model (known as the Ermentrout–Kopell model). After defending the explanatory power of these canonical models, *Ross* then contrasts this kind of noncausal model explanation with the causal–mechanical model approach [4.45, p. 46]:

“The canonical model approach contrasts with Kaplan and Craver’s claims because it is used to explain the shared behavior of neural systems without revealing their underlying causal–mechanical structure. As the neural systems that share this behavior consist of differing causal mechanisms [...] a mechanistic model that represented the causal structure of any single neural system would no longer represent the entire class of systems.”

Her point is that a noncausal explanation is called for in this case because the particular causal details are

irrelevant to the explanation of the universal behavior of class I neurons. The minimal models approach, as we saw above, is designed precisely to capture these sort explanations involving universality.

More generally, many highly abstract or highly mathematical model explanations also seem to fall into this general category of noncausal model explanations. *Pincock*, for example, identifies a type of explanation that he calls *abstract explanation*, which could be extended to model-based explanations. He writes “the best recent work on causal explanation is not able to naturally accommodate these abstract explanations” [4.49, p. 11]. Although some of the explanations *Pincock* cites, such as the topological (graph theory) explanation for why one cannot cross the seven bridges of Königsberg exactly once in a nonbacktracking circuit, seem to be genuinely noncausal explanations, it is not clear that all *abstract* explanations are necessarily noncausal. *Reutlinger* and *Andersen* [4.50] have recently raised this objection against *Pincock*’s account, arguing that an explanation’s being abstract is not a sufficient condition for it being noncausal. They argue that many causal explanations can be abstract too and so more work needs to be done identifying what makes an explanation truly noncausal. This is a particularly pressing issue in model-based explanations, since many scientific models are abstract in this sense of leaving out microphysical or concrete causal details about the explanandum phenomenon.

*Lange* [4.51] has also identified a kind of noncausal explanation that he calls a *distinctively mathematical* explanation. *Lange* considers a number of candidate mathematical explanations, such as why one cannot divide 23 strawberries evenly among three children, why cicadas have life-cycle periods that are prime, and why honeybees build their combs on a hexagonal grid. *Lange* notes that whether these are to count as distinctively mathematical explanations depends on precisely how one construes the explanandum phenomenon. If we ask why honeybees divide the honeycomb into hexagons, rather than other polygons, and we cite that it is selectively advantageous for them to minimize the wax used, together with the mathematical fact that a hexagonal grid has the least total perimeter, then it is an ordinary causal explanation (it works by citing selection pressures). If, however [4.50, p. 500]:

“we narrow the explanandum to the fact that in any scheme to divide their combs into regions of equal area, honeybees would use at least the amount of wax they would use in dividing their combs into hexagons. [...] this fact has a distinctively mathematical explanation.”

As *Lange* explains more generally [4.51, p. 485]:

“These explanations are noncausal, but this does not mean that they fail to cite the explanandum’s causes, that they abstract away from detailed causal histories, or that they cite no natural laws. Rather, in these explanations, the facts doing the explaining are modally stronger than ordinary causal laws.”

The key issue is not whether the explanans cite the explanandum’s causes, but whether the explana-

tion works *by virtue of* citing those causes. Distinctively mathematical (noncausal) explanations show the explanandum to be necessary to a stronger degree than would result from the causal powers alone.

As this literature makes clear, distinguishing causal from noncausal explanations is a subtle and open problem, but one crucial for understanding the wide-spread use of abstract mathematical models in many scientific explanations.

## 4.4 How-Possibly versus How-Actually Model Explanations

Models and computer simulations can often generate patterns or behaviors that are strikingly similar to the phenomenon to be explained. As we have seen, however, that is typically not enough to conclude that the model thereby explains the phenomenon. An important distinction here is that between a *how-possibly* model explanation and a *how-actually* model explanation.

The notion of a how-possibly explanation was first introduced in the 1950s by Dray in the context of explanations in history. Dray conceived of how-possibly explanations as a rival to the DN approach, which he labeled *why-necessarily* explanations [4.52, p. 161]. Dray interpreted how-possibly explanations as ones that merely aim to show why a particular phenomenon or event “need not have caused surprise” [4.52, p. 157]; hence, they are answers to a different kind of question and can be considered complete explanations in themselves. Although Dray’s approach was influential, subsequent authors have interpreted this distinction in different ways. Brandon, in the context of explanations in evolutionary biology, for example, writes [4.53, p. 184]:

“A how-possibly explanation is one where one or more of the explanatory conditions are speculatively postulated. But if we gather more and more evidence for the postulated conditions, we can move the how-possibly explanation along the continuum until finally we count it as a how-actually explanation.”

On this view, the distinction is a matter of the degree of confirmation, not a difference of kind: as we get more evidence that the processes cited in the model are the processes operating in nature, we move from a how-possibly to how-actually explanation.

*Forber* [4.54], however, rejects this interpretation of the distinction as marking a degree of empirical support, and instead defends Dray’s original contention that they mark different kinds of explanations. More specifically, *Forber* distinguishes two kinds of how-possibly

explanations that he labels *global how-possibly* and *local how possibly* explanations [4.54, p. 35]:

“The global how-possibly explanations have theory, mathematics, simulations, and analytical techniques as the resources for fashioning such explanations. [...] The local how-possibly explanations draw upon the models of evolutionary processes and go one step further. They speculate about the biological possibilities relative to an information set enriched by the specific biology of a target system. [...] How-actually explanations, carefully confirmed by empirical tests, aim to identify the correct evolutionary processes that did, in fact, produce the target outcome.”

Although Forber’s distinction is conceptually helpful, it is not clear whether global versus local how-possibly explanations should, in fact, be seen as two distinct categories, rather than simply two poles of a spectrum.

Craver draws a distinction between how-possibly models and how-actually models that is supposed to track the corresponding two kinds of explanations. He notes that how-possibly models purport to explain (unlike phenomenological models, which do not purport to explain), but they are only loosely constrained conjectures about the mechanism. How-actually models, by contrast, describe the detailed components and activities that, in fact, produce the phenomenon. He writes [4.17, p. 361]:

“How-possibly models are [...] not adequate explanations. In saying this I am saying not merely that the description must be true (or true enough) but further, that the model must correctly characterize the details of the mechanism.”

Craver seems to see the distinction resting not just on the degree of confirmation (truth) but also on the degree of detail.

*Bokulich* [4.55] defends another construal of the how-possibly/how-actually distinction and applies it to model-based explanations more specifically. She considers, as an example, model-based explanations of a puzzling ecological phenomenon known as tiger bush. Tiger bush is a striking periodic banding of vegetation in semi-arid regions, such as southwest Niger. A surprising feature of tiger bush is that it can occur for a wide variety of plants and soils, and it is not induced by any local heterogeneities or variations in topography. By tracing how scientists use various idealized models (e.g., Turing models or differential flow models) to explain phenomena such as this, Bokulich argues a new insight into the how-possibly/how-actually distinction can be gained.

The first lesson she draws is that there are different levels of abstraction at which the explanandum phenomenon can be framed, which correspond to different explanatory contexts [4.55, p. 33]. These different explanatory contexts can be clarified by considering the relevant contrast class of explanations (for a discussion of contrast classes and their importance in scientific explanation, see [4.56, Chap. 5]). Second, she argues *pace* Craver that the how-possibly/how-actually distinction

does not track how detailed the explanation is. She explains [4.55, p. 334]:

“It is not the amount of detail that is relevant, but rather whether the mechanism represented in the model is the mechanism operating in nature. Indeed as we saw in the tiger bush case, the more abstractly the explanatory mechanism is specified, the easier it is to establish it as a how-actually explanation; whereas the more finely the explanatory mechanism is specified, the less confident scientists typically are that their particular detailed characterization of the mechanism is the actual one.”

Hence, somewhat counterintuitively, model explanations at a more fine-grained level are more likely to be how-possibly model explanations, even when they are nested within a higher level how-actually model explanation of a more abstract characterization of the phenomenon. She concludes that when assessing model explanations, it is important to pay attention to what might be called the scale of resolution at which the explanandum phenomenon is being framed in a particular explanatory context.

## 4.5 Tradeoffs in Modeling: Explanation versus Other Functions for Models

Different scientists will often create different models of a given phenomenon, depending on their particular interests and aims. Following *Giere*, we might note that “there is no best scientific model of anything; there are only models more or less good for different purposes” [4.57, p. 1060]. If this is right, then it raises the following questions: What are the features that make a model particularly good for the purpose of explanation? Are there tradeoffs between different modeling aims, such that if one optimizes a model for explanation, for example, then that model will fail to be optimized for some other purpose, such as prediction?

One of the earliest papers to explore this theme of tradeoffs in modeling is Levins’ paper *The Strategy of Model Building in Population Biology*. Levins writes [4.58, p. 422]:

“It is of course desirable to work with manageable models which maximize generality, realism, and precision toward the overlapping but not identical goals of understanding, predicting, and modifying nature. But this cannot be done.”

Levins then goes on to describe various modeling strategies that have evolved among modelers, such as sacrificing realism to generality and precision, or sac-

rificing precision to realism and generality. Levins in his own work on models in ecology favored this latter strategy, where he notes his concern was primarily qualitative not quantitative results, and he emphasizes the importance of robustness analyses in assessing these models.

Although Levins’s arguments have not gone unchallenged, Matthewson and Weisberg have recently defended the view that some tradeoffs in modeling are genuine. They focus on precision and generality, given the relevance of this tradeoff to the aim of explanatory power. After a technical demonstration of different kinds of tradeoffs between two different notions of generality and precision, they conclude [4.59, p. 189]:

“These accounts all suggest that increases in generality are, *ceteris paribus*, associated with an increase in explanatory power. The existence of tradeoffs between precision and generality indicates that one way to increase an explanatorily valuable desideratum is by sacrificing precision. Conversely, increasing precision may lead to a decrease in explanatory power via its effect on generality.”

Mapping out various tensions and tradeoffs modelers may face in developing models for vari-

ous aims, such as scientific explanation, remains a methodologically important, though underexplored topic.

More recently, *Bokulich* [4.60] has explored such tradeoffs in the context of modeling in geomorphology, which is the study of how landscapes and coastlines change over time. Even when it comes to a single phenomenon, such as braided rivers (i. e., rivers in which there is a number of interwoven channels and bars that dynamically shift over time), one finds that scientists use different kinds of models depending on whether their primary aim is explanation or prediction. When they are interested explaining why rivers braid geomorphologists tend to use what are known as *reduced complexity models*, which are typically very simple cellular automata models with a highly idealized representation of the fluvial dynamics [4.61]. The goal is to try to abstract away and isolate the key mechanisms responsible for the production of the braided pattern. This approach is contrasted with an alternative approach to modeling in geomorphology known as *reductionist modeling*. Here one tries to simulate the braided river in as much accurate detail and with as many different processes included as is computationally feasible, and then tries to solve the relevant Navier–Stokes equations in three dimensions. These reductionist models are the best available tools for predicting the features of braided rivers [4.61, p. 159], but they are so complex that they yield very little insight into *why* the patterns emerge as they do.

## 4.6 Conclusion

There is a growing realization that the use of idealized models to explain phenomena is pervasive across the sciences. The appreciation of this fact has led philosophers of science to begin to introduce model-based accounts of explanation in order to bring the philosophical literature on scientific explanation into closer agreement with actual scientific practice.

A key question here has been whether the idealizations and falsehoods inherent in modeling are *harmless* in the sense of doing no real explanatory work, or whether they have an essential – maybe even ineliminable – role to play in some scientific explanations. Are such fictions compatible with the explanatory aims of science, and if so, under what circumstances? While some inroads have been made on this question, it remains an ongoing area of research. As we saw, yet another controversial issue concerns the fact that many highly abstract and mathematical models seem to exemplify a noncausal form of explanation, contrary to the current orthodoxy in scientific explanation. Deter-

*Bokulich* uses cases such as these to argue for what she calls a division of cognitive labor among models [4.60, p. 121]:

“If one’s goal is explanation, then reduced complexity models will be more likely to yield explanatory insight than simulation models; whereas if one’s goal is quantitative predictions for concrete systems, then simulation models are more likely to be successful. I shall refer to this as the *division of cognitive labor among models*.”

As Bokulich notes, however, one consequence of this division of cognitive labor is that a model that was designed to optimize explanatory insight might fail to make quantitatively accurate predictions (a different cognitive goal). She continues [4.60, p. 121]:

“This failure in predictive accuracy need not mean that the basic mechanism hypothesized in the explanatory model is incorrect. Nonetheless, explanatory models need to be tested to determine whether the explanatory mechanism represented in the model is in fact the real mechanism operating in nature.”

She argues for the importance of robustness analyses in assessing these explanatory models, noting that while robustness analyses cannot themselves function as a nonempirical mode of confirmation, they can be used to identify those *qualitative* predictions or trends in the model that can appropriately be compared with observations.

mining what is or is not to count as a causal explanation turns out to be a subtle issue.

Finally, just because a model or computer simulation can reproduce a pattern or behavior that is strikingly like the phenomenon to be explained, does not mean that it thereby explains that phenomenon. An important distinction here is that between a how-possibly model explanation and a how-actually model explanation. Despite the wide agreement that such a distinction is important, there has been less agreement concerning how precisely these lines should be drawn.

Although significant progress has been made in recent years in understanding the role of models in scientific explanation, there remains much work to be done in further clarifying many of these issues. However, as the articles reviewed here reveal, exploring just how and when models can explain is a rich and fruitful area of philosophical investigation and one essential for understanding the nature of scientific practice.

## References

- 4.1 C. Hempel: *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science* (Free Press, New York 1965)
- 4.2 W. Salmon: *Scientific Explanation and the Causal Structure of the World* (Princeton Univ. Press, Princeton 1984)
- 4.3 R. Frigg, S. Hartmann: Models in science. In: *The Stanford Encyclopedia of Philosophy*, ed. by E.N. Zalta (Stanford Univ., Stanford 2012)
- 4.4 J. Maynard Smith: *Evolution and the Theory of Games* (Cambridge Univ. Press, Cambridge 1982)
- 4.5 A. Potochnik: *Idealization and the Aims of Science* (Univ. Chicago Press, forthcoming)
- 4.6 E. McMullin: Structural explanation, *Am. Philos. Q.* **15**(2), 139–147 (1978)
- 4.7 E. McMullin: Galilean idealization, *Stud. Hist. Philos. Sci.* **16**(3), 247–273 (1985)
- 4.8 E. McMullin: A case for scientific realism. In: *Scientific Realism*, ed. by J. Leplin (Univ. California Press, Berkeley 1984)
- 4.9 R. Batterman: Critical phenomena and breaking drops: Infinite idealizations in physics, *Stud. Hist. Philos. Modern Phys.* **36**, 25–244 (2005)
- 4.10 A. Bokulich: Explanatory Fictions. In: *Fictions in Science: Philosophical Essays on Modeling and Idealization*, ed. by M. Suárez (Routledge, London 2009) pp. 91–109
- 4.11 N. Cartwright: *How the Laws of Physics Lie* (Clarendon Press, Oxford 1983)
- 4.12 P. Duhem: *The Aim and Structure of Physical Theory* (Princeton Univ. Press, Princeton 1914/1954)
- 4.13 M. Elgin, E. Sober: Cartwright on explanation and idealization, *Erkenntnis* **57**, 441–450 (2002)
- 4.14 D. Cristol, P. Switzer: Avian prey-dropping behavior. II. American crows and walnuts, *Behav. Ecol.* **10**, 220–226 (1999)
- 4.15 R. Batterman: Idealization and modeling, *Synthese* **169**, 427–446 (2009)
- 4.16 A. Kennedy: A non representationalist view of model explanation, *Stud. Hist. Philos. Sci.* **43**(2), 326–332 (2012)
- 4.17 C. Craver: When mechanistic models explain, *Synthese* **153**, 355–376 (2006)
- 4.18 A. Bokulich: How scientific models can explain, *Synthese* **180**, 33–45 (2011)
- 4.19 D.M. Kaplan: Explanation and description in computational neuroscience, *Synthese* **183**, 339–373 (2011)
- 4.20 A. Bokulich: *Reexamining the Quantum-Classical Relation: Beyond Reductionism and Pluralism* (Cambridge Univ. Press, Cambridge 2008)
- 4.21 A. Bokulich: Can classical structures explain quantum phenomena?, *Br. J. Philos. Sci.* **59**(2), 217–235 (2008)
- 4.22 A. Bokulich: Distinguishing explanatory from non-explanatory fictions, *Philos. Sci.* **79**(5), 725–737 (2012)
- 4.23 J. Woodward: *Making Things Happen: A Theory of Causal Explanation* (Oxford University Press, Oxford 2003)
- 4.24 M. Morrison: Models as autonomous agents. In: *Models and Mediators: Perspectives on Natural and Social Science*, ed. by M. Morgan, M. Morrison (Cambridge Univ. Press, Cambridge 1999) pp. 38–65
- 4.25 C. Rice: Moving beyond causes: Optimality models and scientific explanation, *Noûs* **49**(3), 589–615 (2015)
- 4.26 R. Batterman: *Devil in the Details: Asymptotic Reasoning in Explanation, Reduction, and Emergence* (Oxford University Press, Oxford 2002)
- 4.27 J. Reiss: The explanation paradox, *J. Econ. Methodol.* **19**(1), 43–62 (2012)
- 4.28 U. Mäki: On a paradox of truth, or how not to obscure the issue of whether explanatory models can be true, *J. Econ. Methodol.* **20**(3), 268–279 (2013)
- 4.29 M. Strevens: *Depth: An Account of Scientific Explanation* (Harvard Univ. Press, Cambridge 2008)
- 4.30 J. Main, G. Weibusch, A. Holle, K.H. Welge: New quasi-Landau structure of highly excited atoms: The hydrogen atom, *Phys. Rev. Lett.* **57**, 2789–2792 (1986)
- 4.31 H. Vaihinger: *The Philosophy of 'As If': A System of the Theoretical, Practical, and Religious Fictions of Mankind*, 2nd edn. (Lund Humphries, London [1911] 1952), translated by C.K. Ogden
- 4.32 D. Kleppner, J.B. Delos: Beyond quantum mechanics: Insights from the work of Martin Gutzwiller, *Found. Phys.* **31**, 593–612 (2001)
- 4.33 R. Batterman: Response to Belot's "Whose Devil? Which Details?", *Philos. Sci.* **72**, 154–163 (2005)
- 4.34 G. Belot: Whose Devil? Which Details?, *Philos. Sci.* **52**, 128–153 (2005)
- 4.35 G. Belot, L. Jansson: Review of reexamining the quantum-classical relation, *Stud. Hist. Philos. Modern Phys.* **41**, 81–83 (2010)
- 4.36 A. Bokulich: Bohr's correspondence principle. In: *The Stanford Encyclopedia of Philosophy*, ed. by E.N. Zalta (Stanford Univ., Stanford 2014), <http://plato.stanford.edu/archives/spr2014/entries/bohr-correspondence/>
- 4.37 W. Heisenberg: In: *The Physical Principles of the Quantum Theory*, ed. by C. Eckart, F. Hoyt (Univ. Chicago Press, Chicago 1930)
- 4.38 J.C. Maxwell: On Faraday's Lines of Force. In: *The Scientific Papers of James Clerk Maxwell*, ed. by W. Niven (Dover Press, New York [1855/56] 1890) pp. 155–229
- 4.39 A. Bokulich: Maxwell, Helmholtz, and the unreasonable effectiveness of the method of physical analogy, *Stud. Hist. Philos. Sci.* **50**, 28–37 (2015)
- 4.40 S. Schindler: Explanatory fictions – For real?, *Synthese* **191**, 1741–1755 (2014)
- 4.41 D. Weiskopf: Models and mechanism in psychological explanation, *Synthese* **183**, 313–338 (2011)
- 4.42 C. Buckner: Functional kinds: A skeptical look, *Synthese* **192**, 3915–3942 (2015)
- 4.43 C. Hitchcock, J. Woodward: Explanatory generalizations: Part II. Plumbing explanatory depth, *Noûs* **37**(2), 181–199 (2003)

- 4.44 E. Irvine: Models, robustness, and non-causal explanation: A foray into cognitive science and biology, *Synthese* **192**, 3943–3959 (2015), doi:[10.1007/s11229-014-0524-0](https://doi.org/10.1007/s11229-014-0524-0)
- 4.45 L. Ross: Dynamical models and explanation in neuroscience, *Philos. Sci.* **82**(1), 32–54 (2015)
- 4.46 J. Saatsi, M. Pexton: Reassessing Woodward's account of explanation: Regularities, counterfactuals, and noncausal explanations, *Philos. Sci.* **80**(5), 613–624 (2013)
- 4.47 R. Batterman, C. Rice: Minimal model explanations, *Philos. Sci.* **81**(3), 349–376 (2014)
- 4.48 M. Lange: On 'Minimal model explanations': A reply to Batterman and Rice, *Philos. Sci.* **82**(2), 292–305 (2015)
- 4.49 C. Pincock: Abstract explanations in science, *Br. J. Philos. Sci.* **66**(4), 857–882 (2015), doi:[10.1093/bjps/axu016](https://doi.org/10.1093/bjps/axu016)
- 4.50 A. Reutlinger, H. Andersen: Are explanations non-causal by virtue of being abstract?, unpublished manuscript
- 4.51 M. Lange: What makes a scientific explanation distinctively mathematical?, *Br. J. Philos. Sci.* **64**, 485–511 (2013)
- 4.52 W. Dray: *Law and Explanation in History* (Oxford Univ. Press, Oxford 1957)
- 4.53 R. Brandon: *Adaptation and Environment* (Princeton Univ. Press, Princeton 1990)
- 4.54 P. Forber: Confirmation and explaining how possible, *Stud. Hist. Philos. Biol. Biomed. Sci.* **41**, 32–40 (2010)
- 4.55 A. Bokulich: How the tiger bush got its stripes: 'How possibly' vs. 'How actually' model explanations, *The Monist* **97**(3), 321–338 (2014)
- 4.56 B. van Fraassen: *The Scientific Image* (Oxford University Press, Oxford 1980)
- 4.57 R. Giere: The nature and function of models, *Behav. Brain Sci.* **24**(6), 1060 (2001)
- 4.58 R. Levins: The Strategy of model building in population biology, *Am. Sci.* **54**(4), 421–431 (1966)
- 4.59 J. Matthewson, M. Weisberg: The structure of trade-offs in model building, *Synthese* **170**(1), 169–190 (2008)
- 4.60 A. Bokulich: Explanatory models versus predictive models: Reduced complexity modeling in geomorphology. In: *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, ed. by V. Karakostas, D. Dieks (Springer, Cham, Heidelberg, New York, Dordrecht, London 2013)
- 4.61 A.B. Murray: Contrasting the goals, strategies, and predictions associated with simplified numerical models and detailed simulations. In: *Prediction in Geomorphology*, ed. by P. Wilcock, R. Iverson (American Geophysical Union, Washington 2003) pp. 151–165



# Models and Simulations

## 5. Models and Simulations

Nancy J. Nersessian, Miles MacLeod

In this chapter we present some of the central philosophical issues emerging from the increasingly expansive and sophisticated roles computational modeling is playing in the natural and social sciences. Many of these issues concern the adequacy of more traditional philosophical descriptions of scientific practice and accounts of justification for handling computational science, particularly the role of theory in the generation and justification of physical models. However, certain novel issues are also becoming increasingly prominent as a result of the spread of computational approaches, such as nontheory-driven simulations, computational methods of inference, and the important, but often ignored, role of cognitive processes in computational model building.

Most of the philosophical literature on *models and simulations* focuses on computational simulation, and this is the focus of our review. However, we wish to note that the chief distinguishing characteristic between a model and a simulation (model) is that the latter is dynamic. They can be *run* either as constructed or under a range of experimental conditions. Thus, the broad class of simulation models should be understood as com-

5.1	<b>Theory-Based Simulation</b> .....	119
5.2	<b>Simulation not Driven by Theory</b> .....	121
5.3	<b>What is Philosophically Novel About Simulation?</b> .....	124
5.4	<b>Computational Simulation and Human Cognition</b> .....	127
	<b>References</b> .....	130

prising dynamic physical models and mental models, topics considered elsewhere in this volume.

This chapter is organized as follows. First in Sect. 5.1 we discuss simulation in the context of well-developed theory (usually physics-based simulations). Then in Sect. 5.2 we discuss simulation in contexts where there are no over-arching theories of the phenomena, notably agent-based simulations and those in systems biology. We then turn to issues of whether and how simulation modeling introduces novel concerns for the philosophy of science in Sect. 5.3. Finally, we conclude in Sect. 5.4 by addressing the question of the relation between human cognition and computational simulation, including the relationship between the latter and thought experimenting.

### 5.1 Theory-Based Simulation

A salient aspect of computational simulation, and the one which has attracted the most substantial philosophical interest so far, is its ability to extend the power and reach of theories in modern science beyond what could be achieved by pencil and paper alone. Work on simulations has concentrated on simulations built from established background theories or theoretical models and the relations between these simulations and theory. Examples have been sourced mainly from the physical sciences, including simulations in astrophysics, fluid dynamics, nanophysics, climate science and meteorology. *Winsberg* has been foremost in study-

ing theory-driven forms of simulation and promoting the importance of philosophical investigation of it by arguing that such simulations set a new agenda for philosophy of science [5.1–5]. He uses the case of simulation to challenge the longstanding focus of philosophy of science on theories, particularly on how they are justified [5.1, 3, 5]. Simulations, he argues, cannot simply be understood as novel ways to test theories. They are in fact rarely used to help justify theories, rather simulations apply existing theories in order to explore, explain and understand real and possible phenomena, or make predictions about how such phenomena will evolve in

time. Simulations open up a whole new set of philosophical issues concerning the practices and reliability of much modern science.

Winsberg's analysis of theory-based simulation shares much with *Cartwright's* [5.6] and *Morgan and Morrison's* [5.7] challenges to the role of theories. Like them, he starts by strongly disputing the presupposition that simulations are somehow deductive derivations from theory. Simulations are applied principally in the physical sciences when the equations generated from a theory to represent a particular phenomenon are not analytically solvable. The path from a theory to a simulation requires processes of computerization, which transform equations into tractable computable structures by relying on practices of discretization and idealization [5.8]. These practices employ specific transformations and simplifications in combination with those used to make tractable the application of theoretical equations to a specific phenomenon such as boundary conditions and symmetry assumptions. As such simulations are, according to *Winsberg* [5.1], better construed as particular articulations of a theory rather than derivations from theory. They make use of theoretical information and the credibility, explanatory scope and depth, of well-established theories, to provide warrant to simulations of particular phenomena. Inferences drawn by computational simulations have several features in this regard; they are *downward*, *motley* and *autonomous* [5.9]. Inferences are downward because they move from theory to the real world (rather than from the real world to theory). They are motley because they depend not just on theory but on a large range of extra-theoretical techniques and resources in order to derive inferences, such as approximation and simplification techniques, numerical methods, algorithmic methods, computer languages and hardware, and much trial and error. Finally, simulations are autonomous, in the sense of being autonomous from both theory and data. Simulations, according to Winsberg, are principally used to study phenomena where *data is sparse* and unavailable. These three conditions on inference from simulation require a specific philosophical evaluation of their reliability.

Such evaluation is complicated by the fact that relations between theory and inferences drawn from the simulation model are unclear and difficult to untangle. As *Winsberg* [5.1.9] suggests it is a complex task to unpack what role theories play in the final result given all these intervening steps. The fact that much validation of simulations is done through matching simulation outputs to the data, muddies the water further (see also [5.10]). A well-matched simulation constructed through a downward, motley and autonomous process from a nonetheless well-established theory raises the

question of the extent to which the confirmation afforded to the theory flows down to the simulation [5.2]. For instance, although fitting a certain data set might well be the dominant mode of validation of a simulation model, the model could be considered to hold outside the range of that data because the model applies a well-accepted theory of the phenomenon thought to hold under very general conditions.

There is widespread agreement that untangling the relations between theories and simulations, and the reliability of simulations built from theories will require more in depth investigation of the actual practices scientists use to justify the steps they make when building a simulation model. In the absence of such investigations discussions of justification are limited to considerations about whether a simulation fits the observational data or not. Among other things, this limitation hides from view important issues about the warrant of the various background steps that transform theoretical information into simulations [5.10]. In general, what is required is an *epistemology of simulation* which can discover rigorous grounds upon which scientists can and do sanction their results, and more properly the role of theory in modern science.

The concern with practices of simulation has opened up a new angle on the older discussion about the structure of theories. *Humphreys* [5.11] has used the entanglement of theory and simulation in modern scientific practice to reflect more explicitly upon the proper philosophical characterization of the structure of physical theories. Simulations, as with other models, are not logical derivations from theory which is a central, but incorrect, feature of the syntactic view. Humphreys also argues, however, that the now dominant semantic view of theories, which treats theories as nonlinguistic entities, is not adequate either. On the semantic view a syntactical formulation of a theory, and whether different formulations might be solvable or not, is not important for philosophical assessment of relations of representations to the world. Relations of representation are only in fact sensibly held by models not theories. Both Humphreys and Winsberg construe the semantic view as dismissing the role of theories in both normative and descriptive accounts of science, in place of models. But as *Humphreys* [5.12, p. 620] puts it, "the specific syntactic representation used is often crucial to the solvability of a theory's equations", and thus, the solvability of models derived from it. Computational tractability, as well as choices of approximation and simplification techniques, will depend on the particular syntax of a theory. Hence both the semantic and syntactic views are inadequate for describing theory in ways that capture their role in science.

## 5.2 Simulation not Driven by Theory

Investigations, such as those by Winsberg and others discussed in the previous section, have illustrated the importance of close attention to scientific practice and discovery when studying simulations. Simulation manifests application-intensive, rather than theoretical, processes of scientific investigation. As *Winsberg* [5.1] suggests choices about how to model a phenomenon reliably are developed often in the course of the to and fro *blood, sweat and tears* of the model-building process itself. Abstract armchair points of view, distant from an understanding of the contingent, but also technical and technological nature of these practices and their affordances, will not put philosophers in a position to create relevant normative assessments of good simulation practices. What has thus far been established by the accounts of theory-based simulation is that even in the case where there is an established theory of the phenomena, simulation model-building has a degree of independence from theory and theory-building.

However, though the initial focus on theory-based simulation in the study of simulation is not unsurprising given the historical preference in philosophy of science for treating *theory* as the principal unit of philosophical investigation, simulations are not just a tool of theory-driven science alone. Pushing philosophical investigation into model-building practices outside the domain of theory-driven science reveals whole new practices of scientific model production using computational simulations that are not in fact theory-based, in the sense of traditional physical sciences. Some of the most compelling and innovative fields in science today, including, for instance, big-data biology, systems biology and neuroscience, and much modeling in the social sciences, are not theory-driven. As *Winsberg* [5.5] admits (in response to *Parker* [5.13]), his description of simulation modeling is theory-centric, and neither necessarily applicable to understanding the processes by which simulation models are built in the absence of theory, nor an appropriate framework for assessing the reliability and informativeness of models built that way. This is not to say that characteristics of theory-based simulation are irrelevant to simulations that are not. Both theory and nontheory-based simulations share an independence of theory and there are likely to be similarities between them, but there are also profound differences.

One kind of simulation that is important in this regard is agent-based modeling. *Keller* [5.14] has labeled much agent-based modeling as *modeling from above* in the sense that such models are not constructed using a mathematical theory that governs the motions of agents. Agents follow local interactions rules. In many

fields in the social sciences and biology differential equations cannot be used to aggregate accurately agent or population behavior, but it is nonetheless possible to hypothesize or observe the structure of individual interactions. An agent-based model can be used to run those interactions over a large population to test whether the local structures can reproduce aggregate behavior [5.15]. As noted by *Grüne-Yanoff* and *Weirich* [5.16] agent-based modeling facilitates constructing remarkably complex models within computationally tractable constraints that often go well beyond what is possible with equation-based representations.

Agent-based models provide one exemplar of simulations that are not theory-driven. From an epistemological perspective, these simulations exhibit weak emergence [5.17]. The underlying mechanisms are thoroughly opaque to the users, and the way in which emergent properties come about can simply not be reassembled by studying the simulation processes. This opacity raises questions about the purpose and value of agent-based modeling. What kind of explanation and understanding does an agent-based simulation provide if the multiscale mechanisms produced in a simulation are cognitively inaccessible? Further, how is one to evaluate predictions and explanations from agent-based simulations which, in fields like ecology and economics, commonly simplify very complex interactions in order to create computationally tractable simulations. If a simplistic model captures a known behavior, can we trust its predictions? To address questions such as these we need an epistemology that can evaluate proposed techniques for establishing the robustness of agent-based models. One alternative is to argue that agent-based models require a novel epistemology that is able to rationalize their function as types of fictions rather than as representations [5.18, 19]. Another alternative, presented by *Grüne-Yanoff* and *Weirich* [5.16], is to argue that agent-based models provide in many cases functional rather than causal explanations of the phenomena they simulate [5.20]. Agent-based model simulations rarely control for all the potential explanatory factors that might be relevant to a given phenomenon, and any choice of particular interaction mechanism is usually thoroughly underdetermined. In practice, all possible mechanisms cannot be explored. But agent-based models can show reliably how particular lower-level capacities behave in certain ways, when modeled by suitably general interactions rules, and can constitute higher-level capacities no matter how multiply realized those interactions might be. Hence, such models, even though greatly simplified, can extract useful

information despite a large space of potential explananda.

Nontheory-driven forms of simulation such as agent-based models provide a basis for reflecting more broadly on the role theory plays in the production of simulations, and the warrant a theory brings to simulations based on it. Comparative studies of the kinds of arguments used to justify relying on a simulation should expose the roles well-established theories play. Our investigations of integrative systems biology (ISB) have revealed that not all equation-based modeling is theory-driven, if theory is construed in terms of theory in the physical sciences. The canonical meaning based on the physical sciences is something like a background body of laws and principles of a domain.

In the case of systems biology, researchers generally do not have access to such theory and in fact the kinds of theory they do make use of have a function different from what is usually meant by *theory* in fields like physics [5.21]. There are certain canonical theories in systems biology of how to mathematically represent interactions among, for instance, metabolites, in the form of sets of ordinary differential equations. These posit particular canonical mathematical forms for representing a large variety of interactions (see Biochemical Systems Theory [5.22]). In principle, for any particular metabolic network, if all the interactions and reactants are known, the only work for the modeler is to write down the equations for a particular network and calculate the parameters. The mathematics will take care of the rest since the mathematical formulations of interactions are general enough that any potential nonlinear behaviors should be represented if parameters are correctly fixed.

For the most part, however, these canonical frameworks do not provide the basic ontological information from which a representation of a system is ultimately drawn, in the way say that the Navier-Stokes equations of fluid dynamics describe fluids and their component interactions in a particular way. In practice, modelers in systems biology need to assemble that information themselves in the form of pathway diagrams which more or less list the molecules involved and then make their own decisions about how to represent molecular interactions. A canonical framework is better interpreted as a theory of how to approximate and simplify the information that the systems biologist has assembled about a pathway in order to reliably simulate the dominant dynamics of a network given sparse data and complex nonlinear dynamics. Hence, there is no real *theory articulation* in Winsberg's terms. Researchers do not articulate a general theory for a particular application. The challenge for systems biologists is to build a higher level or system level representation out of the

lower level information they possess. We have found that canonical templates mediate this process by providing a possible structure for gluing together this lower level information in a tractable way [5.21]. These theories do not offer any direct explanatory value by virtue of their use.

*Theory* can in fact be used not just to describe a body of laws and theoretical principles, but also to describe principles that instruct scientists on how to reliably build models of given classes of phenomena from a background theory. As Peck puts it [5.18, p. 393]:

“In traditional mathematical modeling, there is a long established research program in which standard methods, such as those used for differential equation modeling, are used to bring about certain ends. Once the variables and parameters and their relationships are chosen for the representation of the model, standard formulations are used to complete the modeling venture.”

If one talks about what physical scientists often start with it is not just the raw theory itself but well-established rules for formulating the theory and applying it with respect to a particular phenomenon. We might refer to this latter sense of *theory* as a theory of how to apply a background theory to reliably represent a phenomenon. The two senses of theory are exclusive. In the case of the canonical frameworks, what is meant by *theory* is something closer to this latter rather than former sense.

Additionally, the modelers we have studied are never in a position to rely on these frameworks uncritically and in fact no theory exists that specifies which representations to use that will reliably lead to a good representation in all data situations. In integrative systems biology the variety of data situations are very complex, and the data are often sparse and are rarely adequate for applying a set mathematical framework. This forces researchers in practice into much more intensive and adaptive model-building processes that certainly share much in common with the back and forth processes Winsberg talks about in the context of theory application. But these processes have the added and serious difficulty that the starting points for even composing the mathematical framework out of which a model should be built are open-ended and need to be decided based on thorough investigation of the possibilities with the specific data available.

Canonical frameworks are just an option for modelers and do not drive the model-building process in the way physical theories do. Currently, systems biology generally lacks effective theory of either kind. Modelers have many different choices about how to confront a particular problem that do not necessarily

involve picking up a canonical framework or sticking to it. MacLeod and Nersessian [5.21] have documented how the nontheory-derived model-building processes work in these contexts. Models are strategic adaptations to a complex set of constraints system biologists are working under [5.23]. Among these constraints are:

- Constraints of the biological problem: A model must address the constraints of the biological problem, such as how the redox environment is maintained in a healthy cell. The system involved is often of considerable complexity.
- Informational/data constraints: There are constraints on the accessibility and availability of experimental data and molecular and system parameters for constructing models.
- Cost constraints: ISB is data-intensive and relies on data that often go beyond what are collected by molecular biologists in small scale experiments. However, data are very costly to obtain.
- Collaboration constraints: Constraints on the ability to communicate effectively with experimental collaborators with different backgrounds or in different fields in order to obtain expert advice or new data. Molecular biologists largely do not understand the nature of simulation modeling, do not understand the data needs of modeling, and do not see the cost-benefit of producing the particular data systems biologists ask from them.
- Time-scale constraints: Different time scales operate with respect to generating molecular experimental data versus computational model testing and construction.
- Infrastructure constraints: There is little in the way of standardized databases of experimental information or standardized modeling software available for systems biologists to rely upon.
- Knowledge constraints: Modelers' lack knowledge of biological systems and experimental methods limits their understanding of what is biologically plausible and what reliable extrapolations can be made from the data sets available.
- Cognitive constraints: Constraints on the ability to process and manipulate models because of their complexity, and thus constraints on the ability to comprehend biological systems through modeling.

Working with these constraints requires them to be *adaptive problem-solvers*. Given the complexity of the systems, lack of data, and the ever-present problem of computational tractability, researchers have to experiment with different mathematical formulations, different parameter-fixing algorithms and approximation techniques in highly intensive trial and error processes.

They build models in nest-like fashion in which bits of biological information and data and mathematical and computational techniques, get combined to create stable models. These processes transform not only the shape of the solutions, but also the problems, as researchers figure out what actual problem can be solved with the data at hand. Simulation plays a central exploratory role in the process. This point goes further than Lenhard's idea of an explorative cooperation between experimental simulation and models [5.8]. Simulation in systems biology is not just for experimenting on systems in order to *sound out the consequences of a model* [5.8, p. 181], but plays a fundamental role in incrementally building the model and learning the relevant known and sometimes unknown features of a system and gaining an understanding of its dynamics. Simulation's roles as a cognitive resource make the construction of representations of complex systems without a theoretical basis possible (see also [5.24, 25]).

Similar conclusions have been drawn by Peck for ecology which shares with systems biology the complexity in its problems and a lack of generalizable theory. As Peck [5.18, p. 393] points out:

“there are no formal methodological procedures for building these types of models suggesting that constructing an ecological simulation can legitimately be described as an art.”

This situation promotes methodological pluralism and creative methodological exploration by modelers. Modelers in these contexts thus focus our attention on the deeper roles (sometimes called heuristic roles [5.5]) that simulation plays in the ability of researchers to explore potential solutions in order to solve complex problems.

These roles have added epistemological importance when it is realized that the *downward* character of simulation can be fact reversed in both senses we have mentioned above. This is a potentially significant difference between cases of theory and nontheory-driven simulation. Consider again systems biology. Firstly, the methodological exploration we witness amongst the researchers we have studied can be rationalized as precisely an attempt by the field to establish a good theory of how to build models of biological systems that work well given a variety of data situations. Since the complexities of these systems and computational constraints make this difficult to know at the outset, the field needs its freedom to explore the possibilities. Lab directors do encourage exploration, and part of the reason they do is to try to glean which practices work well and which do not given a lack of knowledge of what will work well for a given problem.

Secondly, systems biology aspires to a theory of biological systems which will detail general system-level characteristics of biological systems but also the design principles underlying biological networks [5.26, 27]. What is interesting about this theory, if it does emerge, is that it will in fact be theory generated *by* simulation rather than the other way around. Simulation makes possible the exploration of quite complex systems for generalities that can form the basis of a theory of systems biology. As such the use of simulations can also be upwards, not just downwards, to perhaps an unprecedented extent. Upward uses of simulation requires analysis that appears to fit better with

more traditional philosophical analysis of how theories are in fact justified, only in this case robust simulation models will possibly be the more significant source of evidence rather than traditional experiment and observation. How this affects the nature and reliability of our inferences to theory, and what kind of resemblance such theory might have to theory in physics, is something that will need investigation. Thus, further exploration of nontheory-driven modeling practices stand to provide a rich ground for investigation of novel practices that are emerging with simulation, but also for exploring the roles and meanings of *theory*.

### 5.3 What is Philosophically Novel About Simulation?

The question of whether or not simulation introduces new issues into the philosophy of science has emerged as a substantial debate in discussions of computational simulation. *Winsberg* [5.1, 3–5] and *Humphreys* [5.11, 12] are the major proponents of the view that simulation requires its own epistemology. Winsberg, for instance, takes the view that simulations exhibit “distinct epistemological characteristics . . . novel to the philosophy of science” [5.9, p. 443]. Winsberg and Humphreys make this assertion on the basis of the points we outlined in Sec. 5.2; namely, 1) the traditional limited concern of philosophy of science with the justification of theory, and 2) the relative autonomy of simulations and simulation-building from the theory. The steps involved in generating simulations, such as applying approximation methods designed to generate computational tractability, are novel to science. These steps do not gain their legitimacy from a theory but are “autonomously sanctioned” [5.1, p. 837]. Winsberg argues, for instance, that while idealization and approximation methods have been discussed in the literature it has mostly been from a representational perspective in terms of how idealized and approximate models represent or resemble the world and in turn justify the theories on which they are based. But since simulations are often employed where data are sparse, they cannot usually be justified by being compared with the world alone. Simulations must be assessed according to the reliability of the processes used to construct them, and these often distinct and novel techniques require separate philosophical evaluation. Mainstream philosophy of science with its focus on theoretical justification does not have the conceptual resources for accounting for applications using computational methods. Even where theory is concerned, both Humphreys and Winsberg maintain that neither of the established semantic and syntactic conception of theories, conceptions which fo-

cus on justification and representation, can account for how theories are applied or justified in simulation modeling.

However, *Frigg* and *Reiss* [5.28] have countered that these claims were overblown and in fact simulation raises no new questions or problems that are specific to simulation alone. Part of the disagreement might simply come down to whether one construes philosophy of science narrowly or broadly by limiting *philosophical questions* to in-principle and normative issues, while avoiding practical methodological ones. Another part of the disagreement is over how one construes *new issues* or *new questions* for philosophy, since certainly at some level the basic philosophical questions about how representations represent and what makes them reliably do so, are still the same questions.

To some extent, part of the debate might be construed as a disagreement over the relevance of contexts of discovery to philosophy of science. Classically contexts of discovery, the scientific contexts in which model-building takes place, are considered irrelevant to normative philosophical assessments of whether those models are justified or not. *Winsberg* [5.3] and *Humphreys* [5.12] seem willing to assert that one of the lessons for philosophy of science from simulation is that practical constraints on scientific discovery matter for constructing relevant normative principles – both in terms of evaluating current practice, which in the case of simulation-building is driven by all kinds of practical constraints, and in terms of normatively directing practice sensitively within those constraints.

Part of the motivation for using the discovery/justification distinction to define philosophical interest and relevance is the belief that there is a clear distinction between the two contexts. Arguably Frigg and Reiss are reinforcing the idea of a clear distinction by relying on widespread presupposition that

validation and verification are distinct independent processes [5.4]. Validation is the process of establishing that a simulation is a good representation, a quintessential concept of justification. Verification is the process of ensuring that a computational simulation adequately captures the equations from which it is constructed. Verification, according to Frigg and Reiss, represents the only novel aspects of modeling that simulation introduces. Yet it is a purely mathematical exercise that is of no relevance to questions of validation. As such, simulations involve no new issues of justification beyond those of ordinary models. Winsberg [5.3, 4], however, counters that there is, in practice, no clear division between processes of verification and validation. The equations chosen to represent a system are not simply selected on the basis of how valid they are, but also on the basis of decisions about computational tractability. Much of what validates a representation in practice occurs at the end stage, after all the necessary techniques of numerical approximation and discretization have been applied, by comparing the results of simulations with the data. As such, [5.5]:

“If we want to understand why simulation results are taken to be credible, we have to look at the epistemology of simulation as an integrated whole, not as clearly divided into verification and validation – each of which would look inadequate to the task.”

Hence what would otherwise seem to be distinct discovery and justification processes are in the context computational simulation interwoven.

Frigg and Reiss are right at some level that simulations do not change basic epistemological questions connected to the justification of models. They are also right that Winsberg in his downward, motley and autonomous description of simulation, does not reveal any fundamentally new observations on model-building that have not already been identified as issues by philosophers discussing traditional modeling. However, what appears to be really new in the case of simulation is: 1) the complexity of the philosophical problems of representation and reliability, and 2) the different methodological and epistemological strategies that have become available to modelers as a result of simulation.

Winsberg, in reply to Frigg and Reiss, has clarified what he thinks as novel about theory-based simulation as the *simultaneous confluence* of downward, motley and autonomous features of model-building [5.4]. It is the reliability and validity of the complex modeling processes instantiated by these three features that must be accounted for by an epistemology of simulation, and no current philosophical approaches are adequate to do so, particularly not those within traditional philosophical boundaries of analysis.

As a first step in helping with this task of assessing reliability and validity of simulation, philosophers such as Winsberg [5.29] have drawn lessons from comparison with experimentation, which they argue shares much with simulation in both function (enabling, for instance, in silico experiments) and also in terms of how the reliability of simulations is generated. Scientific researchers try to control for error in their simulations, and fix parameters, in ways that seem analogous to how experimenters calibrate their devices. Simulations build up credibility over long time scales and may have lives of their own independent of developments in other parts of science. These observations suggest a potentially rich analogy between simulations and Hacking’s account of experimentation [5.29]. In a normative step, based on these links, Parker [5.10] has suggested that in fact Mayo’s [5.30] rigorous error-statistical approach for experimentation should be an appropriate starting point for more thorough evaluation of the results of simulations. Simulations need to be evaluated by the degree to which they avoid false positives when it comes to testing hypotheses by successfully controlling for potential sources of error that creep in during the simulation process. At the same time a rather vigorous debate has emerged concerning the clarification of the precise epistemological dissimilarities or disanalogies between simulation and traditional experimentation (see for instance [5.31–36]). This question is in itself of independent philosophical interest for assessing the benefits and value of each as alternatives, but should also help define the limits of the relevance of experimentation as a model for understanding and assessing simulation practices.

From our perspective, however, the new methodological and epistemological strategies that modelers are introducing in order to construct and guarantee the reliability of simulation models could prove to be the most interesting and novel aspect of simulation with which philosophers will have to grapple. Indeed, while much attention has focused on the contrasts and similarities between simulations, experiments and simulation experiments, no one has called attention to the fact that real-world experiments and simulations are also being used in concert to enhance the ability of researchers to handle uncertain complex systems. One of the labs we have studied conducts *bimodal* modeling, where the modelers conduct their own experiments in the service of building their models. We have analyzed the case of one modeler’s behavior in which model-building, simulation and experimentation were tightly interwoven [5.37]. She used a conjunction of experiment and simulation to triangulate on errors and uncertainties in her model, thus demonstrating that the two can be combined in practice in sophisticated ways.

Her model-building would not have been possible without the affordances of both simulation and her ability to perform experimentation precisely adapted to test questions about the model as she was in the process of formulating it. Simulation and experiment closely coupled in this fashion offers the possibility of extending the capacity to produce reliable models of complex phenomena.

Bimodal modeling is relatively easy to characterize epistemologically since experimentation is used to validate and check the simulations as the model is being constructed. Simulations are not relied on independent of experimental verification. Often, however, experimental or any kind of observational data are hard to come by for practical or theoretical reasons. More philosophically challenging will be to evaluate the new epistemological strategies researchers are in fact developing for drawing inferences in these often deeply uncertain and complex contexts with the aid of computation. *Parker* [5.38, 39], for instance, identifies the practice in climate science and meteorology of ensemble modeling. No theory of model-building exists that tells climate and weather modelers how to go from physical theory to reliable models. Different formulations using different initial conditions, models structures and different parameterizations of those models that fit the observational data can be developed from the physical theory. In this situation modelers average over results from large collections of models, using different weighting schemas, and argue for the validity of these results on the basis that these models collectively represent the possibility space. However, considerable philosophical questions emerge as to the underlying justifiability of these ensemble practices and the probability weightings being relied upon. Background theory can provide little guidance in this context and in the case of climate modeling there is little chance for predictively testing performance. Further, the robustness of particular ensemble choices is often very low and justifications for picking out particular ensembles are rarely carefully formulated.

The ability to generate and compare large numbers of complex models in this way is a development of modern computational power. In our studies we have also come across novel argumentation, particularly connected with parameter-fixing [5.40]. Because the parameter spaces these modelers have to deal with are so complex, there is almost no chance of getting a best fit solution. Instead modelers produce multiple models often using Monte Carlo techniques that converge on similar behavior and output. These models have different parameterizations and ultimately represent the underlying mechanisms of the systems differently. However, modelers can nonetheless make specific

arguments about network structure and dynamic relationships among specific variables. There is not usually any well-established theory that licenses these arguments. The fact that the models converge on the same relevant results is motivation for inferring that these models are right at least about those aspects of the system for which they are designed to account. Unfortunately, because access to real-world experimentation is quite difficult, it is hard to judge how reliable this technique is in producing robust models. What is novel about this kind of strategy is that it implicitly treats parameter-fixing as an opportunity, not just a problem, for modelers. If instead of trying to capture the dynamics of whole systems modelers just fix their goals on capturing robust properties and relations of a system, the potential of finding results that work within these constraints in large parameter-spaces increases, and from the multiple models obtained modelers can pare down to those that converge. The more complex problem thus seems to allow a pathway for solving a simpler one. Nonetheless, whether we should accept these kinds of strategies as reliable and the models produced as robust remains the fundamental question, and an overarching question for the field itself. It is a reasonable reaction to suspect that something important is being given up in the process, which will affect how well scientists can assess the reliability and importance of the models they produce. Whether the power computational processes can adequately compensate for the potential distortions or errors introduced is one of the most critical and novel epistemological questions for philosophy today.

The kinds of epistemological innovations we have been considering raise deeper questions about the purposes of simulation, particularly in terms of traditional epistemic categories like understanding, explanation and so on. Of course at one extreme some simulations of the purely data-driven kind is purely phenomenological. Theory plays no role in its generation, and is not sought as its outcome. However in other cases some form of understanding at least is sought. In many cases though, where theory might be thought the essential agent of understanding, the complexity of the equations and resulting complexity of the computational processes that instantiate them, simply block any way of decomposing the theory or theoretical model in order to understand how the theory might explain a phenomena and thus assess the accuracy and plausibility of the underlying mechanisms it might prescribe. *Humphreys* labels this *epistemic opacity* [5.11]. *Lenhard* [5.41] in turn identifies a form of pragmatic understanding that can replace theoretical understanding when a simulation model is epistemically opaque. This form of understanding is pragmatic in the sense of being an understanding of how



to *control* and *manipulate* phenomena, rather explain them using background theoretical principles and laws. Settling for this form of understanding is a choice made by researchers in order to handle more complex problems and systems using simulations. But it is a novel one in the context of physics and chemistry. In systems biology we recognize something similar [5.40]. Researchers give up accurate mechanistic understanding of their systems for more pragmatic goals of gaining network control, at least over specific variables. To do so they use simplification and parameter-fitting techniques that obscure the extent to which their models capture the underlying mechanisms. Mechanistic ex-

planation is thus given up, for some weaker form of understanding.

Finally, computational modeling and simulation in the situations we have been considering in this section are driving a profound shift in the nature and level of human cognitive engagement in scientific production processes and their outputs [5.12, 24, 25, 42, 43]. So much of philosophy of science has been based on intuitive notions of human cognitive abilities. Our concepts of explanation and understanding are constructed implicitly on the basis of what we can grasp as humans. With simulation and big-data science those kinds of characterizations may no longer be accurate or relevant [5.44].

## 5.4 Computational Simulation and Human Cognition

It is on this last point that we turn to consider the ways in which human cognitive processes are implicated in processes of simulation model-building. Computational science, of the nonbig data or nonmachine learning kind which we have focused on here, is as Humphrey's calls it, a "hybrid scenario" as opposed to an "automated scenario" [5.12, p. 616]. In his words:

"This distinction is important because in the hybrid scenario, one cannot completely abstract from human cognitive abilities when dealing with representational and computational issues. . . . We are now faced with a problem, which we can call the *anthropocentric predicament*, of how we, as humans, can understand and evaluate computationally-based scientific methods that transcend our own abilities."

Unlike machine-learning contexts, computational modeling is in many cases a practice of using computation to extend traditional modeling practices and our own capabilities to draw insight out of low-data contexts and complex systems for which theory provides at best a limited guide. In this way cognitive capacities are often heavily involved. The *hybrid* nature of computational science thus motivates the need for understanding how human agents cognitively engage with and control opaque computational processes, and in turn draw information out of them. Evaluating these processes – their productiveness and reliability – requires in the first step having some understanding of them. As we will see, although computational calculation processes are beyond our abilities, at least in the case of systems biology the use of computation by modelers is often far more integrated with their own cognitive processes and understanding, and thus far more under their control, than we might think.

As we have seen there are several lines of philosophical research on computational simulation that un-

derscore it is through the processes of model-building – taken to comprise the incremental and interwoven processes of constructing the model and investigating its dynamics through simulation – that the modeler comes to develop at least a pragmatic understanding of the phenomena under investigation. Complex systems, such as investigated in systems biology, present perhaps the extreme case in which these practices are the *primary* means through which modelers, mostly nonbiologists, develop understanding of the systems. In our investigations, modelers called the building and *running* of their models under various conditions *getting a feel for the model*, which enables them to get a feel for the dynamics of the system.

In our investigations we have witnessed that modelers (mainly engineers) with little understanding of biology have been able to provide novel insights and highly significant predictions, later confirmed by biological collaborators, for the systems they are investigating through simulation. How is it possible that engineers with little to no biological training can be making significant biological discoveries? A related question concerns how complete novices are making scientific discoveries through simulations crowdsourced by means of video games such as Foldit and EteRNA, which appear to enable nonscientists to quickly build accurate/veridical structures representing molecular entities they had no prior knowledge of [5.45, 46]. *Nersessian* and *Chadrasekharan*, individually and together [5.24, 25, 42, 47–49], have argued that the answer to this question lies in understanding how computational simulation enhances human cognition in discovery processes. Because of the visual and manipulative nature of the crowdsourcing cases, the answer points in the direction of the *coupling* of the human sensorimotor systems with simulation models. These crowdsourcing models represent conceptual knowledge developed by the sci-

entific community (e.g., structure of proteins) as computational representations with a control interface that can be manipulated through the gamer's actions. The interface enables these novices to build new representations drawing on tacit/implicit sensorimotor processes. Although the use of crowdsourcing simulations in scientific problem solving is new, the human sensorimotor system has been used explicitly to detect patterns, especially in dynamic data generated by computational models, since the dawn of computational modeling. Entire disciplines and methods have been built using visualized patterns on computer screens. Complexity theory [5.50, 51], artificial life [5.52, 53] and computational chemistry [5.54, 55] provide a few exemplars where significant discoveries have been made.

Turning back now to the computational simulations used by scientists that we have been discussing, all of the above suggests that the model-building processes facilitate a close coupling between the model and the researcher's mental modeling processes even in the absence of a dynamic visualization. The building process manipulates procedural and declarative knowledge in the imagination and in the representation, creating a *coupled cognitive system* of model and modeler [5.25, 42, 43, 48, 56, 57]. This coupling can lead to explicit understanding of the dynamics of the system under investigation. The notion of a coupled cognitive system is best understood in terms of the framework of distributed cognition [5.58, 59], which was developed to study cognitive processes in complex task environments, particularly where external representations and other cognitive artifacts and, possibly, groups of people, accomplish the task. The primary unit of analysis is the socio-technical system that generates, manipulates and propagates representations (internal and external to people). Research leading to the formation of the distributed cognition framework has focused largely on the use of existing representational artifacts and less so on the building/creation of the artifacts. The central metaphor is that of the human *offloading* complex cognitive processes such as memory to the artifact, which, for example, in the canonical exemplar of the speed bug that marks critical airspeeds for a particular flight, replaces complex cognitive operations with a perceptual operation and provides a publically available representation that is shared between pilot and co-pilot.

In the research cited above, we have been arguing that *offloading* is not the right metaphor for understanding the cognitive enhancements provided through the building of novel computational representations. Rather, the metaphor should be that of *coupling* between internal and external representations. Delving into the modifications needed of the distributed cognition framework to accommodate the notion of a coupled

cognitive system would take us too far afield in this review (but see [5.25]). Instead, we will flesh out the notion a bit by noting some of the ways in which building and using simulation models enhance human cognitive capabilities and, in particular, extend the capability of the imagination system for simulative model-based reasoning.

A central, but yet not well-researched premise of distributed cognition is, as Hutchins has stated succinctly, that “humans create cognitive powers by creating the environments in which they exercise those powers” [5.58, p. 169]. Since building modeling-environments for problem solving is a major component of scientific research [5.49], scientific practices provide an especially good locus for examining the human capability to extend and create cognitive powers. In the case of simulation model-building, the key question is: *What are the cognitive changes involved in building a simulation model and how do these lead to discoveries?* The key cognitive change is that over the course of many iterations of model-construction and simulation, the model gradually becomes coupled with the modeler's imagination system (mental model simulation), which enables the modeler to explore different scenarios. The coupling allows *what if* questions in the mind of the modeler to be turned into detailed explorations of the system, which would not be possible in the mind alone. The computational model enables this exploration because as it is incrementally built using many data sets, the model's behavior, in the systems biology case, for instance, comes to parallel the dynamics of the pathway. Each replication of experimental results adds complexity to the model and the process continues until the model is judged to fit all available data well. This judgment is complex, as it is based on a large number of iterations where a range of factors such as sensitivity, stability, consistency, computational complexity and so forth are explored. As the model gains complexity it starts to reveal or expose many details of the system's behavior enabling the modeler to interrogate the model in ways that are not possible in the mind alone (thought experimenting) or in real-world experiments. It makes evident many details of the system's behavior that the modeler could not have imagined alone because of the fine grain and complexity of the details.

The parallel between computation simulation experimenting and thought experimenting is one philosophers have commented on, but the current framing of the discussion primarily centers on the issue of interpreting simulations and whether computational simulations should be construed as *opaque* thought experiments [5.60, 61]. Di Paolo et al. [5.60] have argued that computational models are more opaque than thought experiments, and as such, require more *system-*

*atic enquiry* through probing of the model's behavior. In a similar vein, Lenhard [5.61] has claimed that thought experiments are more *lucid* than computational models, though it is left unclear what is meant by *lucid* in this context, particularly given the extensive discussions around what specific thought experiments actually demonstrate. In the context of the discussion of the relation of thought experimenting and computational simulation, we have argued that the discussion should be shifted from issues of interpretation to a process-oriented analysis of modeling [5.47]. Nersessian [5.62] casts thought experimenting as a form of *simulative model-based reasoning*, the cognitive basis of which is the human capacity for mental modeling. Thought experiments (conceptual models), physical models [5.63] and computational models [5.47, 48] form a spectrum of simulative model-based reasoning in that all these types of modeling generate and test counterfactual situations that are difficult (if not impossible) to implement in the real world. Both thought experiments and computational models support simulation of counterfactual situations, however, while thought experiments are built using concrete elements, computational models are built using variables. Simulating counterfactual scenarios beyond the specific one constructed in the thought experiment is difficult and requires complex cognitive transformations to move away from the concrete case to the abstract, generic case. On the other hand, computational simulation constructs the abstract, generic case from the outset. Since computational models are made entirely of variables, they naturally support thinking about parameter spaces, possible variations to the design seen in nature, and why this variation occurs rather than the many others that are possible.

Thought experiments are a product of a resource environment in science where the only tools available were writing implements, paper (blackboards, etc.) and the brain. Computational models create cognitive enhancements that go well beyond those resources and enable scientists to study the complex, dynamic and nonlinear behaviors of the phenomena that are the focus of contemporary science.

Returning to the nature of the cognitive enhancements created, the coupling of the computational model with the modeler's imagination system significantly enhances the researcher's natural capacity for simulative model-based reasoning, particularly in the following ways:

- It allows running many more simulations, with many variables at gradients not perceivable or manipulable by the mind, which can be compared and contrasted.

- It allows testing what-if scenarios with changes among many variables that would be impossible to do in the mind.
- It allows stopping the simulation at various points and checking and tracking its states. If some desirable effect is seen, variables can be tweaked in process to get that effect consistently.
- It allows taking the system apart as modules, simulating them, and putting them together in different combinations.
- It allows changing the time in which intermediate processes kick in.

These complex manipulations expose the modeler to system-level behaviors that are not possible to examine in either thought alone or in real-world experimentation. The processes involved in building the distributed model-based reasoning system comprising simulation model and modeler enhance several cognitive abilities. Here we will conclude by considering three (for a fuller discussion see [5.25]). First, the model-building process brings together a range of experimental data. Given Internet search engines and online data bases, current models synthesize more data than ever before and create a synthesis that exists nowhere in the literature and would not be possible for modelers or biologists to produce on their own. In effect, the model becomes a *running literature review*. Thus, modeling enhances the synthesizing and integrating capabilities of the modeler, which is an important part of the answer as to how a modeler with scant biological knowledge can make important discoveries. Second, an important cognitive effect of the model-building is to enhance the modeler's powers of abstraction. Most significantly, through the gradual process of thousands of runs of simulations and analyses of system dynamics for these, the modeler gains an external, global view of the system as a whole. Such a global view would not be possible to develop just from mental simulation, especially since the interactions among elements are complex and difficult to keep track of separately. The system view, together with the detailed understanding of the dynamics, provides the modeler with an intuitive sense (*a feeling for the model*) of the biological mechanisms that enables her to extend the pathway structure in a constrained fashion to accommodate experimental data that could not be accounted for by the current pathway from which the model started. Additionally, this intuitive sense of the mechanism built from interaction with the model helps to explain the success of the crowdsourcing models noted above (see also [5.64]).

Finally, the model enhances the cognitive capacity for counterfactual or possible-worlds thinking. As noted in our discussion of thought experimenting, the

model-building process begins by capturing the reactions/interactions using variables. Variables provide a place-holder representation, which when interpreted with combinations of numbers for these variables, can generate model data that parallels the known experimental data. One interesting feature of the place-holder representation is that it provides the modeler with a flexible way of thinking about the reactions, as opposed to the experimentalist who works with only one set of values. Once the model is using the experimental values, the variables can take any set of values, as long as they generate a fit with the experimental data. The modeler is able to think of the real-world values as only *one possible scenario*, to examine why this scenario is commonly seen in nature, and envision other scenarios that fit. Thinking in variables supports both the objective modelers often have of altering or redesigning a reaction (such as the thickness of lignin in plant wall for biofuels) and the objective of developing generic design patterns and principles. More broadly, the variable representation significantly expands the imagination space of the modeler, enabling counterfactual explorations of possible worlds that far outstrip the potential of thought experimenting alone.

A more microscopic focus like this one on the actual processes by which computational simulation is coupled with the cognitive processes of the modeler begins to help break down some of the mystery and seeming inscrutability surrounding computation conveyed by the

idea that computational processes are offloaded automated processes from which inferences are derived. The implications of this research into hybrid nature of simulation modeling are that modelers might often have more control over and insight into their models and their alignment with the phenomena than philosophers have realized. Given the emphasis placed in published scientific literature on fitting the data and predictive success for validating simulations, we might be missing out on the important role that these processes internal to the model-building or discovery context appear to be playing (from a microanalysis of practice) in support of the models constructed. Indeed, the ability of computational modeling to support highly exploratory investigative processes makes it particularly relevant for philosophers to have fine-grained knowledge of model-building processes in order to begin to understand why models work as well as they do and how reliable they can be considered to be.

**Acknowledgments.** We gratefully acknowledge the support of the US National Science Foundation grant DRL097394084. Our analysis has benefited from collaboration with members of the Cognition and Learning in Interdisciplinary Cultures (CLIC) Research Group at the Georgia Institute of Technology, especially with Sanjay Chandrasekharan. Miles MacLeod's participation was also supported by a postdoctoral fellowship at the TINT Center, University of Helsinki.

## References

- 5.1 E. Winsberg: Sanctioning models: The epistemology of simulation, *Sci. Context* **12**(2), 275–292 (1999)
- 5.2 E. Winsberg: Models of success vs. the success of models: Reliability without truth, *Synthese* **152**, 1–19 (2006)
- 5.3 E. Winsberg: Computer simulation and the philosophy of science, *Philos. Compass* **4**(5), 835–845 (2009)
- 5.4 E. Winsberg: *Science in the Age of Computer Simulation* (Univ. of Chicago Press, Chicago 2010)
- 5.5 E. Winsberg: Computer simulations in science. In: *The Stanford Encyclopedia of Philosophy*, ed. by E.N. Zalta (Stanford Univ., Stanford 2014), <http://plato.stanford.edu/cgi-bin/encyclopedia/archinfo.cgi?entry=simulations-science>
- 5.6 N. Cartwright: *The Dappled World: A Study of the Boundaries of Science* (Cambridge Univ. Press, Cambridge 1999)
- 5.7 M.S. Morgan, M. Morrison: Models as mediating instruments. In: *Models as Mediators: Perspectives on Natural and Social Science*, ed. by M.S. Morgan, M. Morrison (Cambridge Univ. Press, Cambridge 1999)
- 5.8 J. Lenhard: Computer simulation: The cooperation between experimenting and modeling, *Philos. Sci.* **74**(2), 176–194 (2007)
- 5.9 E. Winsberg: Simulations, models, and theories: Complex physical systems and their representations, *Philos. Sci.* **68**(3), 442–454 (2001)
- 5.10 W. Parker: Computer simulation through an error-statistical lens, *Synthese* **163**(3), 371–384 (2008)
- 5.11 P. Humphreys: *Extending Ourselves: Computational Science, Empiricism, and Scientific Method* (Oxford Univ. Press, New York 2004)
- 5.12 P. Humphreys: The philosophical novelty of computer simulation methods, *Synthese* **169**, 615–626 (2009)
- 5.13 W. Parker: Computer simulation. In: *The Routledge Companion to Philosophy of Science*, ed. by S. Psillos, M. Curd (Routledge, London 2013) pp. 135–145
- 5.14 E. Fox Keller: Models, simulation, and computer experiments. In: *The Philosophy of Scientific Experimentation*, ed. by H. Radder (Univ. of Pittsburgh Press, Pittsburgh 2003) pp. 198–215
- 5.15 S. Peck: Agent-based models as fictive instantiations of ecological processes, *Philos. Theory Biol.* **4**, 1–12 (2012)
- 5.16 T. Grüne-Yanoff, P. Weirich: Philosophy of simulation, simulation and gaming, *Interdiscip. J.* **41**(1), 1–31 (2010)

- 5.17 M.A. Bedau: Weak emergence and computer simulation. In: *Models, Simulations, and Representations*, ed. by P. Humphreys, C. Imbert (Routledge, New York 2011) pp. 91–114
- 5.18 S. Peck: The Hermeneutics of ecological simulation, *Biol. Philos.* **23**(3), 383–402 (2008)
- 5.19 R. Frigg: Models and fiction, *Synthese* **172**(2), 251–268 (2010)
- 5.20 T. Grüne-Yanoff: The explanatory potential of artificial societies, *Synthese* **169**(3), 539–555 (2009)
- 5.21 M. MacLeod, N.J. Nersessian: Building simulations from the ground-up: Modeling and theory in systems biology, *Philos. Sci.* **80**(4), 533–556 (2013)
- 5.22 E.O. Voit: *Computational Analysis of Biochemical Systems: A Practical Guide for Biochemists and Molecular Biologists* (Cambridge Univ. Press, Cambridge 2000)
- 5.23 M. MacLeod, N.J. Nersessian: The creative industry of systems biology, *Mind Soc.* **12**, 35–48 (2013)
- 5.24 S. Chandrasekharan, N.J. Nersessian: Building cognition: The construction of external representations for discovery, *Cogn. Sci.* **39**(8), 1727–1763 (2015), doi:10.1111/cogs.12203
- 5.25 S. Chandrasekharan, N.J. Nersessian: Building cognition: The construction of computational representations for scientific discovery, *Cogn. Sci.* **39**(8), 1727–1763 (2015)
- 5.26 H. Kitano: Looking beyond the details: A rise in system-oriented approaches in genetics and molecular biology, *Curr. Genet.* **41**(1), 1–10 (2002)
- 5.27 H.V. Westerhoff, D.B. Kell: The methodologies of systems biology. In: *Systems Biology: Philosophical Foundations*, ed. by F.C. Boogerd, F.J. Bruggeman, J.S. Hofmeyr, H.V. Westerhoff (Elsevier, Amsterdam 2007) pp. 23–70
- 5.28 R. Frigg, J. Reiss: The philosophy of simulation: Hot new issues or same old stew, *Synthese* **169**, 593–613 (2009)
- 5.29 E. Winsberg: Simulated experiments: Methodology for a virtual world, *Philos. Sci.* **70**(1), 105–125 (2003)
- 5.30 D.G. Mayo: *Error and the Growth of Experimental Knowledge* (Univ. of Chicago Press, Chicago 1996)
- 5.31 N. Gilbert, K. Troitzsch: *Simulation for the Social Scientist* (Open Univ. Press, Philadelphia 1999)
- 5.32 F. Guala: Models, simulations, and experiments. In: *Model-based reasoning: Science, technology, values*, ed. by L. Magani, N.J. Nersessian (Kluwer Academic/Plenum Publishers, New York 2002) pp. 59–74
- 5.33 F. Guala: Paradigmatic experiments: The ultimatum game from testing to measurement device, *Philos. Sci.* **75**, 658–669 (2008)
- 5.34 M. Morgan: Experiments without material intervention: Model experiments, virtual experiments and virtually experiments. In: *The Philosophy of Scientific Experimentation*, ed. by H. Radder (University of Pittsburgh Press, Pittsburgh 2003) pp. 216–235
- 5.35 W. Parker: Does matter really matter? Computer simulations, experiments and materiality, *Synthese* **169**(3), 483–496 (2009)
- 5.36 E. Winsberg: A tale of two methods, *Synthese* **169**(3), 575–592 (2009)
- 5.37 M. MacLeod, N.J. Nersessian: Coupling simulation and experiment: The bimodal strategy in integrative systems biology, *Stud. Hist. Philos. Sci. Part C* **44**, 572–584 (2013)
- 5.38 W.S. Parker: Predicting weather and climate: Uncertainty, ensembles and probability, *Stud. Hist. Philos. Sci. Part B* **41**(3), 263–272 (2010)
- 5.39 W.S. Parker: Whose probabilities? Predicting climate change with ensembles of models, *Philos. Sci.* **77**(5), 985–997 (2010)
- 5.40 M. MacLeod, N.J. Nersessian: Modeling systems-level dynamics: Understanding without mechanistic explanation in integrative systems biology, *Stud. Hist. Philos. Sci. Part C* **49**(1), 1–11 (2015)
- 5.41 J. Lenhard: Surprised by a nanowire: Simulation, control, and understanding, *Philos. Sci.* **73**(5), 605–616 (2006)
- 5.42 N.J. Nersessian: *Creating Scientific Concepts* (MIT Press, Cambridge 2008)
- 5.43 N.J. Nersessian: How do engineering scientists think? Model-based simulation in biomedical engineering research laboratories, *Top. Cogn. Sci.* **1**, 730–757 (2009)
- 5.44 W. Callebaut: Scientific perspectivism: A philosopher of science's response to the challenge of big data biology, *Stud. Hist. Philos. Sci. Part C* **43**(1), 69–80 (2012)
- 5.45 J. Bohannon: Gamers unravel the secret life of protein, *Wired* **17** (2009), [http://www.wired.com/medtech/genetics/magazine/17-05/ff\\_protein](http://www.wired.com/medtech/genetics/magazine/17-05/ff_protein), Last accessed 06–06–2016
- 5.46 F. Khatib, F. DiMaio, Foldit Contenders Group, Foldit Void Crushers Group, S. Cooper, M. Kazmierczyk, M. Gilski, S. Krzywda, H. Zabranska, I. Pichova, J. Thompson, Z. Popovic, M. Jaskolski, D. Baker: Crystal structure of a monomeric retroviral protease solved by protein folding game players, *Nat. Struct. Mol. Biol.* **18**(10), 1175–1177 (2011)
- 5.47 S. Chandrasekharan, N.J. Nersessian, V. Subramanian: Computational modeling: Is this the end of thought experiments in science? In: *Thought Experiments in Philosophy, Science and the Arts*, ed. by J. Brown, M. Frappier, L. Meynell (Routledge, London 2013) pp. 239–260
- 5.48 S. Chandrasekharan: Building to discover: A common coding model, *Cogn. Sci.* **33**(6), 1059–1086 (2009)
- 5.49 N.J. Nersessian: Engineering concepts: The interplay between concept formation and modeling practices in bioengineering sciences, *Mind Cult. Activ.* **19**, 222–239 (2012)
- 5.50 C.G. Langton: Self-reproduction in cellular automata, *Physica D* **10**, 135–144 (1984)
- 5.51 C.G. Langton: Computation at the edge of chaos: Phase transitions and emergent computation, *Physica D* **42**, 12–37 (1990)
- 5.52 C. Reynolds: Flocks, herds, and schools: A distributed behavioral model, *Comp. Graph.* **21**(4), 25–34 (1987)
- 5.53 K. Sims: Evolving 3D morphology and behavior by competition, *Artif. Life* **1**(4), 353–372 (1994)
- 5.54 W. Banzhaf: Self-organization in a system of binary strings. In: *Artificial Life IV*, ed. by R. Brooks, P. Maes (MIT Press, Cambridge MA 2011) pp. 109–119

- 5.55 L. Edwards, Y. Peng, J. Reggia: Computational models for the formation of protocell structure, *Artif. Life* **4**(1), 61–77 (1998)
- 5.56 N.J. Nersessian, E. Kurz-Milcke, W.C. Newstetter, J. Davies: Research laboratories as evolving distributed cognitive systems, *Proc. 25th Annu. Conf. Cogn. Sci. Soc.* (2003) pp. 857–862
- 5.57 L. Osbeck, N.J. Nersessian: The distribution of representation, *J. Theor. Soc. Behav.* **36**, 141–160 (2006)
- 5.58 E. Hutchins: *Cognition in the Wild* (MIT Press, Cambridge 1995)
- 5.59 E. Hutchins: How a cockpit remembers its speeds, *Cogn. Sci.* **19**(3), 265–288 (1995)
- 5.60 E.A. Di Paolo, J. Noble, S. Bullock: Simulation models as opaque thought experiments. In: *Artificial Life VII*, ed. by M.A. Bedau, J.S. McCaskill, N.H. Packard, S. Rasmussen (MIT Press, Cambridge 2000) pp. 497–506
- 5.61 J. Lenhard: When experiments start. Simulation experiments within simulation experiments, *Int. Workshop Thought Exp. Comput. Simul.* (2010)
- 5.62 N.J. Nersessian: In the theoretician's laboratory: Thought experimenting as mental modeling, *Proc. Philos. Assoc. Am.*, Vol. 2 (1992) pp. 291–301
- 5.63 N.J. Nersessian, C. Patton: Model-based reasoning in interdisciplinary engineering. In: *Handbook of the Philosophy of Technology and Engineering Sciences*, ed. by A. Meijers (Elsevier, Amsterdam 2009) pp. 687–718
- 5.64 S. Chandrasekharan: Becoming knowledge: Cognitive and neural mechanisms that support scientific intuition. In: *Rational Intuition: Philosophical Roots, Scientific Investigations*, ed. by L.M. Osbeck, B.S. Held (Cambridge University Press, Cambridge 2014) pp. 307–337

---

# Theoretical **Part B**

## **Part B Theoretical and Cognitive Issues on Abduction and Scientific Inference**

**Ed. by Woosuk Park**

**6 Reorienting the Logic of Abduction**

John Woods, Vancouver, Canada

**7 Patterns of Abductive Inference**

Gerhard Schurz, Dusseldorf, Germany

**8 Forms of Abduction  
and an Inferential Taxonomy**

Gerhard Minnameier, Frankfurt am Main,  
Germany

**9 Magnani's Manipulative Abduction**

Woosuk Park, Daejeon, Korea

In the last century, abduction was extensively studied in logic, semiotics, the philosophy of science, computer science, artificial intelligence, and cognitive science. The surge of interest in abduction derived largely from serious reflection on the neglect of the logic of discovery at the hands of logical positivists and Popper, especially their distinction between the context of discovery and the context of justification. At the same time, the desire to recover the rationality of science that has been seriously challenged by the publication of Kuhn's *The Structure of Scientific Revolutions* might be another important factor. However, the consensus is that researchers have failed to secure the core meaning of abduction, let alone to cover the full range of its applications. The controversial status of abduction can be immediately understood if we consider our inability to answer the following questions satisfactorily:

- What are the differences between abduction and induction?
- What are the differences between abduction and the well-known hypothetico-deductive method?
- What does Peirce mean when he says that abduction is a kind of inference?
- Does abduction involve only the generation of hypotheses or their evaluation as well?
- Are the criteria for the best explanation in abductive reasoning epistemic or pragmatic, or both?
- How many kinds of abduction are there?

Fortunately, the situation has improved much in the last two decades. To say the least, some ambitious attempts to attain a unified overview of abduction have been made, e.g., in Gabbay and Woods (2005), Magnani (2001), and Aliseda (2006). Each of these attempts emphasizes its own strengths and achievements. For example, Aliseda's book represents some logical and computational approaches to abduction quite well. Gabbay and Woods, by introducing the distinction between explanatory/non-explanatory abductions, adopt a broadly logical approach comprehending practical reasoning of real-life logical agents. By introducing his multiple distinctions between different kinds of abduction, i.e., selective/creative, theoretical/manipulative, and sentential/model-based, Magnani (2001, 2009) develops an eco-cognitive view of abduction, according to which instances of abduction are found not only in science and any other human enterprises, but also in animals, bacteria, and brain cells. Part B of this Handbook presents an overview of the most recent research on the foundational and cognitive issues on abduction inspired by all this.

In **Chap. 6** John Woods provides us with the broader context in which the significance of abductive reasoning can be appreciated. He asks whether abduction's epistemic peculiarities can be readily accommodated in philosophy's mainline *theories of knowledge*, and whether abduction provides any reason to question the assumption that the goodness of drawing a conclusion from premises depends on an underlying relation of *logical consequence*. His answer to these questions amounts to a timely response to Hintikka's announcement of abduction as the central problem in contemporary epistemology, as well as the signal of naturalistic turn in logic.

Gerhard Schurz's **Chap. 7** presents a thorough classification of different patterns of abduction. In particular, it attempts the most comprehensive treatment of the patterns of creative abduction, such as *theoretical model abduction*, *common cause abduction*, and *statistical factor analysis*. This is significant, for, compared to selective abductions, creative abductions are rarely discussed, although they are essential in science. By appealing to *independent testability* and *explanatory unification*, a demarcation between scientifically fruitful abductions and speculative abductions is also proposed. Applications of abductive inference in the domains of belief revision and instrumental/technological reasoning represent the author's most recent interest in the border between logic and the philosophy of science.

In **Chap. 8** Gerhard Minnameier, by appropriating all recent studies on abduction, presents a well-rounded overview of the intricate relationships among deduction, induction, and abduction. By taking Peirce's claim seriously that (1) that there are only three kinds of reasoning, i.e. abduction, deduction, and induction, and (2) that these are mutually distinct, he wants to clarify the very notion of abduction. For this purpose, Minnameier carefully examines the fundamental features of the three inferences. He also suggests a novel distinction between two dimensions: i.e., *levels* of abstraction and *domains* of reasoning. To say the least, his taxonomy of inferential reasoning seems to provide us with a nice framework in which different forms of inferences can be systematically accommodated.

Finally, Woosuk Park counts Lorenzo Magnani's discovery of manipulative abduction as one of the most important developments in recent studies on abduction in **Chap. 9**. After briefly introducing Magnani's distinction between theoretical and manipulative abduction, Park discusses how and why Magnani counts diagram-



matic reasoning in geometry as the prime example of manipulative abduction. Among the commentators of Peircean theorematic reasoning, Magnani is unique in equating theorematic reasoning itself as abduction. Park also discusses what he counts as some common characteristics of manipulative abductions, and how and why Magnani views manipulative abduction as a form of practical reasoning. Ultimately, he argues that it is manipulative abduction that enables Magnani to extend abduction to all directions to develop the eco-cognitive model of abduction.

The authors of this part follow the following commonly accepted abbreviation used to refer to the editions of Peirce's work:

- CP: Collected papers: C.S. Peirce: *Reviews, Correspondence, and Bibliography*, Collected Papers of Charles Sanders Peirce, Vol. 8 (Harvard Univ. Press, Cambridge 1958), ed. by A.W. Burks
- NEM: New Elements of Mathematics: C.S. Peirce: *Mathematical Philosophy*, The New Elements of

Mathematics by Charles S. Peirce, Vol. IV (Mouton, The Hague, 1976), ed. by C. Eisele

- MS: manuscript: Peirce manuscript, followed by a number in Richard R. Robin, *Annotated Catalogue of the Papers of Charles S. Peirce* Amherst: University of Massachusetts, 1967.

## References

- D. Gabbay, J. Woods: *A Practical Logic of Cognitive Systems. The Reach of Abduction: Insight and Trial*, Vol. 2 (Elsevier, Amsterdam 2005)
- L. Magnani: *Abduction, Reason, and Science: Processes of Discovery and Explanation* (Kluwer, New York 2001)
- A. Aliseda: *Abductive Reasoning. Logical Investigations into Discovery and Explanation* (Springer, Dordrecht 2006)
- L. Magnani: *Abductive Cognition. The Epistemological and Eco-cognitive Dimensions of Hypothetical Reasoning* (Springer, Berlin, Heidelberg 2009)

# Reorienting

## 6. Reorienting the Logic of Abduction

John Woods

Abduction, still a comparatively neglected kind of premiss-conclusion reasoning, gives rise to the questions I want to consider here. One is whether abduction's epistemic peculiarities can be accommodated happily in the mainline philosophical *theories of knowledge*. The other is whether abduction provides any reason to question the assumption that the goodness of drawing a conclusion from premisses depends on an underlying relation of logical *consequence*. My answer each time is no. I will spend most of my time on the first. Much of what I'll say about the second is a promissory note.

6.1	<b>Abduction</b> .....	138
6.1.1	Peirce's Abduction .....	138
6.1.2	Ignorance Problems .....	138
6.1.3	The Gabbay-Woods Schema .....	139
6.1.4	The Yes-But Phenomenon .....	140
6.2	<b>Knowledge</b> .....	141
6.2.1	Epistemology.....	141
6.2.2	Losing the <i>J</i> -Condition .....	142
6.2.3	The Causal Response Model of Knowledge .....	142
6.2.4	Naturalism .....	143
6.2.5	Showing and Knowing .....	143
6.2.6	Explaining the Yes-Buts .....	144
6.2.7	Guessing .....	144
6.2.8	Closed Worlds.....	146
6.3	<b>Logic</b> .....	148
6.3.1	Consequences and Conclusions .....	148
6.3.2	Semantics .....	148
	<b>References</b> .....	149

Three facts about today's logic stand out:

1. Never has it been done with such technical virtuosity
2. Never has there been so much of it
3. Never has there been so little consensus about its common subject matters.

It would seem that the more we have of it, the less our inclination to get to the bottom of its sprawlingly incompatible provisions. There is nothing remotely like this in real analysis, particle physics or population genetics. There is nothing like it in the premiss-conclusion reasonings of politics and everyday life. Left undealt with, one might see in logic's indifference to its own rivalries some sign of not quite knowing its own mind.

It could be said that one of logic's more stimulating events in our still-young century is the revival of the idea that it is a universal discipline, that when all is said and done there is a core structure to which all the multiplicities of our day are ultimately answerable. If the historical record is anything to go on, the cornerstone

of that core structure is the relation of *logical consequence*. It occasions some sensible operational advice: If in your work you seek to *enlarge* logic's present multiplicities, have the grace to say why you think it qualifies as logic, that is, embodies logic's structural core. This is not idle advice. I hope to give it heed in the pages to follow, as we turn our attention to the logic of abduction.

Although logic's dominant focus has been the consequence relation, in the beginning its centrality owed comparatively little to its *intrinsic* appeal. Consequence was instrumentally interesting; it was thought to be the relation in virtue of which premiss-conclusion reasoning is safe, or whose absence would expose it to risk. Reasoning in turn had an *epistemic* motivation. Man may be many kinds of animal, but heading the list is his cognitive identity. He is a knowledge-seeking and knowledge-attaining being to which his survival and prosperity are indissolubly linked, indispensable to which is his capacity to adjust what he believes to what follows from what. We might say then that as

long as logic has retained its interest in good and bad reasoning it has retained this same epistemic orientation. Accordingly, a logic of good and bad reasoning carries *epistemological* presuppositions that aren't typically explicitly developed.

It would be premature to say that abduction by now has won a central and well-established place in the research programs of modern logic, but there are some hopeful signs of progress (important sources include [6.1–13]). In the literature to date there are two main theoretical approaches, each emphasizing the different sides of a product-process distinction. The logical (or product) approach seeks for truth conditions on abductive consequence relations and of such other properties as may be interdefinable with it. The computational (or process) approach constructs computational models of how hypotheses are selected for use in ab-

ductive contexts. It is not a strict partition. Between the logical and computational paradigms, abductive logic programming and semantic tableaux abduction occupy a more intermediate position. Whatever its precise details, the logic-computer science dichotomy is not something I welcome. It distributes the theory of abductive reasoning into different camps that have yet to learn how to talk to one another in a systematic way. A further difficulty is that whereas abduction is now an identifiable research topic in logic – albeit a minority one – it has yet to attain that status in computer science. Such abductive insights as may occur there are largely in the form of *obiter dicta* attached to the main business at hand (I am indebted to Atocha Aliseda for insightful advice on this point). This leaves us awkwardly positioned. The *foundational* work for a comprehensive account of abductive reasoning still awaits completion.

## 6.1 Abduction

### 6.1.1 Peirce's Abduction

Although there are stirrings of it in Aristotle's notion of *apagogē* [6.14], we owe the modern idea of abduction to Peirce. It is encapsulated in the *Peircean abduction schema*, as follows [6.15, CP 5.189]:

“The surprising fact C is observed.  
But if A were true, C would be a matter of course.  
Hence there is reason to suspect that A is true.”

Peirce's schema raises some obvious questions. One is how central to abduction is the factor of surprise. Another is the issue of how we are to construe the element of suspicion. A third concerns what we are expected to do with propositions that creep thus into our suspicions. A fourth is what we are to make of the idea that an occurrence of something is a matter of course. Like so many of his better ideas and deeper insights, Peirce has nothing like a fully developed account of abduction. Even so, the record contains some important ideas, seven of which I'll mention here:

- P1 Abduction is triggered by surprise [6.15, CP 5.189].
- P2 Abduction is a form of guessing, underwritten innately by instinct ([6.16, p. 128], [6.15, CP 5.171], [6.17, CP 7.220]).
- P3 A successful abduction provides no grounds for believing the abduced proposition to be true [6.16, p. 178].
- P4 Rather than believing them, the proper thing to do with abduced hypotheses is to send them off

to experimental trial ([6.15, CP 5.599], [6.18, CP 6.469–6.473], [6.17, 7.202–219]).

- P5 The connection between the truth of the abduced hypothesis A and the observed fact C is subjunctive [6.15, CP 5.189].
- P6 The inference that the abduction licenses is not to the proposition A, but rather that A's truth is something that might plausibly be suspected [6.15, CP 5.189].
- P7 The *hence* of the Peircean conclusion is ventured defeasibly [6.15, CP 5.189].

Let us note that P3 conveys something of basic importance. It is that successful abductions are *evidentially inert*. They offer no grounds for believing the hypotheses abduced. What, then, is the good of them?

### 6.1.2 Ignorance Problems

Seen in Peirce's way, abductions are responses to ignorance problems. An agent has an ignorance problem in relation to an epistemic target when it can't be attained by the cognitive resources presently at his command, or within easy and timely reach of it. If, for some proposition A, you want to know whether A is the case, and you lack the information to answer this question, or to draw it out by implication or projection from what you currently do know, then you have an ignorance problem with respect to A.

Two of the most common responses to ignorance problems are (1) *subduance* and (2) *surrender*. In the first case, one's ignorance is removed by new knowl-

edge, and an altered position is arrived at, which may serve as a positive basis for new action. In the second case, one's ignorance is fully preserved, and is so in a way that cannot serve as a positive basis for new action (new action is action whose decision to perform is lodged in reasons that would have been afforded by that knowledge). For example, suppose that you've forgotten when Barb's birthday is. If her sister Joan is nearby you can ask her, and then you'll have got what you wanted to know. This is subduance. On the other hand if Joan is traveling incognito in Peru and no one else is about, you might find that knowing Barb's birthday no longer interests you. So you might rescind your epistemic target. This would be surrender.

There is a third response that is sometimes available. It is a response that splits the difference between the prior two. It is abduction. Like surrender, abduction is ignorance-preserving, and like subduance, it offers the agent a positive basis for new action. With subduance, the agent overcomes his ignorance. With surrender, his ignorance overcomes him. With abduction, his ignorance remains, but he is not overcome by it. It offers a reasoned basis for new action in the presence of that ignorance. No one should think that the goal of abduction is to *maintain* that ignorance. The goal is to make the best of the ignorance that one chances to be in.

### 6.1.3 The Gabbay–Woods Schema

The nub of abduction can be described informally. You want to know whether something  $A$  is the case. But you don't know and aren't in a position here and now to get to know. However, you observe that if some further proposition  $H$  were true, then it together with what you already know would enable you to answer your question with regard to  $A$ . Then, on the basis of this subjunctive connection, you infer that  $H$  is a conjecturable hypothesis and, on that basis, you release it provisionally for subsequent inferential work in the relevant contexts.

More formally, let  $T$  be an agent's epistemic target at a time, and  $K$  his knowledge base at that time. Let  $K^*$  be an immediate successor of  $K$  that lies within the agent's means to produce in a timely way. Let  $R$  be an attainment relation for  $T$  and let  $\rightsquigarrow$  denote the subjunctive conditional relation.  $K(H)$  is the revision of  $K$  upon the addition of  $H$ .  $C(H)$  denotes the conjecture of  $H$  and  $H^c$  its activation. Accordingly, the general structure of abduction can be captured by what has come to be known as the Gabbay–Woods schema [6.6, 19, 20]:

1.  $T! E$  [The  $!$  operator sets  $T$  as an epistemic target with respect to some state of affairs  $E$ ]

2.  $\neg R(K, T)$  [fact]
3. Subduance is not presently an option [fact]
4. Surrender is not presently an option [fact]
5.  $H \notin K$  [fact]
6.  $H \notin K^*$  [fact]
7.  $\neg R(H, T)$  [fact]
8.  $\neg R(K(H), T)$  [fact]
9.  $H \rightsquigarrow R(K(H), T)$  [fact]
10.  $H$  meets further conditions  $S_1, \dots, S_n$  [fact]
11. Therefore,  $C(H)$  [sub-conclusion, 1–7]
12. Therefore,  $H^c$  [conclusion, 1–8].

It is easy to see that the distinctive epistemic feature of abduction is captured by the schema. It is a given that  $H$  is not in the agent's knowledge set  $K$ . Nor is it in its immediate successor  $K^*$ . Since  $H$  is not in  $K$ , then the revision of  $K$  by  $H$  is not a knowledge-successor set to  $K$ . Even so,  $H \rightsquigarrow R(K(H), T)$ . But that subjunctive fact is evidentially inert with respect to  $H$ . So the abduction of  $H$  leaves the agent no closer than he was before to achieving the knowledge he sought. Though abductively successful,  $H$  doesn't enable the abducer to attain his epistemic target. So we have it that successful abduction is ignorance-preserving. Of course, the devil is in the details. Specifying the  $S_i$  is perhaps the hardest open problem for abductive logic. In much of the literature it is widely accepted that  $K$ -sets must be consistent and that its consistency must be preserved by  $K(H)$ . This strikes me as unrealistic. Belief sets are often, if not routinely, inconsistent. Also commonly imposed is a minimality condition. There are two inequivalent versions of it. The simplicity version advises that complicated hypotheses should be avoided as much as possible. It is sometimes assumed that truth tends to favor the uncomplicated. I see no reason to accept that. On the other hand, simplicity has a prudential appeal. Simple ideas are more easily understood than complicated ones. But it would be overdoing things to elevate this desideratum to the status of a logically necessary condition. The other version is a form of Quine's maxim of minimum mutilation. It bids the theorist to revise his present theory in the face of new information in ways that leave as much as possible of the now-old theory intact. It advises the revisionist to weigh the benefits of admitting the new information against the costs of undoing the theory's current provisions. This, too, is little more than prudence. No one wants to rule out Planck's introduction of the quantum to physics, never mind the mangling of old physics that ensued. Another of the standard conditions is that  $K(H)$  must entail the proposition for which abductive support has been sought. In some variations inductive implication is substituted. Both I think are too strong. Note also that none of the three – consistency, minimality or implica-

tion – could be thought of as *process* protocols. The  $S_i$  are conditions on hypothesis selection. I have no very clear idea about how this is done, and I cannot but think that my ignorance is widely shared. Small wonder that logicians have wanted to offload the *logic of discovery* to psychology. I will come back to this briefly in due course. Meanwhile let's agree to regard line (10) as a promissory note [6.21, Chap. 11].

#### 6.1.4 The Yes-But Phenomenon

Perhaps it won't come as much of a surprise to learn of the resistance with which the ignorance-preservation claim has been met when the Gabbay–Woods schema has been presented to (what is by now a sizable number of) philosophical audiences. There are those who think that precisely because it strips good abductions of evidential force, the G–W schema misrepresents Peirce. Others think that precisely because it is faithful to Peirce's conditions the G–W schema discredits the Peircean concept of abduction. Of particular interest is the hesitation shown by philosophers who are actually inclined to *accept* the schema, and *accept* the Peircean notion. It may be true, they seem to think, that abduction is ignorance-preserving, but it is not a truth to which they take kindly. Something about it they find unsatisfying. There is a conventional way of giving voice to this kind of reticence. One does it with the words, *Yes, but . . .* So we may speak of this class of resisters as the ignorance-preservation *yes-but*s.

Some philosophers are of the view that there are at least three grades of evidential strength. There is evidential strength of the truth-preserving sort; evidential strength of the probability-enhancing sort; and evidential strength of a weaker kind. This latter incorporates a notion of evidence that is strong in its way without being either deductively strong or inductively strong. It is, as we might say, induction's poor cousin. Proponents of this approach are faced with an interesting challenge. They must try to tell us what it is for premisses nondeductively to favor a conclusion for which there is no strong inductive support. If the weak cousin thesis is false, lots of philosophers are nevertheless drawn to it. So perhaps the better explanation of the *yes-but*s' resistance to the ignorance-preservation claim is that they think that it *overstates* the poor cousin thesis, that it makes of abduction a poorer thing than it actually is. The poor cousin thesis says that abduction is the weakest evidential relation of the family. But the ignorance-preservation thesis says that it is an evidential relation of no kind, no matter how weak. Accordingly, what the *yes-but*s are proposing is tantamount to retention of the G–W schema for abduction *minus* Peirce's clause P3. This would allow success-

fully abduced hypotheses the promise of *poor-cousin* evidential backing; but it wouldn't be backing with no evidential force. It is an attractive idea, but it cuts too far.

There are too many cases in which successful reasoning, indeed brilliant reasoning, has the very characteristic the reformers would wish to suppress. A case in point is Planck's quantum hypothesis. In the physics of 1900s, black body radiation lacked unifying laws for high and low frequencies. Planck was disturbed by this. Notwithstanding his lengthy acquaintanceship with it, the disunification of the black body laws was a surprising event. It was, for physics, not a matter of course. Planck wanted to know what it would take to ease his cognitive irritation. Nothing he knew about physics answered this question. Nothing he would come to know about physics would answer it either, as long as physics was done in the standard way. Planck recognized that he would never attain his target until physics were done in a new way, in a way sufficiently at odds with the present paradigm to get some movement on this question; yet not so excessively ajar from it as to make it unrecognizable as physics. That day in 1900 when he announced to his son that he had overturned Newton, Planck was drawn to the conditional that if the quantum hypothesis  $Q$  were true then  $K(Q)$  – that is, physics as revised by the incorporation of  $Q$  – would enable him to reach his target. So he put it to work accordingly. At no stage did Planck think that  $Q$  was true. He thought it lacked physical meaning. He thought that his reasoning provided no evidence that  $Q$  was true and no grounds for believing it to be true. Peirce wanted a logic that respected this kind of thinking. This is what I want too. The poor cousin thesis doesn't do this, and cannot.

Ignorance removal is prompted by the reasoner's desire to know something he doesn't now know, or to have more knowledge of it than he currently does. What are the conditions under which this happens? It seems right to say that without an appreciation of the general conditions under which a human reasoner is in a state of knowledge, this is a question without a principled answer. If, as I aver, there are abductive modes of reasoning prompted by the desire to improve one's epistemic condition which, even when wholly successful, do not fulfill that objective, there must be two particular considerations thanks to which this is so. One would have to do with abduction. The other has to do with knowledge. A fair part of this first factor is captured by the Gabbay–Woods schema (or so I say). The second is catered for by the right theory of knowledge, if there is one. We asked why, if a philosopher accepted the Gabbay–Woods schema for abduction, would he dislike its commitment to the ignorance-preservation claim? The possibility that we're now positioned to con-

sider is that his yes-but hesitancy flows from how he approaches the general question of knowledge. That is to say, it is his *epistemology* that makes him nervous,

not his *logic*. If so, the *yes* part of *yes, but ...* is directed to the logic, but the *but part* is directed to the epistemology.

## 6.2 Knowledge

### 6.2.1 Epistemology

I said in the abstract that epistemological considerations affecting the goodness or badness of premiss-conclusion reasoning are little in evidence in mainstream logic. In so saying, I intend no slight to the now large growing and prospering literature on epistemic logics [6.22–24]. For the most part these logics construct formal representations of the standard repertoire of properties – consequence, validity, derivability, consistency, and so on – defined for sentences to which symbols for *it is known that*, and *it is believed that* function as sentence operators. A central task for these logics is to construct a formal semantics for such sentences, typically on the assumption that these epistemic expressions are modal operators, hence subject to a possible worlds treatment. Notwithstanding their explicitly epistemic orientation, it remains true that there is in this literature virtually no express contact with any of the going epistemologies. So here, too, if they operate at all epistemological considerations operate tacitly as part of an unrevealed epistemological background information. I intend something different here. I want to bring epistemology to the fore, which is precisely where it belongs in logics of premiss-conclusion reasoning of all kinds.

I want also to move on to what I think may be the right explanation of the yes-but's dissatisfactions. Before getting started, a caveat of some importance should be flagged. The explanation I'm about to proffer attributes to the yes-but's an epistemological perspective that hardly anyone shares; I mean by this hardly any epistemologist shares, a notable exception is [6.25]. There is a good chance that whatever its intrinsic plausibility, this new explanation will lack for takers. Even so, for reasons that will appear, I want to persist with it for awhile. Here is what it proposes.

#### The Right-Wrong Thesis

While the Gabbay–Woods schema gets something right about abduction, it nevertheless gets ignorance-preservation *wrong*. What it gets right is that good abductions are evidentially inert. What it gets wrong is that this lack of evidential heft entails a corresponding failure to lift the abducer in any degree from his present ignorance.

#### Corollary 6.1

There are abductive contexts in which knowledge can be attained in the absence of evidence.

The idea of knowledge without supporting evidence isn't entirely new or in the least shocking. There is a deeply dug-in inclination to apply this characterization to quite large classes of cases. Roughly, these are the propositions knowledge of which is a priori or independent of experience; or, as with Aristotle's first principles, are known without the necessity or even possibility of demonstration; or, as some insist, are the immediate disclosures of sense and introspection. Disagreements have arisen, and still do, about whether these specifications are accurate or sustainable, but it would be a considerable exaggeration to call this sort of evidential indifference shocking, and wildly inaccurate as a matter of historical fact to think of it as new.

In truth, apriorism is beside the point of the right-wrong thesis and its corollary. The knowledge that falls within their intended ambit is our knowledge of contingent propositions, whether of the empirical sciences or of the common experience of life. The right-wrong claim is that there are contingent propositions about the world which, without being in any way *epistemically privileged*, can be ignorance-reducing by virtue of considerations that lend them no evidential weight. So what is wanted is a theory of knowledge that allows this to happen.

The historically dominant idea in philosophy is that knowledge is true belief plus some other condition, usually identified as justification or evidence. This, the *J*-condition, has been with us at least since Plato's *Theaetetus*, and much scholarly ink has been spilled over how it is best formulated and whether it might require the corrective touch of some further condition. But, as a general idea, the establishment bona fides of the *J*-condition are as rock solid as anything in philosophy.

The account of knowledge I am looking for arises at the juncture of two epistemological developments. One is the trend towards naturalism [6.26] and the other is the arrival of reliabilism [6.27]. It is a theory in which the *J*-condition fails as a general constraint on epistemically unprivileged contingent knowledge. Accordingly, my first task is to try to downgrade the condition, to deny it a defining role. Assuming some success with the

first, my second task will be to find at the intersection of these trends an epistemological orientation – perhaps I would better call it an epistemological *sensibility* – which might without too much strain be reconciled to the loss of the *J*-condition. For ease of reference let me baptize this orientation, this sensibility, the *causal response turn*.

Whereupon task number three, which is to identify those further features of the causal response model that link up the notions of evidence and knowledge in the heterodox ways demanded by the right-wrong thesis.

### 6.2.2 Losing the *J*-Condition

The *J*-condition has attracted huge literature and underwritten a good deal of strategic equivocation. On *engaged* readings of the condition, a person's belief is justified or evidenced only if he himself has produced his justification then and there, or he has presented the evidence for it on the spot. On *disengaged* readings, a person is justified in believing if a justification exists but hasn't been invoked, or evidence exists but hasn't been adduced or even perhaps found. The engaged and disengaged readings raise an interesting question. How deeply engaged does one have to be to meet the *J*-condition on knowledge? Most epistemologists formulate the engaged-disengaged distinction as one between internalist and externalist justification.

Engagement here is a matter of *case making*. The two readings of *J* define a spectrum, but for present purposes there is little that needs saying of what lies within. It suffices to note that in its most engaged sense a belief is justified or evidenced only if the believer can himself make the case for it here and now. At the other extreme, the belief is justified or evidenced if a case for it is available in principle to someone or other. In the first case, the individual in question has a high degree of case-making engagement. In the other, his engagement is a gestural, anonymous and proxied one: it is engagement in name only.

Suppose the following were true. Suppose that, for every piece of epistemically unprivileged contingent knowledge *p*, there were a structure of facts in virtue of which *p* is the case. Suppose further that for every such *p* a person knows, it would be possible in principle to discern this structure of the facts and the in-virtue-of-relation it bears to *p*'s truth. (I don't think there is any realistic chance of this being so, but let's assume it for the point at hand.) Suppose, finally, that we agreed to say that when in principle knowledge of that structure and that relation exists with respect to a *p* that a subject *S* knows, there exists a justification of *S*'s belief that *p*. For ease of reference, let's call these *factive* justifications. Factive justifications are justifications at their most

disengaged. They stand in radical contrast to highly engaged justifications, which we may call *forensic*.

By construction of the case presently in view, factive justification will be the constant companion of any piece of epistemically unprivileged contingent knowledge that *S* chances to have. But we have in this constancy not conditionhood but concomitance. Factive justification is a faithful accompaniment of such knowledge, but it is not a constituent of it. Forensic justification is another story. We might grant that if, when *S* knows that *p*, he has a forensic justification for his belief, then his justification will have made a contribution to this knowledge. But in relation to all that *S* knows it is comparatively rare that there is a forensic justification. Here is a test case, with a tip of the hat to Peirce: Do you know who your parents are? Of course you do! Very well, then, let's have your forensic justification.

This is troublesome. If we persist in making forensic justification a condition on knowledge, the result is skepticism on an undesirable scale. If, on the other hand, we decide to go with factive justification, then justifications exist whenever knowledge exists, but they aren't conditions on this knowledge. They are not a structural element of it. Whereupon we are met with the *J*-condition dilemma.

#### *J*-Condition Dilemma

Depending on how it is read, the *J*-condition is either an irrelevant concomitant of knowledge, or a skepticism-inducing discouragement of it.

The forensic-factive ambiguity runs through all the idioms of *J*-attribution. Concerning his belief that *p* there might be evidence for *p* that *S* adduces or there may be evidence for *p* that exists without attribution. There may be reasons for it that *S* gives, or reasons for it that exist without being given. Like confusions repose in careless uses of *have*. If we allow that *S* has a justification or has evidence or has reasons whenever these things exist factively, we mislicense the inference from the factive to the forensic, allowing, in so doing, *S* to have justifications that he's never heard of.

### 6.2.3 The Causal Response Model of Knowledge

The causal response (CR) model of knowledge is rightly associated with reliabilism. In all the going forms of it, the *J*-condition is preserved [6.28]. One of the few places in the reliabilist literature where we see stirrings of the pure version of the causal model is Alvin Goldman's first reliabilist paper, which appeared in 1967. It is a rare place in Goldman's foundational corpus where the *J*-condition, if there at all, is given shortest shrift. In some versions, the *J*-condition is

satisfied when one's belief has been reached by reliable *procedures*. In others, the condition is met when the belief was reliably produced, that is, produced by belief-forming *mechanisms* that were working reliably. In contrast to the standard versions, the *pure* version is one in which the *J*-condition is eliminated, rather than reinterpreted along reliabilist lines. As a first approximation, the pure theory characterizes knowledge as follows:

“*S* knows that if and only if *p* is true, *S* believes that, the belief was produced by belief-forming devices, in good working order, operating as they should on good information and in the absence of Gettier nuisances and other hostile externalities.”

Fundamental to what I've been calling the pure theory is the conviction that knowledge is not in any essential or general way tied to case making, that *knowing* is one thing and *showing* another. This is not to say that case making is never implicated in knowledge. There are lots of beliefs that would not have been had in the absence of the case makings that triggered their formation. Think here of a mother's sad realization that her son is guilty of the crime after all, or a nineteenth century mathematician's grudging acknowledgment of the transfinite. But as a general constraint, case making is rejected by pure causalists; by causalists of the sort that Goldman was trying to be in 1967.

### 6.2.4 Naturalism

Epistemology's naturalized turn supplies a welcoming habitat for the CR model. Naturalism comes in various and competing versions, but at the core of them all is the insistence that human knowledge is a natural phenomenon, achieved by natural beings in accordance with their design and wherewithal, interacting in the causal nexi in which the human organism lives out his life. Unlike the *J* theorist, the CR theorist is a respecter of the passive side of knowledge. He knows that there are large classes of cases in which achieving a knowledge of something is a little more than just being awake and on the scene. Even where some initiative is required by the knower, the resultant knowledge is always a partnership between doing and being done to. So even worked-for knowledge is partly down to *him* and partly down to his *devices*.

It would be wrong to leave the impression that, on the CR model, knowing things is just a matter of doing what comes naturally. There are ranges of cases in which knowledge is extremely difficult to get, if gettable at all. There are cases in which knowledge is unattainable except for the intelligence, skill, training and expertise of those who seek it. Everyone has an

aptitude for knowledge. But there are cases galore in which aptitude requires the supplementation of vocation and talent – and training. CR theorists are no less aware of this than their *J* rivals. The difference between them falls in where the emphasis falls. Among *J* theorists there is a tendency to generalize the hard cases. Among CR theorists there is a contrary tendency to keep the hard cases in their place.

Let me say again that *J*-theories give an exaggerated, if equivocal, place to the role of showing in knowing. Contrary to what might be supposed, the CR model is no disrespecter of the showing-knowing distinction, albeit with a more circumscribed appreciation of showing. I want to turn to this now.

### 6.2.5 Showing and Knowing

Consider the case of Fermat's Last Theorem. The theorem asserts that for integers  $x$ ,  $y$ , and  $z$ , the equation  $x^n + y^n = z^n$  lacks a solution when  $n > 2$ . Fermat famously left a marginal note claiming to have found a proof of his theorem. I want to simplify the example by stipulating that he did not have a proof and did not think or say that he did. The received wisdom is that Fermat went to his grave not knowing that his theorem is true. The received wisdom is that no one knew whether the theorem is true until Andrew Wiles' proof of it in 1995. If the forensically conceived *J* model were true, this would be pretty much the way we would expect the received wisdom to go.

If the *J* model is hard on knowledge, the CR model is a good deal more accommodating. It gives to knowledge a generous provenance. But I daresay that it will come as a surprise that, on some perfectly plausible assumptions, Fermat did indeed know the truth of his theorem, never mind (as we have stipulated) that he was all at sea about its proof. Fermat was no rookie. He was a gifted and experienced mathematician. He was immersed in a sea of mathematical sophistication. He was a mathematical virtuoso. Fermat knew his theorem if the following conditions were met: It is true (as indeed it is), he believed it (as indeed he did), his highly trained belief-forming devices were in good order (as indeed they were) and not in this instance misperforming (as indeed they were not), and their operations were not compromised by bad information or Gettier nuisances (as indeed was the case). So Fermat and generations of others like-placed knew the theorem well before its proof could be contrived.

We come now to a related point about showing and knowing. Showing and knowing mark two distinct goals for science, and a corresponding difference in their satisfaction conditions. Not unlike the law, science is in significant measure a case-making profession –



a forensic profession – made so by the premium it places on *knowing* when knowledge has been achieved, rather than just achieving it. This has something to do with its status as a profession, subject to its own exacting requirements for apprenticeship, advancement and successful practice. These are factors that impose on people in the showing professions expectations that regulate *public announcement*. Fermat may well have known the truth of his theorem and may have had occasion to say so to a trusted friend or his mother. But he was not to say it for publication. Publication is a vehicle for case making, and case making is harder than knowing. Journal editors don't give a toss for what you know. But they might sit up and notice if you can show what you know.

### 6.2.6 Explaining the Yes-Buts

The ignorance-preservation claim is rooted in the idea of the no evidence-no knowledge thesis.

#### The No Evidence–No Knowledge Thesis

Since successful abduction is evidentially inert, it is also epistemically inert. But this is justificationism: No advance in knowledge without some corresponding advance in evidence.

The CR model jettisons justificationism. It denies the very implication in which the ignorance-preservation thesis is grounded. It is not hard to see that the evidence, whose abductive absence Peirce seizes upon, is not evidence in the factive sense. Peirce insists that we have no business believing a successfully abducted hypothesis. Peirce certainly doesn't deny that behind any plausibly conjectured hypothesis there is a structure of facts in virtue to which it owes its truth value. Peirce thinks that our track record as abductive guessers is remarkably good. He is struck by the ratio of right guesses to guesses. He is struck by our aptitude for correcting wrong guesses. The evidence whose absence matters here is forensic, it is evidence by which an abducer could vindicate his belief in the hypothesis at hand. But Peirce thinks that in the abductive context nothing vindicates that belief.

We come now to a critical observation. There is nothing in Peirce's account that tells us that abducted hypotheses aren't believed as a *matter of fact*. Some clearly are not. At the time of their respective advancements, Planck didn't believe the quantum hypothesis and Gell-Mann didn't believe the quark hypothesis. But it takes no more than simple inspection to see that there are masses of cases to the contrary, that abductive success is belief-*inducing* on a large scale.

There is in this commonplace fact something for the CR theories to make something of. Let  $H$  be one

of those successfully abducted hypotheses that happen to be true and, contrary to Peirce's advice, believed by its abducer  $S$ . What would it take to get us seriously to propose that, when these conditions are met,  $S$ 's belief-forming device's are malfunctioning or are in poor operating order. Notice that a commonly held answer is not available here, on pain of question begging. It cannot be said that unevidenced belief is itself evidence of malfunction and disorder. That is, it cannot be said to the CR-theorist, since implicit in his rejection of justificationism is his rejection of this answer.

Is there, then, any reason to suppose that the arousal of unevidenced belief might be some indication of *properly* functioning belief formation? Ironically enough, there is an affirmative answer in Peirce himself. Peirce is much taken with our capacity for right guessing. Our facility with guessing is so impressive that Peirce is driven to the idea that good guessing is something the human animal is built for. But if we are built for good guessing, and good abduction is a form of guessing, how can the abduction of true hypotheses not be likewise something we're built for? Accordingly, there is a case for saying that.

#### Knowledge Enhancement

In the CR model of knowledge, there are numbers of instances in which successful abduction is not only not ignorance preserving, but actually *knowledge enhancing*.

Part of what makes for the irony of Peirce's enthusiasm for right guessing is his insistence that guesses not be indulged by belief. In this he is a justificationist. Abducers have no business in believing unevidenced propositions, never mind their abductive allure. This is enough of a basis to pin the ignorance-preservation thesis on Peirce, but *not* on a CR theorist who accepts the Gabbay–Woods schema. What this shows is that theirs is not a disagreement about abduction. It is a disagreement about knowledge.

There isn't much likelihood that yes-buts will flock to this accommodation. The reason is that hardly anyone (any philosopher anyway) thinks the CR model is true in its pure form. There is no space left to me to debate the ins and outs of this. Suffice it to say that it offers the abductive logician the very relief that the yes-buts pine for. Besides, the CR theory just might be true [6.21].

### 6.2.7 Guessing

In line (10) of the G–W schema the  $S_i$  occur as placeholders for conditions on hypothesis selection. Previously, I said that I didn't know what these conditions

are [6.7]. In point of fact there are two things that I don't know. One is the normative conditions in virtue of which the selection made is a worthy choice. The other is the causal conditions that enable the choice to be made. It is easy to see that there are a good many  $H$ s that could serve as antecedents in line (9)'s  $H \leftrightarrow R(K(H), T)$  without disturbing its truth value. It is also easy to see that a good many of *those*  $H$ s would never be abductively concluded, never mind their occurrence there. It is clear that a reasonable choice of  $H$  must preserve the truth of (9). It is also clear that this is not enough for abductive significance. A reasonable choice must have some further features. I am especially at a loss to describe how beings like us actually go about finding things like that. Perhaps it will be said that my difficulty is a reflection on me, not on the criteria for hypothesis selection. It is true that the number of propositions that could be entertained is at least as large as the number of  $H$ s that slot into the antecedent of (9) in a truth-preserving way. Let's think of these as constituting the hypothesis-selection space. Selection, in turn, is a matter of cutting down this large space to a much smaller proper subset, ideally a unit set. Selection, to this same effect, would be achieved by a search engine operating on the hypothesis-selection space. Its purpose would be to pluck from that multiplicity the one, or few ones, that would serve our purposes.

There is nothing remotely mystifying or opaque about search engines (why else would we bother with Google?). So isn't the problem I'm having with the  $S_i$  that I'm not a software engineer? Wouldn't it be prudent to outsource the hypothesis-selection task to someone equipped to perform it? To which I say: If that is a doable thing we should do it. There is no doubt that algorithms exist in exuberant abundance for search tasks of considerable variety and complexity. There are algorithms that cut down a computer system's search space to one answering to the algorithm's flags. Perhaps such an arrangement could be said to model hypothesis selection. But it is another thing entirely as to whether, when we ourselves are performing them, our hypothesis selections implement the system's algorithms. So I am minded to say that *my* questions about the  $S_i$  are not comprehensively answerable by a software engineer.

Here is where guessing re-enters the picture, which is what Peirce thinks that hypothesis selection is. Peirce is struck by how good we are at it. By this he needn't have meant that we have more correct guesses than incorrect. It is enough that, even if we make fewer correct guesses than incorrect, the ratio of correct to incorrect is still impressively high. We get it right, rather than wrong, with a notable frequency. Our opportunities for getting it wrong are enormous. Relative to the propositions that could have been guessed at, the number of

times that they are rightly guessed is amazing; so much so that Peirce is led to surmise that our proclivity for right guesses is innate. Of course, not all good guessing is accurate. A good guess can be one that puts the guessed-at proposition in *the ball park*, notwithstanding that it might actually not be true. Here, too, good guesses might include more incorrect ones than correct. But as before, the ratio of correct to merely good could be notably high. So the safer claim on Peirce's behalf is that beings like us are hardwired to make for good, although not necessarily correct, guesses with a very high frequency. It is lots easier to make a ball-park guess than a true one; so much so that the hesitant nativist might claim a hardwired proclivity for ball-park, yet not for truth, save as a welcome contingency, which in its own turn presents itself with an agreeable frequency. Thus the safe inference to draw from the fact that  $H$  was selected is that  $H$  is in the ball park. The inference to  $H$ 's truth is not dismissable, but it is weaker.

Needless to say, nativism has problems all its own. But what I want to concentrate on is a problem it poses for Peircian abduction. At the heart of all is what to make of ball-park guesses. The safest thing is to propose is that, even when false, a ball-park hypothesis in a given context is one that bears serious operational consideration there. There might be two overarching reasons for this. *One* is that ball-park hypotheses show promise of having a coherently manageable role in the conceptual spaces of the contexts of their engagement. Take again the example of Planck. The quantum hypothesis was a big wrench to classical physics. It didn't then have an established scientific meaning. It entered the fray without any trace of a track record. Even so, for all its foreignness, it was a ball-park hypothesis. What made it so was that  $P(Q)$  was a theory revision recognizable as *physics*. Contrast  $Q$  with *the gold fairy will achieve the sought-for unification*. Of course, all of this turns on the assumption that Peirce got it right in thinking that hypothesis selection is guessing, and to note that good guessing is innate. Call this the *innateness hypothesis*. The *second* consideration is that the frequency of true hypotheses to ball-park hypotheses is notably high.

Whether he (expressly) knows how it's done, when an abductive agent is going through his paces, there is a point at which he selects a hypothesis  $H$ . If the innateness thesis holds, then the agent has introduced a proposition that has an excellent shot at being ball-park, and a decent shot of being true. On all approaches to the matter, an abduction won't have been performed in the absence of  $H$ ; and on the G-W approach, it won't have been performed correctly unless  $H$  is neither believed nor (however weakly) evidenced by its own abductive success. On the other, our present reflections suggest

that the very fact that  $H$  was *selected* is evidence that it is ball-park, and less good but not nonexistent evidence that it is true. Moreover,  $H$  is the antecedent of our subjunctive conditional (9)  $H \rightsquigarrow R(K(H), T)$ . If  $H$  is true so is  $R(K(H), T)$  by modus ponens; and if  $R(K(H), T)$  holds the original ignorance problem is solved by a form of subduance. In which case, the abduction simply *lapses*. It lapses because the nonevidential weight lent to a successfully *abducted* hypothesis is, on the G–W model, weaker than the evidential support given it by way of the innateness hypothesis as regards its very *selection*.

If, on the other hand,  $H$  is not true, but ball-park – hence favorably evidenced – and being evidenced is closed under consequence, then the reasoning at hand also goes through under the obvious adjustments.

The problem is that there are two matters on which Peirce can't have his cake and eat it too. If he retains the innateness thesis he can't have the ignorance-preservation thesis. Equally, if he keeps ignorance preservation he must give up innateness, which *nota bene* is not the thesis that guessing is innate but that *good* guessing is innate. Yet if we give up innateness we're back to where we started, with less than we would like to say about the actual conditions for which the G–W  $S_i$  are mere placeholders. I leave the innateness-ignorance preservation clash as an open problem in the abduction research program. Since, by our earlier reasoning, there is an epistemology (CR) that retains ignorance preservation only as a contingent property of some abductions, my present uncertain inclination is to retain G–W as modified by CR and to rethink innateness. But I'm open to offers. I'll get back to this briefly in the section to follow.

Having had my say about the epistemological considerations that play out in the logic of abduction, I want to turn to the question of how, or to what extent, a logic of abduction will meet universalist conditions on logic. I want to determine whether or to what extent abductive theories embody the structural core assumed by universalists to be common to any theory that qualifies for admittance to the province of logic.

Whatever the details, abduction is a form of premiss-conclusion reasoning. By the conclusions-consequence thesis, whenever the reasoning is good the conclusion that's drawn is a consequence of those premisses. As logics have proliferated, so too the consequences, albeit not exactly in a strict one-to-one correspondence. If today there are more logics than one can shake a stick at, there is a concomitant plenitude of consequences relations. Much of what preoccupies logicians presently is the classification, individuation, and interrelatedness of this multiplicity. Whatever their variations, there is one distinction to which they all an-

swer. Some consequence relations are truth-preserving; all the others aren't. Truth-preserving consequence is (said to be) monotonic. (It isn't. To take an ancient example, Aristotle's *sylogistic* consequence is truth-preserving but nonmonotonic.) Premises from which a conclusion follows can be supplemented at will and the conclusion will still follow. One way of capturing this point is that truth-preserving consequence is impervious to the openness of the world. As far as consequencehood is concerned, the world might as well be closed. Once a consequence of something, always a consequence of it. It is strikingly otherwise with non-truth-preserving consequence. It is precisely this indifference to the openness of the world that is lost.

### 6.2.8 Closed Worlds

When we were discussing the  $J$  condition on knowledge, we called upon a distinction between the factive justification of a belief and its forensic justification. In a rough and ready way, a factive justification is down to the *world*, whereas a forensic justification is down to *us*. We find ourselves at a point at which the idea of factivity might be put to further good use. To see how, it is necessary to acknowledge that the distinction between open and closed worlds is systematically ambiguous. In one sense it marks a contrast between *information states* at a time, with the closed world being the state of total information, and open ones states of incomplete information. In the other sense, a closed world can be called factive. A closed world at  $t$  is everything that is the case at  $t$ . It is the totality of facts at  $t$ . A closed world is also open at  $t$ , not with regard to the facts that close it at  $t$ , but in respect of the facts thence to come. We may suppose that the world will cease to open at the crack of doom, and that the complete inventory of all the facts that ever were would be logged in the right sort of Domsday Book. It is not, of course, a book that any of us will get to read. Like it or not, we must make do with openness. Both our information states and the world are open at any  $t$  before the crack. But the diachronics of facticity outpace the accuracy of information states. When there is a clash, the world at  $t$  always trumps our information about it at  $t-n$ .

At any given time the world will be more closed than its concurrent information states. At any given time the state of the world outreaches the state of our knowledge of it. When we reason from premisses to conclusions we are not negotiating with the world. We are negotiating with informational reflections of the world. We are negotiating with information states. Given the limitations on human information states, our representations of the world are in virtually all respects

open, and most premises-conclusion relations are susceptible to rupture. Truth-preserving consequences are an interesting exception. The world can be as open as openness gets, but a truth-preserving consequence of something is always a consequence of it, never mind the provisions at any  $t$  of our information states. Nonmonotonic consequence is different: Today a consequence tomorrow a nonconsequence.

We might think that the more prudent course is to cease drawing conclusions and postpone the decisions they induce us to make until our information state closes, until our information is permanently total. The ludicrousness of the assumption speaks for itself. Cognitive and behavioral paralysis is not an evolutionary option. Thus arises the closed world assumption. Given that belief and action cannot await the arrival of total information, it behooves us to draw our conclusions and take our decisions when the likelihood of informational defeat is least high, at which point we would invoke the assumption that for the matter at hand the world might just as well be closed.

The key question about the closed world assumption is the set of conditions under which it is reasonable to invoke it. The follow-up question is whether we're much good at it. I am not much inclined to think that we have done all that well in answering the first question. But my answer to the second is that, given the plenitude of times and circumstances at which to invoke it, our track record is really quite good; certainly good enough to keep humanity's knowledge-seeking project briskly up and running. Even so, the closed world assumption is vulnerable to two occasions of defeat. One is by way of later information about later facts. Another is by way of later information about the facts now in play. It is easy to see, and no surprise at all, that new facts will overturn present information about present facts with a frequency that matches the frequency of the world's own displacement of old facts by new. Less easy to see is how we manage as well as we do in invoking closure in the absence of information about the present destructive facts currently beyond our ken. Here, too, we have a cut-down problem. We call upon closure in the hopeful expectation that no present unannounced fact will undo the conclusions we now draw and the decisions they induce us to make. Comparatively speaking, virtually all the facts there are now are facts that no one will ever know. That's quite a lot of facts, indeed it is nondenumerably many (for isn't it a fact that, for any real number, it is a number, and is self-identical, and so on?).

There is a point of similarity between hypothesis selection and the imposition of world closure. Our good

track record with both invites a nativist account each time. Oversimplified, we are as good as we are at selecting hypotheses because that's the way we were built. We are as good as we are at closing the world because that too is the way we were built. I suggested earlier that in abductive contexts the very fact that  $H$  has been selected is some evidence that it is true (and even better evidence that it is ball-park). But this seems to contradict the Peircian thesis that abductive success confers on  $H$  nothing stronger than the suspicion that it might be true. Since Peirce's account of abduction incorporates both the innateness thesis and the no-evidential-support thesis, it would appear that Peirce's account is internally inconsistent. I said a section ago that I had a slight leaning for retaining the no-evidence thesis and lightening up on the innateness thesis. Either way is Hobson's choice. That, anyhow, is how it appears.

In fact, however, the appearance is deceptive. There is no contradiction. Peirce does not make it a condition on abductive hypothesis-selection that  $H$  enter the fray entirely untouched by reasons to believe it or evidence that supports it. He requires that the present support-status of  $H$  has no role to play in the abductive process. That  $H$  is somewhat well supported doesn't, if true, have any premissory role here. Moreover, it is not the goal of abduction to make any kind of case for  $H$ 's truth. The goal is to find an  $H$  which, independently of its own epistemic status, would if true enable a reasoner to hit his target  $T$ . But whatever the target is, it's not the target of wanting to know whether  $H$  is true. It is true that, if all goes well, Peirce asserts that it may be defeasibly concluded that there is reason to suspect that  $H$  might be true. But, again, abduction's purpose is not to make a case for  $H$ , no matter how weakly. The function of the suspectability observation is wholly retrospective. It serves as a hypothesis-selection vindicator. You've picked the (or a) right hypothesis only if the true subjunctive conditional in which it appears as antecedent occasions the abducer's satisfaction that that, in and of itself, would make it reasonable to suspect that  $H$  might be so. In a way, then, the G-W schema misrepresents this connection. It is not that the abduction implies  $H$ 's suspectability, but rather that the abduction won't succeed unless the truth of line (9) induces the suspectability belief [6.21] (for more on the causal role in inference, readers could again consult [6.21]). And that won't happen if the wrong  $H$  has been selected, never mind that it preserves (9)'s truth. For the point at hand, however, we've arrived at a good result. The innateness thesis and the no-support thesis are both implicated in the Peircean construal of abduction, but are in perfect consistency.

## 6.3 Logic

### 6.3.1 Consequences and Conclusions

I said at the beginning that for nearly two and a half millennia the central focus of logic has been the consequence relation. More basic still was a concomitant preoccupation with premiss-conclusion reasoning. For a very long time logicians took it as given that these two matters are joined at the hip.

#### Conclusions and Consequences

When someone correctly draws a conclusion from some premisses, his conclusion is a consequence of them.

#### Corollary 6.2

If a conclusion drawn from some premisses is not a consequence of them, then the conclusion is incorrectly drawn.

If this were so, it could be seen at once that there is a quite intuitive distinction between the consequences that a premiss set has and the consequences that a reasonable reasoner would *conclude* from it. Any treatment of logic in which this distinction is at least implicitly present, there is a principled role for agents, for the very beings who draw what conclusions they will from the consequences that flow from the premisses at hand. In any such logic there will be at least implicit provision for the nature of the agent's involvement. In every case the involvement is epistemically oriented. People want to know what follows from what. They want to know how to rebut an opponent. They want to know whether, when this follows from that that, they can now be said to know that. In a helpful simplification, it could be said that logic got out of the agency business in 1879. It is not that agency was overlooked entirely, but rather that it was scandalously short-sheeted. For consequence, the having-drawing distinction would fold into having; and having, it would be said, would be the very things drawn by an ideally rational reasoner. Of course, this downplaying of cognitive agency was never without its dissenters. Indeed today we are awash in game theoretic exuberance, to name just one development of note.

### 6.3.2 Semantics

Consequence derives its semantic character from its attachment to truth, itself a semantic property in an odd baptismal bestowal by Tarski. In the deductive case, it is easy to see how truth is implicated in consequence and how, in turn, consequence assumes its status as a semantic relation. Not only does truth ground the very definition of consequence, but it makes for a re-

lation that is also truth-preserving. The monotonicity of consequence provides the sole instance in which a consequence is impervious to the informational openness of the world. It is the one case in informational openness at  $t$  that is indifferent to the world's factive closure at  $t$ , to say nothing of its final closure at the crack of doom. It has long been known that logicians, then and now, harbor an inordinate affection for deductive consequence. It's not hard to see why. Deductive consequence has proved more responsive to theoretical treatment than any of the nondeductive variety. But more centrally, it is the only consequence relation that captures permanent chunks of facticity.

Whatever else we might say, we can't say that nonmonotonic relations are relations of *semantic* consequence. If  $B$  is a nonmonotonic consequence of  $A$  it holds independently of whatever makes for the *truth* of  $A$  and  $B$ . Sometimes perhaps it holds on account of probability conditions on  $A$  and  $B$ , but probability has nothing to do with truth. If there is such a thing as probabilistic consequence – think here of Carnap's partial entailment – it is not a semantic relation. We may have it now that the evidence strongly supports the charge against Spike in last night's burglary. We might come to know better tomorrow. We might learn that at the time of the offense Spike was spotted on the other side of town. So the world at  $t$  didn't support then the proposition that Spike did do it, never mind the state of information the day after  $t$ .

No one doubts that yesterday there existed between the evidence on hand and the charge against Spike a relation of epistemic and decisional importance, a kind of relation in whose absence a survivable human life would be impossible. But a fair question nevertheless presses for attention: Where is the gain in conceptualizing these vital premiss-conclusion relations as relations of logical *consequence*? Where is the good of trying to construe nonmonotonic relations on the model of attenuated and retrofitted monotonic consequences? My own inclination is to say that talk of nonmonotonic consequence misconceives the *import* of nonmonotonicity. We tend to think of it as a distinguishing feature of consequence relations, when what it really is is the defining feature of nontruth preservation.

When premiss-conclusion reasoning is good but not truth-preserving, it is made so by an underlying relation. Any theory of premiss-conclusion reasoning had better have something to say about this, about its nature and how it operates. We should give it the name it both deserves and better reflects how it actually does function. Let's call it *conclusionality*. Conclusionality is an epistemic or epistemic/prudential relation. It is a relation

that helps rearrange our belief states, hence possessing decisional significance. Any struggle to discern whether it is also a consequence relation seems to me to be sailing into the wind.

Abductive conclusions are on the receiving end of this relation; they are occupants of its converse domain. If our present reflections can be made to stand, there is no relation of abductive consequence; and it will cause us no end of distraction trying to figure out how to make it one. It hardly needs saying that depriving a logic of abduction of its own relation of abductive consequence must of necessity rearrange how abductive logic is conceptualized. There are plenty of logicians more than ready to say that a logic without consequence relations is a logic in name only – a logic *façon de parler*, hence a logic that fails universalistic prescriptions. I am otherwise minded. Logic started with *conclusionality* relations. It was adventitiousness, not essence, that brought it about that the ones first considered were also consequence relations. Logic has had a good innings right from the beginning. In a way, this has been unfortunate. The success we've had with consequence has obscured our view of conclusionality. It has led us to think that the more we can get conclusionality *in gen-*

*eral* to be a species of consequence, the faster we'll achieve some theoretical respectability. We would be better served to place conclusionality at the core of logic and to place consequence in an annex of less central importance. If we did that, we could reinstate the logic of abduction and equip it for admittance into universalist respectability. But we could also reinvest to good effect all that energy we've devoted to consequentializing the conclusionality relation, in a refreshed effort to figure how conclusionality actually works in the epistemically sensitive environments in which, perforce, the human organism must operate.

**Acknowledgments.** I would know a good deal less than I presently do about abduction without stimulating instruction from Dov Gabbay, Lorenzo Magnani, Atocha Aliseda, Ahti-Veikko Pietarinen, Peter Bruza, Woosuk Park, Douglas Niño and more recently – especially in relation to sections 11 and 12 – Madeleine Ransom. To all my warmest thanks. My student Frank Hong has also pitched in with astute suggestions; equal gratitude to him. For technical support and everything else that matters, Carol Woods is my go-to gal. Without whom not.

## References

- 6.1 J.R. Josephson, S.G. Josephson (Eds.): *Abductive Inference: Computation, Philosophy, Technology* (Cambridge University Press, Cambridge 1994)
- 6.2 T. Kapitan: Peirce and the structure of abductive inference. In: *Studies in the Logic of Charles Sanders Peirce*, ed. by N. Houser, D.D. Roberts, J. Van Evra (Indiana University Press, Bloomington 1997) pp. 477–496
- 6.3 J. Hintikka: What is abduction? The fundamental problem of contemporary epistemology, *Trans. Charles S. Peirce Soc.* **34**, 503–533 (1998)
- 6.4 P.A. Flach, C.K. Antonis: *Abduction and Induction: Essays on Their Relation and Interpretation* (Kluwer, Dordrecht 2000)
- 6.5 L. Magnani: *Abduction, Reason and Science: Processes of Discovery and Explanation* (Kluwer, Dordrecht 2001)
- 6.6 D.M. Gabbay, J. Woods: *The Reach of Abduction: Insight and Trial, A Practical Logic of Cognitive Systems*, Vol. 2 (North-Holland, Amsterdam 2005)
- 6.7 S. Paavola: Peircean abduction: Instinct or inference?, *Semiotica* **153**, 131–154 (2005)
- 6.8 A.-V. Pietarinen: *Signs of Logic: Peircean Themes on the Philosophy of Language, Games and Communication* (Springer, Dordrecht 2006)
- 6.9 A. Aliseda: Abductive reasoning. In: *Logical Investigation into the Processes of Discovery and Evaluation*, (Springer, Dordrecht 2006)
- 6.10 P.D. Bruza, D.W. Song, R.M. McArthur: Abduction in semantic space: Towards a logic of discovery, *Logic J. IGPL* **12**, 97–110 (2004)
- 6.11 G. Schurz: Patterns of abduction, *Synthese* **164**, 201–234 (2008)
- 6.12 P. Bruza, A. Barros, M. Kaiser: Augmenting web service discovery by cognitive semantics and abduction, *Proc. IEEE/WIC/ACM Int. Joint Conf. on Web Intelligence and Intell. Agent Technol.* (IET, London 2009) pp. 403–410
- 6.13 L. Magnani: *Abductive Cognition. The Epistemological and Eco-Cognitive Dimensions of Hypothetical Reasoning* (Springer, Heidelberg 2009)
- 6.14 Aristotle: Categories. In: *The Complete Works of Aristotle*, ed. by J. Barnes (Princeton Univ. Press, Princeton 1985)
- 6.15 C.S. Peirce: *Pragmatism and Pragmaticism*, Collected Papers of Charles Sanders Peirce, Vol. 5 (Harvard Univ. Press, Cambridge 1934), ed. by C. Hartshorne, P. Weiss
- 6.16 C.S. Peirce: *Reasoning and the Logic of Things: The Cambridge Conference Lectures of 1898* (Harvard University Press, Cambridge 1992), ed. by K.L. Kettner
- 6.17 C.S. Peirce: *Science and Philosophy*, Collected Papers of Charles Sanders Peirce, Vol. 7 (Harvard Univ. Press, Cambridge 1958), ed. by A.W. Burks
- 6.18 C.S. Peirce: *Scientific Metaphysics*, Collected Papers of Charles Sanders Peirce, Vol. 6 (Harvard Univ. Press, Cambridge 1935), ed. by C. Hartshorne, P. Weiss

- 6.19 J. Woods: Peirce's abductive enthusiasms, *Protosociol.* **13**, 117–125 (1999)
- 6.20 J. Woods: Cognitive economics and the logic of abduction, *Review of Symbolic Logic* **5**, 148–161 (2012)
- 6.21 J. Woods: *Errors of Reasoning: Naturalizing the Logic of Inference*, Vol. 45 (College Publications, London 2013), Studies in Logic Ser.
- 6.22 P. Gochet, G. Gribomont: Epistemic logic. In: *Logic and the Modalities in the Twentieth Century*, Handbook of the History of Logic, Vol. 7, ed. by M. Dov Gabbay, J. Woods (North-Holland, Amsterdam 2006) pp. 99–195
- 6.23 R. Fagin, J.Y. Halpern, Y. Moses, Y.M. Vardi: *Reasoning About Knowledge* (MIT, Cambridge 1995)
- 6.24 J. Van Benthem: *Logical Dynamics of Information and Interaction* (Cambridge University Press, New York 2011)
- 6.25 N.E. Sahlin, W. Rabinowitz: The evidentiary value model. In: *Handbook of Defeasible Reasoning and Uncertainty Management Systems*, Vol. 1, ed. by D.M. Gabbay, P. Smets (Kluwer, Dordrecht 1998) pp. 247–265
- 6.26 W.V. Quine: Epistemology naturalized. In: *Ontological Relativity and Other Essays*, ed. by W.V. Quine (Columbia University Press, New York 1969)
- 6.27 A. Goldman: What is justified belief? In: *Justification and Knowledge Philosophical Studies Series in Philosophy* **17**, ed. by G. Pappas (Reidel, Dordrecht 1979) pp. 1–23
- 6.28 A. Goldman: A causal theory of knowing, *J. Philosophy* **64**, 357–372 (1967)

# Patterns of Abductive Inference

Gerhard Schurz

This article understands abductive inference as encompassing several special patterns of inference to the best explanation whose structure *determines* a promising explanatory conjecture (an abductive conclusion) for phenomena that are in need of explanation (Sect. 7.1). A *classification* of different patterns of abduction is given in Sect. 7.2, which is intended to be as complete as possible. A central distinction is that between *selective abductions*, which choose an optimal candidate from a given multitude of possible explanations (Sects. 7.3 and 7.4), and *creative abductions*, which introduce new theoretical models or concepts (Sects. 7.5–7.7). While the discussion of selective abduction has dominated the literature, creative abductions are rarely discussed, although they are essential in science. This paper introduces several kinds of creative abduction, such as *theoretical model abduction*, *common-cause abduction*, and *statistical factor analysis*. A demarcation between scientifically fruitful abductions and speculative abductions is proposed, by appeal to two interrelated criteria: *independent testability* and *explanatory unification*. Section 7.8 presents applications of abductive inference in the domains of belief revision and instrumental/technological reasoning.

7.1	<b>General Characterization of Abductive Reasoning and Inference</b> .....	152
7.2	<b>Three Dimensions for Classifying Patterns of Abduction</b> .....	154
7.3	<b>Factual Abduction</b> .....	155
7.3.1	Observable–Fact Abduction .....	155
7.3.2	First–Order Existential Abduction.....	156
7.3.3	Unobservable–Fact Abduction .....	156
7.3.4	Logical and Computational Aspects of Factual Abduction .....	157
7.4	<b>Law Abduction</b> .....	158
7.5	<b>Theoretical–Model Abduction</b> .....	159
7.6	<b>Second–Order Existential Abduction</b> ....	161
7.6.1	Micro–Part Abduction .....	161
7.6.2	Analogical Abduction .....	161
7.6.3	Hypothetical (Common) Cause Abduction .....	162
7.7	<b>Hypothetical (Common) Cause Abduction Continued</b> .....	162
7.7.1	Speculative Abduction Versus Causal Unification: A Demarcation Criterion.....	163
7.7.2	Strict Common–Cause Abduction from Correlated Dispositions and the Discovery of New Natural Kinds	164
7.7.3	Probabilistic Common–Cause Abduction and Statistical Factor Analysis .....	167
7.7.4	Epistemological Abduction to Reality ....	168
7.8	<b>Further Applications of Abductive Inference</b> .....	169
7.8.1	Abductive Belief Revision .....	169
7.8.2	Instrumental Abduction and Technological Reasoning.....	170
	<b>References</b> .....	171



## 7.1 General Characterization of Abductive Reasoning and IBE

This article is based upon the work in *Schurz* [7.1] where abductive inferences are described as special patterns of inference to the best explanation (IBE) whose structure determines a promising explanatory conjecture (an abductive conclusion), and thus serves as a *search strategy* for finding explanations for given phenomena. However, depending on the explanatory *target* and the background knowledge, there are rather *different patterns* of abductive inference. Sections 7.2–7.7 give a detailed reconstruction of these patterns of abductive inference. Section 7.8 presents applications of abduction in the domains of belief revision and instrumental reasoning. The introductory section explains three general theses that underlie the analysis and terminology of this chapter.

### Thesis 7.1 (Induction versus abduction)

*Peirce* [7.2, CP 2.619–2.644], [7.3, CP 5.14–5.212] distinguished between three families of reasoning patterns: deduction, induction, and abduction. Deductions are *non-ampliative* and *certain*: given the premises are true, the conclusion *must* be true. In contrast, inductions and abductions are *ampliative* and *uncertain*, which means that even if the truth of the premises is taken for granted, the conclusion may be false, and is therefore subject to *further testing*.

My first thesis is that induction and abduction are two *distinct families* of ampliative reasoning that are *not reducible* to each other. For this reason, I do *not* regard induction as an umbrella term for all kinds of ampliative (non-deductive) inferences [7.4, p. 42], [7.5], [7.6, p. 28]. Rather, I understand induction in the *narrow Humean sense* in which a property or regularity is *transferred* from the past to the future, or from the observed to the unobserved.

Inductions and abductions can be distinguished by their different *targets*. Both serve the target of extending our knowledge beyond observation, but in different respects. Inductions serve the goal of inferring something about the *future course of events*, which is important for planning, that is, adapting our actions to the course of events. In contrast, abductions serve the goal of inferring something about the unobserved *causes* or *explanatory reasons* of observed events, which is of central importance for manipulating the course of events, that is, adapting the course of events to our wishes [7.3, CP 5.189], [7.7, p. 35]. That abductions cannot be reduced to inductions follows from the fact that inductions cannot introduce new concepts or conceptual models; they merely transfer them to new instances. In contrast, some kinds of ab-

ductions *can* introduce new concepts [7.3, CP 5.170]. Following *Magnani* [7.8, p. 20], [7.9], I call abductions that introduce new concepts or models *creative*, in contrast to *selective* abductions whose task is to choose the best candidate from among a given multitude of possible explanations.

### Thesis 7.2 (Inference to the best available explanation may not be good enough)

Most authors agree that Harman's IBE [7.10] has to be modified in (at least) the following respect: Nobody knows *all* possible explanations for a given phenomenon, and therefore, what one really has instead of an IBE is an inference to the *best available* explanation, in short an IBAE.

However, as *Lipton* [7.11, p. 58] has pointed out – and this is my second thesis: the best available explanation is not always *good enough* to be rationally acceptable. If a phenomenon is poorly understood, then one's best available explanation is usually *pure speculation*. In the early history of human mankind, the best available explanations of otherwise unexplainable natural phenomena, such as the rising of the Sun or the coming of the rain, was in terms of the actions of supernatural agents. Speculative explanations of this sort fail to meet important scientific demands that are discussed in Sect. 7.1.

Summarizing, the rule IBE is not feasible, and the rule IBAE is not generally acceptable. What is needed for more satisfying versions of abductive rules are (1) *minimal criteria* for the *acceptability* of scientific abductions, and (2) *comparative criteria* for the *quality* of the abducted explanations. Concerning (2), many authors have pointed out [7.12, pp. 443] that a *unique* criterion for the quality of an explanation does not exist – we rather have several criteria that may come into mutual conflict. For example, *Lipton* [7.11, pp. 61] argued that in scientific abductions we do not prefer the *likeliest* (most probable) explanation, but the *loveliest* explanation (that with highest explanatory strength), while *Barnes* [7.13] objected that loveliness without likeliness is scientifically unacceptable. One result of the present article, which has a direct bearing on this debate, is that there is *no general answer* to these questions, because the evaluation criteria for different kinds of abductions are different. For example, in the area of selective factual abductions, comparative plausibility criteria are important, while in the area of creative second-order existential abductions, one needs only minimal acceptability criteria.

### *Thesis 7.3 (The strategic role of abduction as means for discovery)*

All inferences have a *justificatory* (or *inferential*) and a *strategic* (or *discovery*) function, but to a different degree. The justificatory function consists of the justification of the conclusion, *conditional* on the justification of the premises. The strategic function consists of *searching* for the most promising conjecture (conclusion), which is set out for further empirical testing, or in *Hintikka's* words, which stimulates new *questions* [7.14, p. 528], [7.15, Sect. 14].

In deductive inferences the justificatory function is fully realized, because the premises guarantee the truth of the conclusion. Deductive inferences may also serve important strategic functions, because many different conclusions can be derived from the same premises. In inductive inferences, there is not much search strategy involved, because the inductive conclusions of a premise set are narrowly defined by the operations of generalization over instances. So the major function of inductive inferences is justificatory, but their justificatory value is uncertain. In contrast, in abductive inferences, the strategic function becomes *dominant*. Different from the situation of induction, in abduction problems we are often confronted with thousands of possible explanatory conjectures – anyone in the village might be the murderer. The essential function of abductions is their role as *search* strategies that tell us which explanatory conjecture we should set out *first* for further inquiry [7.14, p. 528] – or more generally, which suggest a *short* and *most promising* (though not necessarily successful) path through the exponentially excessive *search space* of possible explanatory reasons.

In contrast, the justificatory function of abductions is minor. *Peirce* pointed out that abductive hypotheses are *prima facie* not even probable, as inductive hypotheses, but merely possible [7.3, CP 5.171]. Only upon being confirmed by further tests may an abductive hypothesis become probable. However, I cannot completely agree with *Peirce* or other authors [7.14, 16], [7.17, p. 192] who think that abductions are merely a discovery procedure and whose justificatory value is zero. As *Niiniluoto* pointed out, “abduction as a motive for pursuit cannot always be sharply distinguished from considerations of justification” [7.12, S442]. *Niiniluoto's* point is confirmed by a Bayesian analysis: If a hypothesis  $H$  explains an explanandum  $E$  (where  $P(H), P(E) \neq 0, 1$ ), then  $P(E|H)$  ( $E$ 's posterior probability) has been raised compared to  $P(E)$  ( $E$ 's prior probability), which implies (by probability theory) that  $P(H|E) > P(H)$ , i. e.,  $E$  raises  $H$ 's probability, if only a little bit.

It is essential for a good search strategy that it leads us to an optimal conjecture in a reasonable amount of

time. In this respect, the rule of IBE fails completely. It just tells us that we should choose the best (available) explanation without giving us any clue of how to find it. To see the problem, as presented by a humorous example, think of someone in a hurry who asks an IBE-philosopher for the right way to the railway station and receives the following answer: *Find out which is the shortest way among all ways between here and the train station – this is the route you should choose.* In other words, IBE merely reflects the justificatory but misses the strategic function of abduction, which in fact is its *essential* function. For this reason, the rule of IBE is, epistemically, rather uninformative [7.18, p. 281].

*Peirce* once remarked that there are sheer myriads of possible hypotheses that would explain a given experimental phenomena, yet scientists usually manage to find the true hypothesis after only a small number of guesses [7.19, CP 6.5000]. But *Peirce* did not propose any abductive rules for conjecturing new theories; he rather explained this miraculous ability of human minds by their *abductive instincts* [7.20, CP 5.47, fn. 12; 5.172; 5.212]. The crucial question seems to be whether there can be anything like a *logic* of discovery. *Popper* and the logical positivists correctly observed that the justification of a hypothesis is independent from the way it was discovered. This does not imply, however, that it would not be *desirable* to have *in addition* good rules for discovering explanatory hypotheses – if there only *were* such rules [7.16]. This paper intends to show that there *are* such rules; in fact, every kind of abduction pattern presented in this paper constitutes such a rule.

The majority of the recent literature on abduction has aimed at *one* most general schema of abduction that matches every particular case. I do not think that good heuristic rules for generating explanatory hypotheses can be found along this route, because such rules are dependent on the *specific type* of abductive scenario, for example, concerning whether the abduction is mainly selective or creative (etc.). In the remainder of this paper, I will pursue a different route to characterizing abduction, which consists of modeling various particular schemata of abduction, each fitting a particular kind of conjectural situation. Two major results of my paper can be summarized as follows:

#### *Result 7.1*

There exist rather *different kinds* of abductive patterns. While some of them enjoy a broad discussion in the literature, others have been neglected, although they play an important role in science. The epistemological role and the evaluation criteria of abduction are different for the different patterns of abduction.

**Result 7.2**

In all cases, the crucial function of a pattern of abduction or IBE consists in its function as a *search strategy* that leads us, for a given kind of *scenario*, to the most promising explanatory conjecture, which may then subject to further test. In selective abductions, the difficulty usually lies in the fact that the search space of possible conjectures is very *large*. On the other hand, in creative

abductions the difficulty often consists in finding just *one* conjecture that meets the required constraints.

More important than my general theses and results are the *particular results* of the following sections, in which I model each kind of abduction as a specific inference *schema* in which the most promising explanatory conjecture is structurally determined.

**7.2 Three Dimensions for Classifying Patterns of Abduction**

I classify patterns of abduction along three dimensions:

1. Along the kind of *hypothesis* that is abduced, i. e., that is produced as a conjecture
2. Along the kind of *evidence* that the abduction is meant to explain
3. According to the *beliefs* or *cognitive mechanisms* that *drive* the abduction.

I signify the different kinds of abduction according to the first dimension. But the three dimensions are not

independent: the properties of an abductive pattern in the second and third dimension are in characteristic covariance with its status in the first dimension. Also, how the evidence together with the background knowledge conveys *epistemic support* to the abduced hypothesis, and by which follow-up procedures the abduced hypotheses is put to further test, depend crucially on the kind of abduced hypothesis. Figure 7.1 anticipates my classification of kinds of abductive patterns as a first orientation for the reader – the listed kinds of abduction are explained in Sects. 7.2–7.7.

Kind of abduction	Evidence to be explained	Abduction produces	Abduction is driven by
<b>Factual abduction</b>	Singular empirical facts	New facts (reasons/causes)	Known laws or theories
— <i>Observable-fact-abduction</i>	"	Factual reasons	Known laws
— <i>First-order existential abduction</i>	"	Factual reasons postulating new unknown individuals	"
— <i>Unobservable-fact-abduction</i> (Historical abduction)	"	Unobservable facts (facts in the past)	"
<b>Law abduction</b>	Empirical laws	New laws	"
<b>Theoretical-model-abduction</b>	General empirical phenomena (laws)	New theoretical models of these phenomena	Known theories
<b>Second-order existential abduction</b>	"	New laws/theories with new concepts	Theoretical background knowledge (b.k.)
— <i>Micro-part abduction</i>	"	Microscopic composition	Extrapolation of b.k.
— <i>Analogical abduction</i>	"	New laws/theories with analogous concepts	Analogy with b.k.
— <i>Hypothetical cause abduction</i>	"	Hidden (unobservable) causes	(see below)
— <i>Speculative abduction</i>	(")	(")	Speculation
— <b>Common-cause abduction</b>	"	Hidden <i>common</i> causes	Causal unification
— <i>Strict common-cause abduction</i>	"	New theoretical concepts	"
— <i>Statistical factor analysis</i>	"	"	"
— <i>Abduction to reality</i>	Introspective phenomena	Concept of external reality	"

**Fig. 7.1** Classification of kinds of abduction

### 7.3 Factual Abduction

In factual abductions, both the evidence to be explained and the abducted hypothesis are *singular facts*. Factual abductions are always *driven* by known implicational laws going from causes to effects, and the abducted hypotheses are found by backward reasoning, inverse to the direction of the law-like implications. Factual abduction may also be called *retroduction*; *Chisholm* [7.21, Ch. IV.2] speaks of *inverse induction*. This kind of abduction has the following structure (the double line = indicates that the inference is uncertain and preliminary)

*Known law:* If  $Cx$ , then  $Ex$   
*Known evidence:*  $Ea$  has occurred  
 =====  
*Abducted conjecture:*  $Ca$  could be the reason.

One may call the factual abduction schema *the official Peirce abduction schema*, since *Peirce* [7.2, CP 2.619–2.644] formalized abduction in this way and named it *hypothesis*; later he generalized abduction in the way described in Sect. 7.1. Factual abductions are omnipresent in common sense reasoning, and presumably rely on inborn abductive instincts of hominids. Prototypical examples are detective stories [7.22], or more generally, all sorts of *causal interpretations of traces*. The AI literature is focused almost exclusively on factual abductions (Sect. 7.3.4). Depending on the epistemological nature of the abducted fact, one can distinguish between the following three subpatterns.

#### 7.3.1 Observable-Fact Abduction

Here one reasons according to the fact-abduction schema from observed effects ( $Ea$ ) to non-observed but observable causes ( $Ca$ ) in the background of known laws. The follow-up test procedure consists of the attempt to gain direct evidence for the abducted conjecture. In the example of a murder case, such direct evidence would be given, for example, by a confession of the putative murderer.

In the setting of factual abduction, the problem often consists of the *combinatorial explosion* of the search space of possible causes, in the presence of a rich background store of laws but in the *absence* of a rich factual knowledge. Thus, factual abductions are primarily *selective* in the sense of *Magnani* [7.8, p. 20], and their epistemic support depends on the degree to which the background knowledge increases their *probability* in comparison to alternative possible causes. Consider the following example: If your evidence consists of the trace of the imprints of sandals

on an elsewhere empty beach, then your immediate conjecture is that somebody was recently walking here. How did you arrive at this conjecture? Classical physics allows for myriads of ways of imprinting footprints into the sand of the beach, which reach from cows wearing sandals on their feet to footprints that are drawn into the sand or blown by the wind, etc. The majority of these possible abductive conjectures will never be considered by us because they are extremely improbable. The major strategic algorithm that we apply in factual abduction cases of this sort is a *probabilistic elimination technique*, which usually works in an unconscious manner: our mind quickly scans through our large memory store containing millions of memorized possible scenarios and only those that have minimal plausibility pop up in our consciousness.

So, probabilistic evaluation of possible causes given certain effects and elimination of implausible causes plays a central role in factual abductions. *Fumerton* [7.23, p. 592] has gone further and argued that factual abduction can even be *reduced* to ordinary inductive-statistical inference. More precisely, he argues that the first inference pattern (below) can be reduced to the second inference pattern (below) in the following way (where  $P(-)$  denotes subjective-epistemic and  $p(-)$  statistical probability, and  $K$  expresses background knowledge).

*Abductive inference :*  
 $L : \forall x(Fx \rightarrow Gx)$   
 $Ga$   
 =====  
 $Fa$  (*presupposition :*  
 $P(Fa|Ga \wedge L \wedge K) = \text{high}$ )

Fumerton's  
 reduction: ↓

*Inductive – statistical inference :*  
 $L' : p(Fx|Gx) = \text{high}$   
 $Ga$   
 =====  $P(Fa|Ga \wedge L') = \text{high}$   
 $Fa$

Although Fumerton's reduction seems reasonable in some cases, I see two reasons why his argument is not generally correct. Firstly, the abductive hypothesis is probabilistically evaluated not merely in the light of the evidence  $Ga$  and an inverse statistical law  $L'$ , but in the light of the entire background knowledge  $K$ . Fumerton may reply that the inference pattern on the right may be appropriately extended so that it includes background

knowledge. But secondly, Fumerton’s proposed transformation does not correspond to psychological reality, nor would it be strategically recommendable. Every individual case (or effect) is *different*, and hence, only a small fraction of possible cause-effect scenarios are encountered frequently enough in a human lifetime in order to be represented by Fumerton-like conditional probabilities. For example, if you are *not* a turtle expert and you observe the trace of a turtle in the sand, then the only way in which you may arrive at the right guess that there was a turtle crawling here is by careful backward reasoning combined with elimination. Unless you are a turtle hunter, it is unlikely that you will have explicitly stored information concerning the typical sand traces of turtles via a corresponding forward conditional of the sort proposed by Fumerton.

### 7.3.2 First-Order Existential Abduction

This subcase of factual abduction occurs when the antecedent of a law contains so-called *anonymous* variables, i. e., variables that are not contained in the consequent of the law. In the simplest case, the formal structure of first-order existential abduction is as follows (cf. also [7.24, p. 57]).

L :  $\forall x \forall y (Ryx \rightarrow Hx)$   
 logically equivalent :  $\forall x (\exists y Ryx \rightarrow Hx)$   
 Ha  
 =====  
 Conjecture :  $\exists y Rya$

Instantiating the consequent of the law with *a* and backward chaining yields a law-antecedent in which one variable remains uninstantiated (*Rya*). In such a case, the safest abductive conjecture is one in which we existentially quantify over this variable. We have already discussed an example of this sort in Sect. 7.3.1: from the footprint in the sand we abductively infer that *some* man was walking on the beach. *Prendinger* and *Ishizuka* [7.25, p. 322] call first-order existential abduction *element-creative abduction*, because here the existence of an unknown object is hypothesized.

Note, however, that only in some cases will we be satisfied with the existential conjecture. In other cases, in particular in criminal cases, all depends on finding out *which* individual is the one whose existence we conjecture – who was the murderer? Here one is not satisfied with a first-order existential abduction but want to have a proper, fully-instantiated fact abduction.

In observable-fact abduction, the abducted hypothesis may at later stages of inquiry be confirmed by direct observation – for example, when we later meet the man

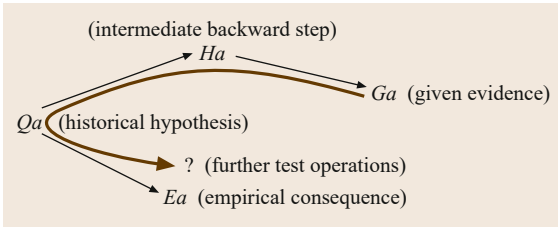
who had walked yesterday on this beach. In this case, the weak epistemic support that the abductive inference conveys to the conjecture gets replaced by the strong epistemic support provided by the direct evidence: abduction has played an important strategic role, but it no longer plays a justificatory role. This is different, however, in all of the following patterns of abduction, in which the abductive hypothesis is not directly observable, but only indirectly confirmable via its empirical consequences.

### 7.3.3 Unobservable-Fact Abduction

This kind of abduction has the same formal structure as observable-fact abduction, but the abducted fact is unobservable. The *typical* case of unobservable-fact abductions are *historical-fact abductions*, in which the abducted fact is unobservable because it is located in the distant past. The abducted fact may also be unobservable in principle, because it is a theoretical fact. However, in such a case the abduction is usually not driven by simple implicational laws, but by a quantitative theory, and the abducted theoretical fact corresponds to a theoretical model of the observed phenomenon: This sort of abduction differs crucially from law-driven factual abduction and is therefore treated under the separate category of *theoretical-model abduction* (Sect. 7.5).

Historical-fact abductions are of obvious importance for all historical sciences [7.12, p. 442]. Assume, for example, that biologists discover marine fossil records, say fish bones, in the ground of dry land. They conjecture abductively, given their background theories, that some geological time span ago there was a sea here. Their hypothesis cannot be directly verified by observations. So the biologists look for further empirical consequences that follow from the abducted conjecture plus background knowledge – for example, further geological indications such as calcium deposits, or marine shell fossils, etc. If the latter findings are observationally verified, the abductive conjecture is confirmed. Logically speaking, an unobservable-fact abduction performs a combination of abductive backward reasoning and deductive or probabilistic forward reasoning to consequences that can be put to further test. This is graphically displayed by the *bold arrow* in Fig. 7.2.

If the empirical consequence *Ea* is verified, then both pieces of evidence *Ga* and *Ea* provide epistemic support for the abducted hypothesis *Ha* (modulo probabilistic considerations in the light of the background knowledge). So, the initial abductive inference has not only a strategic value, but *keeps* its justificatory value.



**Fig. 7.2** Historical-fact abduction (the *bold arrow* indicates the route of the abduction process)

### 7.3.4 Logical and Computational Aspects of Factual Abduction

If one’s background knowledge does not contain general theories but just a finite set of (causal) implicational laws, then the set of possible abductive conjectures is finite and can be generated by *backward-chaining* inference procedures. In this form, abductive inference has frequently been studied in AI research [7.26, 27]. Given is a knowledge base  $\mathbf{K} = \langle \mathbf{L}[x], \mathbf{F}[a] \rangle$  in form of a finite set  $\mathbf{L}[x]$  of monadic implicational laws going from conjunctions of open literals to literals, and a finite set  $\mathbf{F}[a]$  of facts (closed literals) about the individual case  $a$ . (A literal is an atomic formula or its negation.) Given is moreover a certain *goal*, which is a (possibly conjunctive) fact  $G_a$  that needs to be explained. One is not interested just in any hypotheses that (if true) would explain the goal  $G_a$  given  $\mathbf{K}$ , but only in those hypotheses that are not further potentially explainable in  $\mathbf{K}$  [7.28, p. 133], [7.29].

So formally, the candidates for abducible hypotheses are all closed literals  $A[a]$  such that  $A[a]$  is neither a fact in  $\mathbf{F}[a]$ , nor the consequent (*head*) of a law, i. e.,  $A[a]$  cannot be further explained by other laws in  $\mathbf{K}$ . The set of all possible abductive conjectures  $A[a]$  for arbitrary abduction tasks in  $\mathbf{K}$  is called the set of *abducibles*  $\mathbf{H}[a]$ . The *abductive task* for goal  $G_a$  is then defined as follows: Find *all possible* explanations, i. e., all *minimal* sets  $\mathbf{E}[a]$  of singular statements about  $a$  such that:

- (i)  $\mathbf{E}[a] \subseteq \mathbf{F}[a] \cup \mathbf{H}[a]$

- (ii)  $\mathbf{L}[x] \cup \mathbf{F}[a] \cup \mathbf{E}[a]$  is consistent
- (iii)  $\mathbf{L}[x] \cup \mathbf{E}[a]$  logically implies  $G[a]$  (by forward chaining).

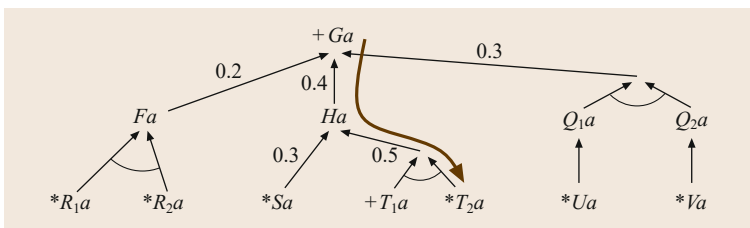
Those elements of the explanatory sets  $\mathbf{E}[a]$  that are abducibles are the abductive hypotheses for  $G[a]$ .

This kind of abduction problem is graphically displayed in Fig. 7.3 in the form of a so-called *and-or tree* [7.30, Ch. 13]. The *labeled* nodes of an and-or tree correspond to literals, unlabeled nodes represent conjunctions of them, and the directed edges (arrows) correspond to laws in  $\mathbf{L}[x]$ . Arrows connected by an *arc* are and-connected; those without an arc are or-connected. Written as statements, the laws underlying Fig. 7.3 are  $\forall x(Fx \rightarrow Gx)$ ,  $\forall x(Hx \rightarrow Gx)$ ,  $\forall x(Q_1x \wedge Q_2x \rightarrow Gx)$ ,  $\forall x(R_1x \wedge R_2x \rightarrow Fx)$ ,  $\forall x(Sx \rightarrow Hx)$ ,  $\forall x(T_1x \wedge T_2x \rightarrow Hx)$ ,  $\forall x(Ux \rightarrow Q_1x)$ ,  $\forall x(Vx \rightarrow Q_2x)$ . Besides the goal  $G_a$ , the only known fact is  $T_1a$ .

Algorithms for this sort of task have been implemented, for example, in the programming language Prolog in the form of backward-chaining with backtracking to all possible solutions.

The task of finding *all* possible explanations has exponential complexity and, thus, is intractable (that is, the time of this task increases exponentially in the number of data points and possible hypotheses). Only the complexity of finding some explanation has polynomial complexity and is tractable [7.26, Ch. 7, p. 165, Th. 7.1, 7.2]. Therefore it is crucial to constrain the search space by probabilistic (or plausibilistic) evaluation methods. A simple heuristic strategy is the *best-first* search: for each or-node one processes only that successor that has the highest plausibility value (among all successors of this node). The route of a best-first abduction search is depicted in Fig. 7.3 by the *bold arrow*.

A related but more general framework for factual abductions via backward reasoning is abduction within Beth tableaux [7.15], [7.7, Ch. 4]. Even more general frameworks are Gabbay’s *labeled deductive systems* [7.31, Part III] and abduction within *epistemic dynamic logic* [7.32]. An alternative approach for computing abductive hypotheses is abductive rea-



**Fig. 7.3** Search space for a factual abduction problem. + indicates a known fact, \* indicates possible abductive hypotheses. The *numbers* are probability values (they do not add up to 1 because of an unknown residual probability). The *bold arrow* indicates the route of a best-first search, which leads to the abductive conjecture  $T_2a$

soning in the framework of *adaptive logic* [7.33, 34]. Here one infers singular explanatory hypotheses by backward chaining defeasibly and excludes them as soon as contradicting abnormality statements turn out to be derivable in later stages of the proof. As a result, one doesn't compute all possible minimal explanations (consistent with the knowledge base) but only those that are undefeated.

Besides probabilistic elimination, the second major technique of constraining the search space is *intermediate information acquisition*: not only the ultimately abducted conjectures, but also intermediate conjectures (nodes) along the chosen search path can be set out for further empirical test – or in the framework

of Hintikka et al. [7.15], they may stimulate further interrogative inquiry [7.35, Ch. 6]. As an example, consider again a criminal case: If backward reasoning leads to the possibility that the butler could have been the murderer, and along an independent path, that the murderer must have been left handed, then before continuing the abductive reasoning procedure one better finds out first whether the butler is indeed left handed. There are also some AI abduction systems that incorporate question-asking modules. For example, the RED system, designed for the purpose of red-cell antibody identification based on antigen-reactions of patient serum, asks intermediate questions of a database [7.26, pp. 72].

### 7.4 Law Abduction

In this kind of abduction, both the evidence to be explained and the abducted hypothesis are implicational laws, and the abduction is driven by one (or several) known implicational laws. Because of the latter fact, this kind of abduction is more similar to factual abductions than to theory-driven abductions that are discussed in Sect. 7.5. Law abductions can already be found in Aristotle, and they correspond to what Aristotle has called the mind's power of *hitting upon the middle term* of a syllogism (An. Post., I, 34). Here is an example.

Background law:  $\forall x(Cx \rightarrow Ex)$   
 Whatever contains sugar tastes sweet

Empirical law to be explained:  $\forall x(Fx \rightarrow Ex)$   
 All pineapples taste sweet

=====

Abducted conjecture:  $\forall x(Fx \rightarrow Cx)$   
 All pineapples contain sugar.

A more general example of law abduction in qualitative chemistry is this,

All substances that contain molecular groups of the form  $C$  have property  $E$ .

All substances of empirical kind  $S$  have certain empirical properties  $E$ .

=====

Conjecture: Substances of kind  $S$  have molecular characteristics  $C$ .

In the case where there are several causal laws of the form  $\forall x (C_i x \rightarrow Ex)$ , one has to select the most *plausible* one. In any case, the conclusions of law abductions are conjectural and in strong need of further support.

Flach and Kakas [7.27, pp. 21] have argued that a law abduction can be *reduced* to the following combination of a fact abduction and an inductive generalization

Background law:  $\forall x(Cx \rightarrow Ex)$

Observed facts:  $Fa_i \wedge Ea_i \quad 1 \leq i \leq n$   
 $\rightarrow$  Induction basis for:  $\forall x(Fx \rightarrow Ex)$

===== Factual abduction

Abducted hypotheses:  $Ca_i \quad 1 \leq i \leq n$   
 hence:  $Fa_i \wedge Ca_i \quad 1 \leq i \leq n$   
 $\rightarrow$  Induction basis for:  $\forall x(Fx \rightarrow Cx)$ .

This decomposition, however, is somewhat artificial. Law abductions are usually performed in one single conjectural step. We don't form the abductive hypothesis that  $x$  contains sugar for each observed pineapple  $x$ , one after the other, and then generalize it, but we form the law-conjecture *pineapples contain sugar* at once.

All patterns of abduction that we have discussed so far are driven by known qualitative implication laws, and they are mainly *selective*, i. e., their driving algorithm draws a most promising candidate from a class of possible conjectures, which is very large but in principle constructible. Discussion of such patterns dominates the abduction literature. The patterns of abductions to be discussed in the next sections are rarely discussed in the literature. They are not driven by implicational laws, but either by scientific theories, or by (causal) unification procedures. Moreover, they are not mainly selective but mainly *creative*, that is, the underlying abduction operation constructs something new, for example a new theoretical model or a new theoretical concept.

## 7.5 Theoretical-Model Abduction

The explanandum of a theoretical-model abduction is typically a well-confirmed and reproducible empirical phenomenon expressed by an *empirical law* – for example, the phenomenon that wood floats in water but a stone sinks in it. The abduction is driven by an *already established* scientific theory that is usually quantitatively formulated. The abductive task consists in *finding theoretical* (initial and boundary) *conditions* that describe the causes of the phenomenon in the theoretical language and that allow the mathematical derivation of the phenomenon from the theory. Halonen and Hintikka [7.36] argue that this task makes up the essential point of the scientist’s explanatory activity. Formally, these theoretical conditions are expressed by factual or law-like statements, but their semantic content corresponds to what one typically calls a *theoretical model* for a particular kind of phenomenon within a *given* theory, whence I speak of *theoretical-model abduction*. Note also that with my notion of a *model* I do not imply a particular kind of formalization of models: they can be represented by *statements* or by set-theoretical *models* [7.37, p. 109]. A general translation between sentential and model-theoretic theory representations is developed in Schurz [7.38].

As an example, consider Archimedes’ theoretical model of the phenomenon of buoyancy. Here one searches for a theoretical explanation of the fact that certain substances like stones and metals sink in water while others like wood and ice float on water, *solely in terms of mechanical and gravitational effects*. Archimedes’ ingenious abductive conjecture was that the amount of water that is supplanted by the floating or sinking body tends to lift the body upwards, with a force  $f_w$  that equals the weight of the supplanted water (Fig. 7.4). If this force is greater than the weight of the body ( $f_B$ ) the body will float, otherwise it will sink. Since the volume of supplanted water equals the volume of the part of the body that is underwater, and since the weight is proportional to the mass of a body, it follows that the body will sink exactly if its density (mass per volume) is greater than the density of water.

The situation of theoretical-model abduction is rather different from the situation of factual abductions: one does *not* face here the problem of a huge multitude of possible theoretical models or conjectures, since the given theory *constrains* the space of possible causes to a small class of basic parameters (or generalized *forces*) by which the theory models the domain of phenomena that it intends to explain. In the Archimedean case, the given theory presupposes that the ultimate causes can only be contact forces and gravitational forces – other ultimate causes such as intrinsic capacities of bodies

to float or invisible water creatures, etc., are excluded. Therefore, the real difficulty of theoretical-model abduction does not consist in the elimination of possible explanations – this elimination is already achieved by the given theory – but of finding just *one* plausible theoretical model that allows the derivation of the phenomenon to be explained. If such a theoretical model is found, this is usually celebrated as a great scientific success.

Theoretical-model abduction is the typical theoretical activity of *normal science* in the sense of Kuhn [7.39], that is, the activity of extending the application of a given theory core (or paradigm) to new cases, rather than changing a theory core or creating a new one. If the governing theory is *classical physics*, then examples of theoretical model abduction come in the hundreds, and physics text books are full of them. Examples are the theoretical models underlying:

1. The trajectories (paths) of rigid bodies in the constant gravitational field of the Earth (free fall, parabolic path of ballistic objects, gravitational pendulum, etc.)
2. The trajectories of cosmological objects in position-dependent gravitational fields (the elliptic orbits of planets, Kepler’s laws, the Moon’s orbit around the Earth, and the lunar tides, interplanet perturbations, etc.)
3. The behavior of solid, fluid or gaseous macroscopic objects viewed as systems of more-or-less coupled mechanical atoms (the modeling of pressure, friction, viscosity, the thermodynamic explanation of heat and temperature, etc.); and finally
4. The explanation of electromagnetic phenomena by incorporating electromagnetic forces into classical physics [7.40, Ch. 5.3].

While for all other kinds of abductions we can provide a *general* formal pattern and algorithm by which one can *generate* a most promising explanatory hypothesis, we cannot provide such a general pattern for theoretical model abduction because here all depends

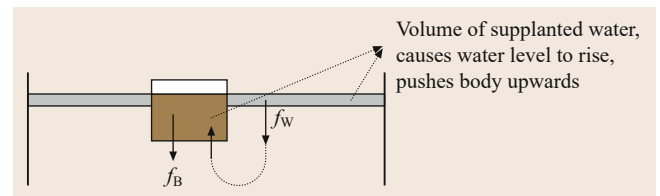


Fig. 7.4 Theoretical conditions that allow the mechanical derivation of the law of buoyancy



on which theory is assumed in the background. But if the theory is specified, then such patterns can often be provided: they are very similar to what Kitcher [7.41, p. 517] calls a schematic explanatory argument, except that the explanandum is now given and the particular explanatory premises have to be found within the framework of the given theory. See the example in Tab. 7.1.

Theoretical-model abduction can also be found in higher and more special sciences than physics. In chemistry, the explanations of the atomic component ratios (the chemical gross formulae) by a three-dimensional molecular structure are the results of theoretical-model abductions; the given theory here is the periodic table plus Lewis' octet rule for forming chemical bonds. A computational implementation is the automatic abduction system DENDRAL [7.42, pp. 234], which abducts the chemical structure of organic molecules, given their mass spectrum and their gross formula.

Theoretical model abductions also take place in evolutionary theory. The reconstruction of evolutionary trees of descent from given phenotypic similarities is a typical abductive process. The basic evolutionary-theoretical premise here is that different biological species descend from common biological ancestors from which they have split apart by discriminative mutation and selection processes. The alternative abductive conjectures about trees of descent can be evaluated by probability considerations. Assume three species  $S_1$ ,  $S_2$ , and  $S_3$ , where both  $S_1$  and  $S_2$  but not  $S_3$  have a new property  $F$  – in Sober's example,  $S_1$  is sparrows,  $S_2 =$  robins,  $S_3 =$  crocodiles, and  $F =$  having wings [7.43, pp. 174–176]. In this case, the tree of descentance  $T_1$  where the common ancestor  $A$  first splits into  $S_3$  and the common ancestor of  $S_1$  and  $S_2$ , which has already  $F$ , requires only one mutation-driven change of non- $F$  into  $F$ , while the alternative tree of

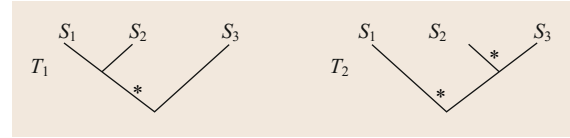


Fig. 7.5 Two alternative trees of descentance; \* = mutation of non- $F$  into  $F$

descentance  $T_2$  in which  $A$  first splits into  $S_1$  and a common  $F$ -less ancestor of  $S_2$  and  $S_3$  requires two such mutations (Fig. 7.5). So probabilistically  $T_1$  is favored as against  $T_2$ .

There are some well-known examples where closeness of species due to common descent does not go hand in hand with closeness in terms of phenotypic similarities: Examples of this sort are recognized because there are several independent kinds of evidence that the tree of descentance must simultaneously explain, in particular:

1. Phenotypic similarities
2. Molecular similarities
3. The fossil record [7.44, Ch. 17].

An example of qualitative-model abduction in the area of humanities is interpretation [7.31, Sect. 4.1]. The explanandum of interpretations are the utterances, written text, or the behavior of given persons (speakers, authors, or agents). The abduced models are conjectures about the beliefs and intentions of the given persons. The general background theory is formed by certain parts of folk psychology, in particular the general premise of all rational explanations of actions, namely, that normally or ceteris paribus, persons act in a way that is suited to fulfill their goals given their beliefs about the given circumstances [7.45, Sect. 1]. More specific background assumptions are hermeneu-

Table 7.1

#### Abduction pattern of Newtonian particle mechanics:

*Explanandum:* A kinematical process involving (a) some moving particles whose position, velocity and acceleration at a variable time  $t$  is an empirical function of their initial conditions, and (b) certain objects defining constant boundary conditions (e.g., a rigid plane on which a ball is rolling, or a large object that exerts a gravitational force, or a spring with Hooke force, etc.)

Generate the abduced conjecture as follows: (i) Specify for each particle its mass and all non-neglectible forces acting on it in dependence on the boundary conditions and on the particle's position at the given time. (ii) Insert these specifications into Newton's second axiom (which says that for each particle  $x$  and time  $t$ , the sum of all forces on  $x$  at  $t$  equals the mass of  $x$  times the acceleration of  $x$  at  $t$ ). (iii) Try to solve the resulting system of differential equations. (iv) Check whether the resulting time-dependent trajectories fit the empirical function mentioned in the explanandum; if yes, the conjecture is preliminarily confirmed; if no, then search for (perturbing) boundary conditions and/or forces that may have been overlooked.

tic rationality presumptions [7.46], Grice’s maxims of communicative cooperation [7.47], and common contextual knowledge. The investigation of interpretation as an abductive process is also an important area in AI [7.48].

What all abduction schemata discussed so far have in common is that they are driven by *known* laws or

theories, and hence, they work within a *given conceptual space*. In other words, the abduction schemata discussed so far cannot introduce *new concepts*. In the next section we turn to abduction schemata that can do this: Since their explanans postulate the existence of a new kind of property or relation, we call them *second-order existential abductions*.

## 7.6 Second-Order Existential Abduction

The explanandum of a second-order existential abduction consists of one or several general empirical phenomena, or laws. What one abducts is an at least a *partly* new property or kind of concept governed by an at least partly new theoretical law. Depending on whether the concept is merely partly or completely new, the abduction is driven by extrapolation, analogy, or by pure unification. We discuss these kinds of abductions in the following subsections Sects. 7.6.1–7.7.4.

### 7.6.1 Micro-Part Abduction

In this most harmless case of second-order existential abduction, one abducts a hypothesis about the unobservable micro-parts of observable objects that obey the *same* laws as the macroscopic objects, in order to explain various observed empirical phenomena. The prototypical example is the *atomic hypothesis* that was already conjectured in antiquity by Leucippus and Democritus and was used to explain such phenomena as the dissolution of sugar in water. These philosophers have abducted a new natural kind term: *atoms*, which are the smallest parts of all macroscopic bodies, being too small to be observable, but otherwise obeying the *same* mechanical laws as macroscopic bodies. So what one does here is to *extrapolate* from macroscopic concepts and laws to the microscopic domain – whence we may also speak here of *extrapolative* abduction. In the natural sciences after Newton, the atomic hypothesis turned out to have enormous explanatory power. For example, Dalton’s atomic hypothesis had successfully explained Avogadro’s observation that equal volumes of gases contain the same number of gas particles. Dalton also postulated that all substances are composed of molecules built up from certain atoms in certain integer-valued ratios, in order to explain the laws of constant proportions in chemical reactions [7.49, pp. 259]. The different states of aggregation of substances (solid, fluid, and gaseous) are explained by different kinds of intermolecular distances and interactions. We conclude our list of examples here, although many more applications of the atomic hypothesis could be mentioned.

Extrapolative micro-part abductions differ from analogical abductions insofar as the atoms are not merely viewed as *analogical* to mechanical particles; they are literally taken as tiny mechanical particles (though too small to be observable). Nevertheless one may view extrapolative abductions as a *pre-stage* of analogical abductions, which we are going to discuss now.

### 7.6.2 Analogical Abduction

Here one abducts a partially new concept together with partially new laws that connect this concept with given (empirical) concepts, in order to explain the given law-like phenomenon. The concept is only partly new because it is analogical to familiar concepts, and this is the way in which this concept was discovered. So analogical abduction is *driven* by analogy. We first consider *Thagard’s* [7.24] example of sound waves.

*Background knowledge:* Laws of propagation and reflection of water waves.

*Phenomenon to be explained:* Propagation and reflection of sound

=====

*Abductive conjecture:* Sound consists of atmospheric waves in analogy to water waves.

According to *Thagard* [7.24, p. 67] analogical abduction results from a *conceptual combination*: the already possessed concepts of wave and sound are combined into the combined concept of a sound wave. I think that this early analysis of *Thagard* [7.24] is too simple. In my view, the crucial process that is involved in analogical abduction is a *conceptual abstraction* based on an *isomorphic* or *homomorphic mapping*. What is abducted by this analogy is not only the combined concept of sound wave, but at the same time the *theoretical* concept of a *wave in abstracto* (the later paper of *Holyoak* and *Thagard* [7.50] supports this view).

A clear analysis of analogy based on conceptual abstraction has been given by *Gentner* [7.51]. According

to Gentner's analysis, an analogy is a *partial isomorphic* mapping  $m$  between two relational structures, the *source* structure  $(D, (F_i : 1 \leq i \leq m), (R_i : 1 \leq i \leq n))$  and the *target* structure  $(D^*, (F_i^* : 1 \leq i \leq m^*), (R_i^* : 1 \leq i \leq n^*))$ , where the  $F_i$  are monadic predicates and the  $R_i$  are relations. Gentner argues convincingly [7.51, p. 158] that an analogical mapping preserves only the *relations* of the two structures (at least many of them, including second-order relations such as *being-a-cause-of*), while monadic properties are not preserved. This is what distinguishes an *analogy* from a *literal similarity*. For example, our solar system is literally similar to the star system X12 in the Andromeda galaxy, inasmuch as the X12 central star is bright and yellow like our sun, and surrounded by planets that are similar to our planets. Thus, our sun and the X12 star have many (monadic) properties in common. On the other hand, an atom (according to the Rutherford theory) is merely analogical to our solar system: the positively charged nucleus is surrounded by electrons just as the sun is surrounded by planets, being governed by a structurally similar force law. But concerning its monadic properties, the atomic nucleus is very different from the sun and the electrons are very different from the planets. Formally, then, an analogical mapping  $m$  maps a subset  $D'$  of  $D$  bijectively into a subset  $D'^*$  of  $D^*$ , and many (but not necessarily all) relations  $R_i$ , with  $i \in I \subset \{1, \dots, n\}$ , into corresponding relations  $R_{m(i)}^*$ , such that the following holds: For all  $a, b \in D'$  and  $R_i$  with  $i \in I$ ,  $aR_i b$  iff  $m(a)R_{m(i)}^*m(b)$ , where *iff* stands short for *if and only if*. In this sense, the Rutherford-analogy maps *sun* into *nucleus*, *planet* into *electron*, *gravitational attraction* into *electrical attraction*, *surrounding* into *surrounding*, etc. It follows from the existence of such a partial isomorphic mapping that, for every explanatory law  $L$  expressed in terms of mapping-preserved relations that hold in the  $D'$ -restricted source structure, its starred counterpart  $L^*$  will hold in the  $D'^*$ -restricted target structure. In this way, explanations can be transferred from the source to the target structure.

Every partial isomorphism gives rise to a *conceptual abstraction* by putting together just those parts of both structures that are isomorphically mapped into each other: the resulting structure  $(D', (R_i : i \in I))$ , which is determined up to isomorphism, is interpreted in an *abstract* system-theoretic sense. In this way, the

abstract model of a *central force system* arises, with a *central body*, *peripheral bodies*, a *centripetal* and a *centrifugal force* [7.51, p. 160]. So, finding an abductive analogy consists in finding the *theoretically essential* features of the source structure that can be generalized to other domains, and this goes hand-in-hand with forming the corresponding conceptual abstraction. In our example, the analogical transfer of water waves to sound waves can only work if the theoretically essential features of (water) waves have been identified, namely, that waves are produced by *coupled oscillations*. The abductive conjecture of sound waves also stipulates that sound consists of coupled oscillations of the molecules of the air. Only after the theoretical model of sound waves has been formed, does a *theoretical* explanation of the propagation and reflection of sound waves become possible.

### 7.6.3 Hypothetical (Common) Cause Abduction

This is the most fundamental kind of conceptually creative abduction. The explanandum consists either (a) in *one* phenomenon or (b) in *several* mutually *inter-correlated* phenomena (properties or regularities). One abductively conjectures in case (a) that the phenomenon is the effect of one *single* hypothetical (unobservable) cause, and in case (b) that the phenomena are effects of one *common* hypothetical (unobservable) cause. I will argue that only case (b) constitutes a scientifically worthwhile abduction, while (a) is a case of pure speculation. In both cases, the abductive conjecture postulates a *new unobservable entity* (property or kind) together with *new laws* connecting it with the observable properties, without drawing on analogies to concepts with which one is already familiar. This kind of abduction does not presuppose any background knowledge except knowledge about those phenomena that are in need of explanation. What drives hypothetical-cause abduction is the search for explanatory *unification*, usually in terms of hidden or common causes – but later on, we will also meet cases where the unifying parameters have a merely instrumentalistic interpretation. Hypothetical (common) cause abduction is such a large family of abduction patterns that we treat it separately in the next section.

## 7.7 Hypothetical (Common) Cause Abduction Continued

Salmon [7.52, pp. 213] has emphasized the importance of finding common-cause explanations for the justification of scientific realism. However, Salmon does not inform us about the crucial difference between scien-

tific common-cause abduction and speculative (cause) abduction. In the next two subsections I argue that the major criterion for this distinction is *causal unification* and (connected with this) *independent testability*.

### 7.7.1 Speculative Abduction Versus Causal Unification: A Demarcation Criterion

Ockham’s razor is a broadly accepted maxim among IBE-theorists: an explanation of observed phenomena should postulate as few new theoretical (unobservable) entities or properties as possible [7.53, pp. 97–100]. Upon closer inspection this maxim turns out to be a gradual optimization criterion, for an explanation is better, the *fewer* theoretical (*hidden*) entities that it postulates, and the *more* phenomena it explains. However, by introducing sufficiently many hidden variables one can *explain* anything one wants. Where is the borderline between *reasonably many* and *too many* hidden variables postulated for the purpose of explanation? Based on Schurz [7.1, p. 219], [7.54, pp. 246], [7.40, pp. 112] I suggest the following demarcation criterion.

DC (Demarcation Criterion) for 2nd-Order Abduction:

The introduction of *one* new theoretical variable (entity or property) merely for the purpose of explaining *one* phenomenon is always *speculative* and post facto, i. e., has no independently testable empirical consequences. Only if the postulated theoretical variable explains *many intercorrelated* but analytically independent phenomena, and in this sense yields a *causal* or *explanatory unification*, is it a legitimate scientific abduction that is independently testable and, hence, worthy of further investigation.

Let us explain the criterion (DC) by way of examples. The simplest kind of speculative abduction *explains* every particular phenomenon *P* by a special power who (or which) has caused this phenomenon. In what follows, read  $\psi(\varphi)$  as *some power of kind  $\psi$  intends that  $\varphi$  happens*, where the formula  $\varphi$  may either express a singular fact or an empirical regularity or law that is frequently clothed in the form of an empirical disposition. Accordingly, we have two kinds of speculative abductions – see Table 7.2.

Speculative abductions have been performed by our human ancestors since the earliest times. All sorts of unexpected events can be pseudo-explained by speculative fact-abductions. They do *not* achieve unification, because for every event (*E*) a special hypothetical wish of God ( $\psi(E)$ ) has to be postulated [7.55, p. 86]. For the same reason, such pseudo-explanations are entirely *post hoc* and don’t entail any empirical predictions by which they could be independently tested.

Speculative law-abductions were especially common in the Middle Ages: every special healing capacity of a certain plant (etc.) was attributed to a special power that God had implanted in nature for human benefit. The example of the *virtus dormitiva* was ironically employed by Molière, and famous philosophers have used it as a paradigm example of a pseudo-explanation [7.56, Book 5, Ch. 7, Sect. 2], [7.57, Ch. 6, Sect. 2]. Speculative law-abductions violate Ockham’s principle since we have already a sufficient cause for the disposition to make one sleepy, namely the natural kind *opium*, so that the postulated power amounts to a redundant multiplication of causes. More formally, the schema does not offer unification because for every elementary empirical law one has to introduce two elementary hypothetical laws to explain it [7.55, p. 87]. For the same reason, the abductive conjecture has no predictive power that goes beyond the predictive power of the explained law.

I do not want to diminish the value of cognitive speculation by this analysis. Humans have an inborn instinct to search for causes [7.58, Ch. 3], and cognitive speculations are the predecessor of scientific inquiry. However, it was pointed out in Sect. 7.1 that the best available *explanations* are often not good enough to count as rationally acceptable. The above speculative abduction patterns can be regarded as the *idling* of our inborn explanatory search activities when applied to events for which a proper explanation is out of reach.

In contrast to these empty causal speculations, scientific common cause abductions have usually led to

Table 7.2

Speculative Fact-Abduction :	Example:
Explanandum E: $Ca$	John got a cold.
=====	
Conjecture H: $\psi(Ca) \wedge \forall \varphi(\psi(\varphi) \rightarrow \varphi)$	God wanted John to get a cold, and whatever God wants, happens.
Speculative Law-Abduction:	Example:
Explanandum E: $\forall x(Ox \rightarrow Dx)$	Opium has the disposition to make people sleepy (after consuming it).
=====	
Conjecture H: $\forall x(Ox \rightarrow \psi(Dx)) \wedge \forall x(\psi(Dx) \rightarrow Dx)$	Opium has a special power (a <i>virtus dormitiva</i> ) that causes its disposition to make people sleepy.

genuine theoretical progress. The leading principle of causal unification is the following.

(CC) Causal Connection Principle: If two properties or kinds of events are probabilistically dependent, then they are *causally connected* in the sense that either one is a cause of the other (or vice versa), or both are effects of a common cause (where  $X$  is a cause of  $Y$  iff a directed path of cause-effect relations leads from  $X$  to  $Y$ ).

The causal connection principle (CC) does not entail that every phenomenon must have a sufficient cause – it merely says that all correlations result from causal connections. This principle has been empirically corroborated in *almost every* area of science, in the sense that conjectured common causes have been identified in later stages of inquiry; the only known exception is quantum mechanics. Here we treat (CC) not as a *dogma*, but as a meta-theoretical principle that *guides* our causal abductions. (CC) is a consequence of the more general *causal Markov* condition, which is the fundamental axiom of the *theory of causal nets* [7.59, pp. 396], [7.40, p. 16], [7.60, pp. 29]. Schurz and Gebharter [7.61, Sect. 2] demonstrate that the causal Markov condition can itself be justified via an explanatory abduction inasmuch as it yields the best and only plausible explanation of two (in)stability properties of statistical correlations: screening-off and linking-up.

The way that the causal connection principle leads to common-cause abduction is as follows: Whenever we encounter several *intercorrelated phenomena*, and – for some reason or other – we can *exclude* that one causes the other(s), then (CC) requires that these phenomena must have some (unobservable) common cause that simultaneously explains all of them. The most important scientific example of this sort is common cause abduction from *correlated dispositions*: since dispositions cannot cause other dispositions, their correlations must have a common intrinsic cause.

### 7.7.2 Strict Common-Cause Abduction from Correlated Dispositions and the Discovery of New Natural Kinds

In this section I analyze common-cause abduction in a *deductivistic* setting, which is appropriate when the domain is ruled by *strict* laws. Probabilistic generalizations are treated in Sect. 7.7.3. Recall the schema of speculative law-abduction, where *one* disposition  $D$  occurring in one (natural) kind  $F$  was pseudo-explained by a causal *power*  $\psi(D)$ . In this case of a *single* disposition, the postulate of a causal power  $\psi(D)$  that mediates between  $F$  and  $D$  is an unnecessary multiplication of causes. But in the *typical* case of a scientifically productive common-cause abduction, we have several (natural) kinds  $F_1, \dots, F_n$  all of which have a set of characteristic dispositions  $D_1, \dots, D_m$  in common – with the result that all these dispositions are correlated. Assuming that it is excluded that one disposition can cause another one, then by principle (CC) these correlated dispositions must be the common effects of a certain intrinsic structure that is present in all of the kinds  $F_1 \dots, F_n$  as their common cause. For example, the following dispositional properties are common to certain substances such as *iron, copper, tin, etc.* (Fig. 7.6): a characteristic glossing, smooth surface, characteristic hardness, elasticity, ductility, high conductivity of heat and of electricity. Already before the era of modern chemistry, craftsmen had abducted that there exists a characteristic intrinsic property of substances that is the common cause of all these (more-or-less strictly) correlated dispositions, and they called it *metallic character*  $Mx$ . To be sure, the natural kind term *metal* of premodern chemistry was theoretically hardly understood. But the introduction of a new (theoretical) natural kind term is the first step in the development of a *new research program*. The next step was then to construct a *theoretical model* of the postulated kind *metal*, by which one can give an explanation of *how* the structure of a *metal* can cause all these correlated

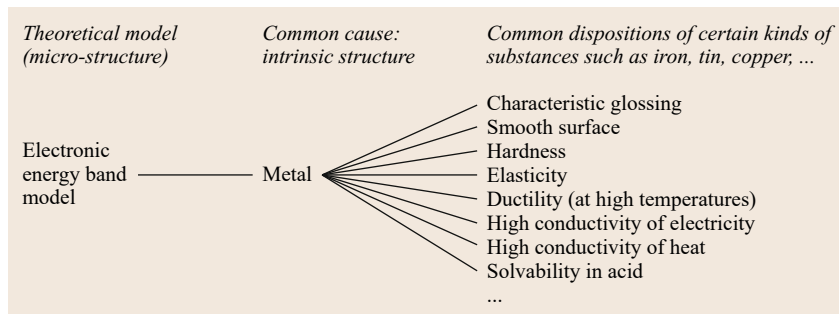


Fig. 7.6 Common-cause abduction of the chemical kind term *metal*

dispositions at once. In combination with *atomic* (and molecular) hypotheses the abducted natural kind terms of chemistry became enormously fruitful. In modern chemistry, the molecular microstructure of metals is modeled as an *electron band* of densely layered energy levels among which the electrons can easily shift around [7.62, pp. 708].

The structural pattern of the example in Fig. 7.6 can be formalized as in Tab. 7.3. The abductive conjecture *H* logically implies the explanandum *E* and yields a *unification* of  $n \cdot m$  empirical (elementary) laws by  $n + m$  theoretical (elementary) laws, which is a *polynomial reduction* of elementary laws. At the same time, *H* generates novel empirical consequences by which it can be *independently tested*, as follows. *H* does not only postulate the theoretical property  $\psi x$  to be a merely *sufficient* cause of the dispositions; it also assumes that these dispositions are an *empirical indicator* of the theoretical property  $\psi x$  [7.54, p. 111]. This indicator relation may either be reconstructed in a strict form (as indicated in brackets [ $\leftrightarrow$ ]), or in a weaker probabilistic form. In either case, if we know for some new kind  $F^*$  that it possesses *some* of the dispositions, then the abducted common-cause hypothesis predicts that  $F^*$  will also possess all the other dispositions. This is a novel (qualitatively new) prediction. For example, we can predict solubility in acid for a new kind of metallic ore, even if this ore has never been put into acid before. Novel predictions are a characteristic virtue of genuine scientific theories that go beyond simple empirically inductive generalizations [7.54, p. 112].

Having described the basic principles and mechanisms of common-cause abduction, we must add four important clarifications:

1. *Instrumentalist versus realist interpretations:* In common-cause abduction, a new theoretical property is postulated or *discovered* that accomplishes a unified explanation of correlated empirical phenomena. Even if the theory-core describing this property it is independently confirmed in later experiments, there is no guarantee that the hypothesized theoretical property has *realistic* refer-

ence. Also a purely *instrumentalistic* interpretation of a theoretical concept is possible, as a useful means of unifying empirical phenomena (which is defended, for example, by *van Fraassen* [7.63]). However, the more empirically successful a theory becomes, the more plausible it is to assume that the theoretical concept producing this success actually *does* refer to something real [7.64, Sect. 6.2].

2. *Addendum on dispositions:* I understand dispositions as *conditional* (or *functional*) properties: That an object  $x$  has a disposition  $D$  means that whenever certain initial conditions  $C$  are (or would be) satisfied for  $x$ , then a certain reaction  $R$  of  $x$  will (or would) take place. This understanding of dispositions is in accordance with the *received view* [7.65, p. 44], [7.66]. Dispositional properties are contrasted with *categorical* properties, which are not defined in terms of conditional effects, but in terms of *occurrent* intrinsic structures or states (in the sense of *Earman* [7.67, p. 94]). Dispositional properties can have categorical properties such as molecular structures as their *causal basis*, but they are *not identical* with them. Although the received view is not uncontroversial [7.68, 69], it is strongly supported by the situation that underlies common-cause abduction (Fig. 7.7): Here *different* dispositions have the *same* molecular structure as their common cause; so they cannot be identical with this molecular structure.

*Prior et al.* [7.66, p. 255] argue that since dispositions are *functional* properties, they can only be the effects of (suitable) categorical causes, but cannot themselves act as causes. If this argument is not convincing enough, here is a more detailed argument showing why, at least normally, dispositions cannot cause each other. Assume  $D_1$  and  $D_2$  are two correlated dispositions, each of them being equivalent with an empirical regularity of the above-mentioned form:  $D_i x \leftrightarrow_{\text{def}} \forall t (C_i x t \rightarrow R_i x t)$  (for  $i \in \{1, 2\}$ ,  $t$  for the time variable). The implication expresses a cause-effect relation. Now, if  $D_1$  would cause  $D_2$ , the only possible causal reconstruction would be to assume that  $C_1$  causes  $C_2$ , which causes

Table 7.3

<b>Common-cause abduction (abducted theoretical concept: <math>\psi</math>).</b>
<i>Explanandum E:</i> All kinds $F_1, \dots, F_n$ have the dispositions $D_1, \dots, D_m$ in common. $\forall i \in \{1, \dots, n\} \forall j \in \{1, \dots, m\} : \forall x (F_i x \rightarrow D_j x)$ .
=====
<i>Abductive conjecture H:</i> All $F_1s, \dots, F_ns$ have a common intrinsic and structural property $\psi$ that is a cause and an indicator of all the dispositions $D_1, \dots, D_m$ . $\forall i \in \{1, \dots, n\} : \forall x (F_i x \rightarrow \psi x) \wedge \forall j \in \{1, \dots, m\} : \forall x (\psi x \rightarrow [\leftrightarrow] D_j x)$ .

$R_2$ , which in turn causes  $R_1$  ( $C_1 \rightarrow C_2 \rightarrow R_2 \rightarrow R_1$ ). But normally this is impossible, since the initial conditions are freely manipulable and, hence, causally independent. For example, my decision to irradiate a substance with light (to test for its glossing) can in no way cause my decision to heat it (to test for its ductibility).

A final remark: When I speak of a molecular structure as being the *cause* of a disposition, I understand the notion of *cause* in a more general sense than the narrow notion of event causation. This extended usage of *cause* is *reducible* to the notion of event causation as follows: A disposition  $Dx$ , being defined as the conditional property  $\forall t(Cxt \rightarrow Rxt)$ , is *caused* by a categorical property  $\psi x$  iff each manifestation of the disposition's reaction,  $Rxt$ , is caused by  $\psi x$  together with the initial conditions  $Cxt$ , or formally, iff  $\forall x\forall t(Lx \wedge Cxt \rightarrow Rxt)$ .

3. The precise explication of causal unification – many *effects* explained by one or just a few *causes* – presupposes formal ways of *counting* elementary phenomena, expressed by elementary statements. There are some technical difficulties involved with this. Solutions to this problem have been proposed in Schurz [7.70], Gemes [7.71], and more recently in Schurz and Weingartner [7.72]. The following explication is sufficient for our purpose: We represent every given belief system (or set of statements)  $K$  by the set of all those elementary statements  $S$  of the underlying language that are *relevant* consequences of  $K$ , in the sense that no predicate in  $S$  is replaceable by another arbitrary predicate (of the same place-number), salva validitate of the entailment  $K \parallel -S$ . Here, a statement  $S$  is called *elementary* iff  $S$  is not logically equivalent to a nonredundant conjunction of statements  $S_1 \wedge \dots \wedge S_n$ , each of which is shorter than  $S$ . Borrowing the terminology from Gemes [7.71], we call the elementary relevant consequences of  $K$   $K$ 's *content elements*. It can be shown that every set of sentences is classically equivalent with the set of its content elements; so no information is lost by this representation [7.72, Lemma 7.2]. But note that our analysis of common-cause abduction does *not depend* on this particular representation method; it merely depends on the assumption that a natural method of decomposing the classical consequence class of a belief system  $K$  into a nonredundant set of *smallest* content elements exists.

Many more examples of common-cause abduction in the natural sciences can be given. For example, Glauber's discovery of the central chemical concepts of *acids*, *bases*, and *salts* in the 17th century was based

on a typical common-cause abduction [7.49, pp. 196]. The fundamental common-cause abduction of Newtonian physics was the abduction of the *gravitational force* as the common cause of the disposition of bodies on the Earth to fall and the disposition of the planets to move around the Sun in elliptic orbits. Here, Newton's qualitative stipulation of the gravitational force as the counterbalance of the centrifugal force that acts on the circulating planets was his *abductive* step, while his quantitative calculation of the mathematical form of the gravitational law was a deduction from Kepler's third law plus this abductive conjecture [7.73, p. 203]. Another example is the abduction of the goal(s) of a person as the common cause of her behavior under various conditions. Prendinger and Ishizuka [7.25, p. 324] have utilized common-cause abduction in *automated web usage mining* to infer the interests of Internet users based on their browsing activities. Kornmesser [7.74] shows that common-cause abduction was the leading principle in the development of the *principles and parameters approach* in theories of generative grammar.

Common-cause abduction can also be applied to ordinary, nondispositional properties or (kinds of) events that are correlated. However, in this case one has first to consider more parsimonious causal explanations that do not postulate an unobservable common cause but stipulate one of these events or properties to be the cause of the others. For example, if the three kinds of events  $F$ ,  $G$ , and  $H$  (for example, eating a certain poison, having difficulties in breathing, and finally dying) are strictly correlated and always occur in the form of a temporal chain, then the most parsimonious conjecture is that these event types form a causal chain. Only in the special case where two (or several) correlated event types, say  $F$  and  $G$ , are strongly correlated, but we know that there *cannot* be a direct causal mechanism that connects them, is a common-cause abduction the most plausible conjecture. An example is the correlation of lightning and thunder: we know by induction from observation that light does not produce sound, and hence, we conjecture that there must be a common cause of both of phenomena.

We finally discuss our demarcation criterion (DC) in the light of *Bayesian confirmation* theory. According to criterion (DC), a speculative hypothetical cause explanation can *never* be regarded as confirmed by the evidence; only a common-cause explanation can. A Bayesian would probably object that our demarcation between single- and common-cause abduction is just a matter of degree [7.75, p. 141]. Recall from Sect. 7.1 that a given piece of evidence  $E$  raises the probability of *every* hypothesis  $H$  that increases  $E$ 's probability (since by Bayes' theorem,  $P(H|E)/P(H) =$

$P(E|H)/P(E)$ ). So according to standard Bayesian notions of confirmation [7.76] the evidence  $E$ : *John got a cold* does indeed confirm the speculative post facto speculation  $H_{\text{spec}}$ : *God wanted John to get a cold* – just to a minor degree in comparison to the scientific hypothesis  $H_{\text{sci}}$ : *John was infected by a virus*. So it seems that our demarcation criterion is in conflict with Bayesian confirmation theory.

Schurz [7.77] suggests a way of embedding (DC) into the Bayesian perspective. In all cases of *post facto* explanations, the hypothesis  $H$  results from fitting a latent (unobserved) first- or second-order variable  $X$  to the observed evidence  $E$ . So  $H$  entails the more general background hypothesis  $\exists XH(X)$  in which the latent variable  $X$  is unfitted and existentially generalized. In our example,  $\exists XH_{\text{spec}}(X)$  says: *There exists a God who wants some  $X$ , and whatever God wants, happens*.  $H_{\text{spec}}$  results from  $\exists XH_{\text{spec}}(X)$  by replacing  $X$  by *John got a cold* and omitting the existential quantifier. Note that  $\exists XH_{\text{spec}}(X)$  is a content element of  $H_{\text{spec}}$  that transcends (is not contained in) the evidence.

Schurz [7.77] argues that the probability-raising of  $H_{\text{spec}}$  by  $E$  is a case of *pseudo*-confirmation and not of genuine confirmation, because the probability increase of  $H_{\text{spec}}$  by  $E$  does not spread to  $H_{\text{spec}}$ 's evidence-transcending content element  $\exists XH_{\text{spec}}(X)$ . This follows from the fact that  $\exists XH_{\text{spec}}(X)$  can be fitted to *every possible* piece of evidence whatsoever. Therefore the probability of  $\exists XH_{\text{spec}}(X)$  remains as low as it was before conditionalization, for arbitrarily many pieces of evidence, whence the posterior probability of  $H_{\text{spec}}$  (which entails  $\exists X_{\text{spec}}H(X)$ ) also remains low.

Also note that the scientific hypothesis  $H_{\text{sci}}$  contains a latent variable  $X$  that has been fitted to the evidence  $E$ . In our example the unfitted hypothesis  $\exists XH_{\text{sci}}(X)$  says that *every disease is caused by some pathogenic agent  $X$* , which is fitted to John's cold by replacing  $X$  by *a virus*. In this case, however,  $H_{\text{sci}}$  implies further empirical consequences  $E'$  (e.g., that John's immune system will contain characteristic antibodies) by which it can independently tested. If such independent evidence  $E'$  raises the probability of  $H_{\text{sci}}$  as well, this is no longer the result of a post facto fitting, but an instance of *genuine* confirmation, because now this probability increase spreads to  $H_{\text{sci}}$ 's evidence-transcending content element  $\exists X_{\text{sci}}H(X)$  [7.77, Sect. 4].

### 7.7.3 Probabilistic Common-Cause Abduction and Statistical Factor Analysis

Statistical factor analysis is an important branch of statistical methodology whose analysis (according to my knowledge) has been neglected by philosophers of sci-

ence (with the exception of Haig [7.78], who shares my view of factor analysis). In this section I want to show that factor analysis is a certain generalization of hypothetical common-cause abduction, although sometimes it may be better interpreted in a purely instrumentalistic way. For this purpose, I assume that scientific concepts are represented as statistical random variables  $X, Y, \dots$ , each of which can take several values  $x_i, y_j$ . (A random variable  $X: D \rightarrow \mathbb{R}$  assigns to each individual  $d$  of the domain  $D$  a real-valued number  $X(d)$ ; a dichotomic property  $Fx$  is coded by a binary variable  $X_F$  with values 1 and 0.) The variables are assumed to be at least interval-scaled, and the statistical relations between the variables are assumed to be monotonic – the *linearity* assumption of factor analysis yields good approximations only if these conditions are satisfied.

Let us start from the example of the previous section, where we have  $n$  empirically measurable and highly intercorrelated variables  $X_1, \dots, X_n$ , i.e.,  $\text{cor}(X_i, X_j) = \text{high}$  for all  $1 \leq i, j \leq n$ . An example would be the scores of test persons on  $n$  different intelligence tests. We assume that none of the variables screens off the correlations between any other pair of variables (i.e.,  $\text{cor}(X_i, X_j|X_r) \neq 0$  for all  $r \neq i, j$ ), which makes it plausible that these  $n$  variables have a common cause, distinct from each of the variables – a theoretical factor, call it  $F$ . In our example,  $F$  would be the theoretical concept of intelligence. Computationally, the abductive conjecture asserts that for each  $1 \leq i \leq n$ ,  $X_i$  is approximated by a linear function  $f_i$  of  $F$ ,  $f_i(F(x)) = a_i F(x)$ , for given individuals  $x$  in the domain  $D$  (since we assume the variables  $X_i$  to be  $z$ -standardized, the linear function  $f_i$  has no additive term  $+b_i$ ). The true  $X_i$ -values are scattered around values predicted by this linear function,  $f_i(F)$ , by a remaining random dispersion  $s_i$ ; the square  $s_i^2$  is the *remainder variance*. According to the standard linear regression technique, the optimally fitting coefficients  $a_i$  are computed so as to *minimize* this remainder variance. Visually speaking, the  $X_i$ -values form a stretched cloud of points in an  $n$ -dimensional coordinate system, and  $F$  is a straight line going through the middle of the cloud such that the squared normal deviations of the points from the straight line are minimized.

So far we have described the linear-regression statistics of the abduction of *one* factor or cause. In factor analysis one also takes into account that the mutually intercorrelated variables may have not only one but *several* common causes. For example, the variables may divide into two subgroups with high correlations within each subgroup, but low correlations between the two subgroups. In such a case the reasonable abductive conjecture is that there are two independent common causes  $F_1$  and  $F_2$ , each responsible for the variables in one of the two subgroups. In the general picture of



factor analysis, there are  $n$  given empirical variables  $X_i$ , which are explained by  $k < n$  theoretical factors (or common causes)  $F_j$  as follows (for the following [7.79, Ch. 3]; note that I can describe here only the most common method of factor analysis without discussing subtle differences between different methods):

$$\begin{aligned} X_1 &= a_{11}F_1 + \dots + a_{1k}F_k + s_1 \\ &\dots \\ X_n &= a_{n1}F_1 + \dots + a_{nk}F_k + s_n. \end{aligned}$$

This is usually written in a matrix formulation:  $\mathbf{X} = \mathbf{F} \cdot \mathbf{A}'$ . While each variable  $X_i$  and factor  $F_j$  takes different values for the different individuals of the sample, the factor loadings  $a_{ij}$  are constant and represent the causal contribution of factor  $F_j$  to variable  $X_i$ . Given the further assumption that the factor variables  $F_j$  are standardized, each *factor loading*  $a_{ij}$  expresses the correlation between variable  $X_i$  and factor  $F_j$ ,  $\text{cor}(X_i, F_j)$ . Since the variance of each variable  $X_i$  equals the sum of the squared factor loadings  $a_{ij}^2$  and the remainder variance  $s_i$ , each squared factor loading  $a_{ij}^2$  measures the *amount of the variance of  $X_i$  explained* (i. e., statistically predicted) by factor  $F_j$ . The sum of the squared loadings of a factor  $F_j$ ,  $\sum_{1 \leq i \leq n} a_{ij}^2$ , measures the amount of total variance of the variables which is explained by  $F_j$ , and the sum of all of the squared loadings divided through  $n$  equals the percentage of variance explained by the extracted factors, which is a measure for the explanatory success of the factor-statistical analysis.

The major mathematical technique to find those  $k < n$  factors that explain a maximal amount of the total variance is the so-called principal component analysis. Instead of providing a detailed mathematical explanation, I confine myself to the following remarks. The  $k$  factors or axes are determined according to two criteria:

- (i) They are probabilistically independent (or orthogonal) to each other
- (ii) The amount of explained variance is maximized (i. e., the remainder variances are minimized).

Visually speaking, the first factor  $F_1$  is determined as an axis going through the stretched cloud of points in the  $n$ -dimensional coordinate system; then the next factor  $F_2$  is determined as an axis orthogonal to  $F_1$ , and so on, until the  $k < n$  factor axes are determined by the system of coefficients  $a_{ij}$ .

The success of an explanation of  $n$  variables by  $k < n$  factors is greater, the less the number  $k$  compared to  $n$ , and the higher the amount of the total variance explained by the  $k$  factors. This fits perfectly with my account of unification of a given set of  $n$

empirical variables by a small set of  $k$  theoretical variables, as explained in Sect. 7.7.1. While the amount of explained variance of the first factor is usually much greater than 1, this amount becomes smaller and smaller when one introduces more and more factors (in the trivial limiting case  $k = n$  the amount of explained variance becomes 100%). According to the Kaiser–Guttman criterion one should introduce new factors only as long as their amount of explained variance is greater than 1 [7.79, p. 75]. Hence, a theoretical factor is only considered nontrivial if it explains more than the variance of just one variable and, in this sense, offers a unificatory explanation to at least *some* degree. This is the factor analytic counterpart of my suggested demarcation criterion for hypothetical-cause abduction (DC).

We have seen in Sect. 7.7.2 that not only realistic but also instrumentalistic interpretations of the abducted factors are possible. In fact, several statisticians tend to interpret the results of a factor analysis cautiously as a merely instrumentalistic means of *data reduction* in the sense of representing a large class of intercorrelated empirical variables by a small class of independent theoretical variables. In spite of this fact, I think that the intended interpretation of the factors of a factor analysis is their realistic interpretation as common causes, for that is how they are designed. I regard the instrumentalistic perspective as an *important warning* that not every empirically useful theoretical structure must correspond to an existing structure of reality.

### 7.7.4 Epistemological Abduction to Reality

The relevance of abduction for realism is usually discussed within the context of theories and theoretical-entity realism. For many epistemologists [7.4, p. 44], [7.21, Chap. IV.5–6], the fundamental problem of common sense realism – the reasoning from introspective sense data to an external reality causing these perceptions – is an inference *sui generis*, and its justification is a problem of its own. In contrast to this position, I wish to point out that reasoning from introspective sense data to common-sense realism is in perfect fit with the pattern of common-cause abduction [7.53, p. 98]. The hypothesis of external objects that cause our sensual experience yields a common-cause explanation of a huge set of intercorrelations between our introspective experiences.

*First*, there are the *intra-sensual* intercorrelations, in particular those within our system of visual perceptions. There are potentially infinitely many 2-D visual images of a perceptual object, but all these 2-D images are strictly correlated with the position and angle at which we look at that object; so these cor-

relations have a common-cause explanation in terms of three-dimensional external objects by the laws of the perspectival projection. To be sure, these common-cause abductions are mainly *unconscious* and rely on inborn computations performed by the visual cortex of our brains. What we consciously experience are the *abducted* three-dimensional objects that make up the mind of the *naive realist*. However, certain situations – for example, the case of visual illusions caused by 3-D pictures – make it plain that what underlies our three-dimensional visual appearances is a complicated abductive computational process [7.80]. Moreover, since in our ordinary visual perceptions some objects partly conceal other objects that are behind them, our visual abductions always include the task of Gestalt complementation. Identification of three-dimensional objects based on two-dimensional projective images is an important abductive task in the AI field of visual object recognition [7.81, Ch. 24.4]. Scientifically advanced versions of visual abduction where one abducts the shape of entire objects from sparse fragments have been analyzed in the field of archeology [7.82].

The *inter-sensual* correlation between different sensual experiences, in particular between visual perceptions and tactile perceptions, is the *second* important

basis for the unconscious abduction to an outer reality – in fact, these correlations seem even to be the major fundament of our naive belief in the outer reality. If you have a visual appearance of an object, but you are unsure whether it is a mere visual illusion or not, then you will probably go to the object and try touch it – and if you can, then your realistic desires are satisfied. On the other hand, visual appearances that do not correspond to tactile ones, so-called *ghosts*, have frightened the naively realistic mind and occupied its fantasy since the earliest times.

This concludes my analysis of patterns of abduction. Instead of a conclusion, I refer to the classification of abduction patterns in Fig. 7.1, and to my main theses and results as explained in Sect. 7.1, which are densely supported by the details of my analysis. As a final conclusion, I propose the following: As Peirce once remarked [7.20, CP 6.500], the success of scientists at finding true hypotheses among myriads of possible hypotheses seems to be a sheer miracle. I think that this success becomes much less miraculous if one understands the strategic role of patterns of abduction. In the concluding section, I present applications of abductive reasoning in two neighboring fields: belief revision and instrumental or technological reasoning.

## 7.8 Further Applications of Abductive Inference

### 7.8.1 Abductive Belief Revision

The theory of belief revision formulates rules for the rational change of a given belief system (*knowledge*)  $K$  upon receiving new evidence.  $K$  is understood as a set of (believed) sentences that is closed under deductive consequence ( $K = Cn(K)$ ) and the pieces of evidence  $E_i$  are usually (though not always) taken as certainly true (the so-called *axiom of success*). According to the well-known AGM-theory of belief revision (after Alchourrón, Gärdenfors and Makinson [7.83, 84]), the rational change of a belief system  $K$  upon receiving new evidence  $E$  is characterized as follows:

1. If  $E$  is consistent with  $K$ , this change is called the *expansion* of  $K$  by  $E$ , abbreviated as  $K + E$ , and defined as  $K + E =_{\text{def}} Cn(K \cup \{E\})$ , i. e., the logical closure of the union of  $K$  and  $\{E\}$ .
2. If  $E$  is inconsistent with  $K$  (i. e.,  $K$  entails  $\neg E$ ), this change is called the *revision* of  $K$  by  $E$ , abbreviated as  $K * E$  and defined as  $K * E =_{\text{def}} Cn((K - \neg E) + E)$ , where  $K - \neg E$  is called the *contraction* of  $K$  by  $\neg E$  and intuitively characterized as a *minimal* yet *reasonable* subset of  $K$  that is consistent with  $E$ .

Thus,  $K * E$  is the closure of the expansion of  $K$ 's contraction by  $\neg E$  with  $E$ ; this definition is called the *Levi identity*. There is no consensus about a *most reasonable* or *right* contraction operation in the literature.

A general deficiency of standard definitions of belief expansion and revision is their failure to represent learning on the basis of evidence. This is a consequence of the idea that the expansion or revision should always be as *minimal* as possible [7.85, p. 80]. An alternative is *abductive* belief expansion and revision. It rests on the idea that the expansion or revision of  $K$  by new evidence  $E$  should be the result of adding  $E$  to  $K$  and abducting a suitable explanation of  $E$  (assuming that such an explanation is not already present in  $K$ ).

In what follows, the abductive belief expansion or revision of a belief system  $K$  by new evidence  $E$  is abbreviated as  $K ++ E$  or  $K * * E$ , respectively. A first approach of this sort was initiated by Pagnucco [7.86], who characterized the space of abductible explanations in a maximally liberal way: an abduction (explanation) of  $E$  given  $K$  may be *any* sentence  $S$  such that  $S$  entails  $E$  within  $K$ ; in the most trivial case,  $S$  is identical with  $E$ . Thus, Pagnucco defines an abductive belief

expansion operation as any operation  $++$  satisfying  $K ++ E = Cn(K \cup \{S\})$  for some sentence  $S$  such that  $K \cup \{S\}$  is consistent and entails  $E$  (provided  $K \cup \{E\}$  is consistent). Belief revision is then defined by the Levi identity  $K * * E =_{\text{def}} (K - \neg E) ++ E$ . A similar approach to abductive belief revision is suggested in *Aliseda* [7.7, pp. 184].

Schurz [7.85, pp. 88] argues that from the viewpoint of philosophy of science, Pagnucco's notion of abduction is too weak, since not just any sentence that logically entails  $E$  constitutes a scientific explanation of  $E$ . Schurz [7.85] utilizes the abduction patterns classified in Fig. 7.1 to explicate corresponding operations of abductive belief expansion and revision. He defines abductive belief expansion as  $K ++ E =_{\text{def}} K \cup \{E\} \cup \text{abd}(K, E)$ , where the triple  $(E, \text{abd}(E, K), K)$  expresses one of the abduction situations classified in Fig. 7.1:  $E$  is the evidence to be explained,  $\text{abd}(E, K)$  is the abductive conjecture, and  $K$  is the given background belief system that drives the abduction [7.85, p. 93]. Next, Schurz [7.85, pp. 94–6] discovers that an appropriate definition of abductive belief revision fails to meet the Levi identity for two reasons, which he calls the problems of *old evidence* and *incremental belief revision*. He defines abductive belief revision based on a suitable notion of the abductive revision of an explanatory hypothesis by contradicting evidence [7.85, p. 96]. Operations of intelligent abductive belief revision that are congenial in spirit have been implemented in computer programs by Bharathan and Josephson [7.87].

A related account to abductive belief revision has been developed by Cevolani [7.88]. Cevolani agrees with Schurz that the characterization of an explanation as any sentences that entails the explanandum is too weak. He suggests that the abductive hypothesis  $S$  that in the given belief system  $K$  explains  $E$  should be such that the *expected verisimilitude* of the resulting belief systems  $K ++ E$  or  $K * * E$ , respectively, increases. Similar to Schurz, Cevolani observes that his so-defined notion of abductive belief revision fails to satisfy the Levi identity [7.88, p. 011].

## 7.8.2 Instrumental Abduction and Technological Reasoning

Abductions in the standard meaning of this word are *explanatory* in the sense that it is the task of the abductive hypothesis to *explain* a given phenomenon. All patterns of abductive inference that are classified in Fig. 7.1 and discussed in Sects. 7.2–7.7 are explanatory abductions. Tomiyama et al. [7.89, 90] and Tuzet [7.91, p. 152] have proposed to extend the meaning of the notion of *abduction* so that it also includes technological or, more generally, any sort of instrumental reasoning.

In instrumental *abduction*, the proposition that the abductive hypothesis entails in the given background system is not yet true, but rather represents a *goal*, i. e., something that one wishes to realize, and the abductive hypothesis (*conclusion*) expresses a conjectured *means* to *realize* this goal. Referring to the patterns of abductions classified in a predecessor paper of Schurz [7.1], Tomiyama et al. [7.89] show how abductive reasoning can be applied to create the design of a refrigerator. For Tuzet [7.91, p. 152], the reasoning schema *We want E. If C then E. Therefore, we should try to bring about C* expresses the basic form of instrumental abduction, constrained by the restriction that  $C$  should be practically *appropriate*: for example  $C$  must be practically realizable and must not have unwanted side-effects (etc.).

Generally speaking, often – though not always – instrumental reasoning proceeds by similar cognitive operations as abductive reasoning, the only difference being that the *abductandum* is not yet realized but expresses a goal. Therefore I propose to call this sort of reasoning *instrumental abduction*, provided one is aware that this notion *extends* the proper meaning of abduction to a different domain of application. (Note that Tuzet [7.91] speaks of *epistemic* versus *projectual* abduction; I prefer the designations *explanatory* versus *instrumental* abduction because they are more specific.) The distinction between explanatory and instrumental abduction is summarized in Fig. 7.7.

In Fig. 7.7 the superordinate concept that covers explanatory as well as instrumental abduction is called

*Generalized abduction:*

$C$ : Abductandum, conclusion to be inferred

$K$ : System of background beliefs

$P$ : Abductans, abducted conjecture such that  $\{P\} \cup K$  entails  $C$

*Ordinary explanatory abduction*

$C$  is an observed (true) phenomenon

$P$  is an optimal explanation of  $C$  given  $K$ .

*Instrumental abduction*

$C$  expresses a goal to be realized

$P$  expresses an optimal means for realizing  $P$  given  $K$ .

**Fig. 7.7** Generalized, explanatory and instrumental abduction

*generalized abduction*. What is common to both forms of abduction is a process that searches for missing premises  $P$  for a given conclusion  $C$ . Gabbay and Woods [7.17, p. 191] and Woods [7.92, p. 153] go so far as to call every cognitive process an abduction so long as it generates a premise  $P$  from which a given sentence  $C$  can be derived or obtained in the given background system  $K$  (thereby the authors generalize the notion of *consequence* to an arbitrary *closure relation*  $R$ ). According to this view, a process that *confirms*  $C$ , or that *predicts*  $C$  (via finding suitable premises), would also be called an abduction. From the viewpoint of philosophy

of science, I am inclined to think that this generalization overstretches the notion of abduction. On the other hand, from a logical viewpoint it seems to make sense to call any process that searches for premises in order to infer a given *inference goal* an abduction in the *logically generalized* sense.

**Acknowledgments.** For valuable help I am indebted to Ilkka Niiniluoto, Theo Kuipers, Gerhard Brewka, Gustavo Cevolani, Lorenzo Magnani, Helmut Prendinger, Tetsuo Tomiyama, and Erik Olsson.

## References

- 7.1 G. Schurz: Patterns of abduction, *Synthese* **164**, 201–234 (2008)
- 7.2 C.S. Peirce: Deduction, induction, and hypothesis. In: *Elements of Logic*, Collected Papers of Charles Sanders Peirce, Vol. 2, (Harvard Univ. Press, Cambridge 1932), ed. by C. Hartshorne, P. Weiss (2.619–2.644)
- 7.3 C.S. Peirce: Lectures on pragmatism. In: *Pragmatism and Pragmaticism*, Collected Papers of Charles Sanders Peirce, Vol. 5, (Harvard Univ. Press, Cambridge 1934), ed. by C. Hartshorne, P. Weiss (5.14–5.212)
- 7.4 J. Pollock: *Contemporary Theories of Knowledge* (Rowman Littlefield, Maryland 1986)
- 7.5 J. Earman: *Bayes or Bust?* (MIT, Cambridge 1992)
- 7.6 J. Ladyman: *Understanding Philosophy of Science* (Routledge, London 2002)
- 7.7 A. Aliseda: *Abductive Reasoning* (Springer, Dordrecht 2006)
- 7.8 L. Magnani: *Abduction, Reason, and Science* (Kluwer, Dordrecht 2001)
- 7.9 L. Magnani: *Abductive Cognition. The Epistemological and Eco-Cognitive Dimensions of Hypothetical Reasoning* (Springer, Heidelberg, Berlin 2009)
- 7.10 G.H. Harman: The inference to the best explanation, *Philos. Rev.* **74**, 173–228 (1965)
- 7.11 P. Lipton: *Inference to the Best Explanation* (Routledge, London 1991)
- 7.12 I. Niiniluoto: Defending abduction, *Proc. Phil. Sci.*, Vol. 66 (1999) pp. S436–S451
- 7.13 E. Barnes: Inference to the loveliest explanation, *Synthese* **103**, 251–277 (1995)
- 7.14 J. Hintikka: What is abduction? The fundamental problem of contemporary epistemology, *Trans. Charles Sanders Peirce Soc.* **34**(3), 503–533 (1998)
- 7.15 J. Hintikka, I. Halonen, A. Mutanen: Interrogative logic as a general theory of reasoning. In: *Handbook of Practical Reasoning*, ed. by R.H. Johnson, J. Woods (Kluwer, Dordrecht 2000)
- 7.16 N.R. Hanson: Is there a logic of discovery? In: *Current Issues in the Philosophy of Science*, ed. by H. Feigl, G. Maxwell (Holt Rinehart Winston, New York 1961) pp. 20–35
- 7.17 D. Gabbay, J. Woods: Advice on abductive logic, *Logic J. IGPL* **14**(1), 189–220 (2006)
- 7.18 T. Day, H. Kincaid: Putting inference to the best explanation in its place, *Synthese* **98**, 271–295 (1994)
- 7.19 C.S. Peirce: *Scientific Metaphysics*, Collected Papers of Charles Sanders Peirce, Vol. 6 (Harvard Univ. Press, Cambridge 1935), ed. by C. Hartshorne, P. Weiss
- 7.20 C.S. Peirce: *Pragmatism and Pragmaticism*, Collected Papers of Charles Sanders Peirce, Vol. 5 (Harvard Univ. Press, Cambridge 1934), ed. by C. Hartshorne, P. Weiss
- 7.21 R.M. Chisholm: *Theory of Knowledge* (Prentice Hall, Englewood Cliffs, N.J 1966)
- 7.22 T.A. Sebeok, J. Umiker-Sebeok: *You Know My Method. A Juxtaposition of Charles S. Peirce and Sherlock Holmes* (Gaslight, Bloomington/Ind. 1980)
- 7.23 R.A. Fumerton: Induction and reasoning to the best explanation, *Phil. Sci.* **47**, 589–600 (1980)
- 7.24 P. Thagard: *Computational Philosophy of Science* (MIT, Cambridge 1988)
- 7.25 H. Prendinger, M. Ishizuka: A creative abduction approach to scientific and knowledge discovery, *Knowl. Based Syst.* **18**, 321–326 (2005)
- 7.26 J. Josephson, S. Josephson (Eds.): *Abductive Inference* (Cambridge Univ. Press, New York 1994)
- 7.27 P. Flach, A. Kakas (Eds.): *Abduction and Induction* (Kluwer, Dordrecht 2000)
- 7.28 G. Paul: Approaches to abductive reasoning, *Artif. Intell. Rev.* **7**, 109–152 (1993)
- 7.29 L. Console, D.T. Dupre, P. Torasso: On the relationship between abduction and deduction, *J. Logic Comput.* **1**(5), 661–690 (1991)
- 7.30 I. Bratko: *Prolog Programming for Artificial Intelligence* (Addison-Wesley, Reading/Mass 1986)
- 7.31 D. Gabbay, J. Woods: *The Reach of Abduction: Insight and Trial (A Practical Logic of Cognitive Systems)*, Vol. 2 (North-Holland, Amsterdam 2005)
- 7.32 A. Nepomuceno, F. Soler-Toscano, F. Velasquez-Quezada: An epistemic and dynamic approach to abductive reasoning: Selecting the best explanation, *Logic J. IGPL* **21**(6), 962–979 (2013)
- 7.33 J. Meheus, D. Batens: A formal logic for abductive reasoning, *Logic J. IGPL* **14**, 221–236 (2006)

- 7.34 A. Aliseda, L. Leonides: Hypotheses testing in adaptive logics: An application to medical diagnosis, *Logic J. IGPL* **21**(6), 915–930 (2013)
- 7.35 D. Walton: *Abductive Reasoning* (Univ. of Alabama Press, Tuscaloosa 2004)
- 7.36 I. Halonen, J. Hintikka: Towards a theory of the process of explanation, *Synthese* **143**(1/2), 5–61 (2005)
- 7.37 L. Magnani: Multimodal abduction, *Logic J. IGPL* **14**(1), 107–136 (2006)
- 7.38 G. Schurz: Criteria of theoreticity: Bridging statement and non statement view, *Erkenntnis* **79**(8), 1521–1545 (2014)
- 7.39 T.S. Kuhn: *The Structure of Scientific Revolutions* (Chicago Univ. Press, Chicago 1962)
- 7.40 G. Schurz: *Philosophy of Science: A Unified Approach* (Routledge, New York 2013)
- 7.41 P. Kitcher: Explanatory unification, *Phil. Sci.* **48**, 507–531 (1981)
- 7.42 B. Buchanan, G.L. Sutherland, E.A. Feigenbaum: Heuristic dendral – A program for generating explanatory hypotheses in organic chemistry, *Mach. Intell.* **4**, 209–254 (1969)
- 7.43 E. Sober: *Philosophy of Biology* (Westview, Boulder 1993)
- 7.44 M. Ridley: *Evolution* (Blackwell Scientific, Oxford 1993)
- 7.45 G. Schurz: What is normal? An evolution–theoretic foundation of normic laws, *Phil. Sci.* **28**, 476–497 (2001)
- 7.46 D. Davidson: *Inquiries into Truth and Interpretation* (Oxford Univ. Press, Oxford 1984)
- 7.47 H.P. Grice: Logic and conversation. In: *Syntax and Semantics*, Vol. 3: Speech Acts, ed. by P. Cole, J. Morgan (Academic Press, New York 1975) pp. 41–58
- 7.48 J.R. Hobbs, M. Stickel, P. Martin, D. Edwards: Interpretation as abduction, *Artif. Intell. J.* **63**(1/2), 69–142 (1993)
- 7.49 P. Langley, H.A. Simon, G.L. Bradshaw, J.M. Zytkow: *Scientific Discovery. Computational Explorations of the Creative Process* (MIT Press, Cambridge 1987)
- 7.50 K. Holyoak, P. Thagard: Analogical mapping by constraint satisfaction, *Cogn. Sci.* **13**, 295–355 (1989)
- 7.51 D. Gentner: Structure–mapping: A theoretical framework for analogy, *Cogn. Sci.* **7**, 155–170 (1983)
- 7.52 W. Salmon: *Scientific Explanation and the Causal Structure of the World* (Princeton Univ. Press, Princeton 1984)
- 7.53 P.K. Moser: *Knowledge and Evidence* (Cambridge Univ. Press, Cambridge 1989)
- 7.54 G. Schurz: When empirical success implies theoretical reference, *Br. J. Phil. Sci.* **60**(1), 101–133 (2009)
- 7.55 G.K.L. Schurz: Outline of a theory of scientific understanding, *Synthese* **101**(1), 65–120 (1994)
- 7.56 J.St. Mill: *System of Logic*, 6th edn. (Parker Son Bourn, London 1865)
- 7.57 J. Ducasse: *A Critical Examination of the Belief in a Life After Death* (Charles Thomas, Springfield 1974)
- 7.58 D. Sperber, D. Premack, A. James Premack (Eds.): *Causal Cognition – A Multidisciplinary Approach* (Clarendon, Oxford 1995)
- 7.59 J. Pearl: *Causality*, 2nd edn. (Cambridge Univ. Press, Cambridge 2009)
- 7.60 P. Spirtes, C. Glymour, R. Scheines: *Causation, Prediction, and Search*, 2nd edn. (MIT Press, Cambridge 2000)
- 7.61 G. Schurz, A. Gebharter: Causality as a theoretical concept, *Synthese* **193**(4), 1071–1103 (2014)
- 7.62 D.W. Octoby, H.P. Gillis, N.H. Nachtriello: *Principles of Modern Chemistry* (Saunders College, Orlando 1999)
- 7.63 B. Van Fraassen: *The Scientific Image* (Clarendon, Oxford 1980)
- 7.64 T.A.F. Kuipers: *From Instrumentalism to Constructive Realism* (Kluwer, Dordrecht 2000)
- 7.65 A. Pap: Disposition concepts and extensional logic. In: *Dispositions*, ed. by R. Tuomela (Reidel, Dordrecht 1978) pp. 27–54
- 7.66 E.W. Prior, R. Pargetter, F. Jackson: Three theses about dispositions, *Am. Philos. Q.* **19**, 1251–1257 (1982)
- 7.67 J. Earman: *A Primer on Determinism* (Reidel, Dordrecht 1986)
- 7.68 D.M. Armstrong: Dispositions as causes, *Analysis* **30**, 23–26 (1969)
- 7.69 S. Mumford: *Dispositions* (Oxford Univ. Press, Oxford 1998)
- 7.70 G. Schurz: Relevant deduction, *Erkenntnis* **35**, 391–437 (1991)
- 7.71 K. Gemes: Hypothetico–deductivism, content, and the natural axiomatization of theories, *Phil. Sci.* **54**, 477–487 (1993)
- 7.72 G. Schurz, P. Weingartner: Zwart and Franssen’s impossibility theorem, *Synthese* **172**, 415–436 (2010)
- 7.73 C. Glymour: *Theory and Evidence* (Princeton Univ. Press, Princeton 1981)
- 7.74 S. Kornmesser: Model–based research programs, *Conceptus* **41**(99/100), 135–187 (2014)
- 7.75 C. Howson, P. Urbach: *Scientific Reasoning: The Bayesian Approach*, 2nd edn. (Open Court, Chicago 1996)
- 7.76 B. Fitelson: The plurality of bayesian confirmation measures of confirmation, *Proc. Phil. Sci.*, Vol. 66 (1999) pp. S362–S378
- 7.77 G. Schurz: Bayesian pseudo–confirmation, use–novelty, and genuine confirmation, *Stud. Hist. Phil. Sci.* **45**, 87–96 (2014)
- 7.78 B. Haig: Exploratory factor analysis, theory generation, and scientific method, *Multivar. Behav. Res.* **40**(3), 303–329 (2005)
- 7.79 P. Kline: *An Easy Guide to Factor Analysis* (Routledge, London 1994)
- 7.80 I. Rock: *Perception* (Scientific American Books, New York 1984)
- 7.81 S.J. Russell, P. Norvig: *Artificial Intelligence* (Prentice Hall, Englewood–Cliffs 1995)
- 7.82 C. Shelley: Visual abductive reasoning in archaeology, *Phil. Sci.* **63**, 278–301 (1996)
- 7.83 C.E. Alchourrón, P. Gärdenfors, D. Makinson: On the logic of theory change, *J. Symbolic Logic* **50**, 510–530 (1985)
- 7.84 P. Gärdenfors: *Knowledge in Flux* (MIT Press, Cambridge 1988)
- 7.85 G. Schurz: Abductive belief revision. In: *Belief Revision Meets Philosophy of Science*, ed. by E. Olsson, S. Enqvist (Springer, New York 2011) pp. 77–104

- 7.86 M. Pagnucco: The Role of Abductive Reasoning within the Process of Belief Revision, Dissertation (Univ. Sydney, Sydney 1996)
- 7.87 V. Bharathan, J.R. Josephson: Belief revision controlled by meta-abduction, *Logic J. IGPL* **14**(1), 271–286 (2006)
- 7.88 G. Cevolani: Truth approximation via abductive belief change, *Logic J. IGPL* **21**(6), 999–1016 (2013)
- 7.89 T. Tomiyama, H. Takeda, M. Yoshioka, Y. Shimomura: Abduction for creative design, Proc. ASME 2003 DETC (2003), paper no. DETC2003/DTM-48650, ASME (CD-ROM)
- 7.90 T. Tomiyama, P. Gu, Y. Jin, D. Lutters, Ch. Kind, F. Kimura: Design methodologies: Industrial and educational applications, *CIRP Ann. – Manuf. Technol.* **58**(2), 543–565 (2009)
- 7.91 G. Tuzet: Projectual abduction, *Logic J. IGPL* **14**(2), 151–160 (2006)
- 7.92 J. Woods: Cognitive economics and the logic of abduction, *Rev. Symb. Logic* **5**(1), 148–161 (2012)

## 8. Forms of Abduction and an Inferential Taxonomy

Gerhard Minnameier

In recent years, the Peircean concept of *abduction* has been differentiated into different forms and made fruitful in a variety of contexts. However, the very notion of abduction still seems to be in need of clarification. The present contribution takes very seriously Peirce's claim (1) that there are only three kinds of reasoning, that is, abduction, deduction, and induction, and (2) that these are mutually distinct. Therefore, the fundamental features of the three inferences canvassed, in particular as regards inferential subprocesses and the validity of each kind of reasoning. It is also argued that forms of abduction have to be distinguished along two dimensions: one concerns *levels* of abstraction (from elementary embodied and perceptual levels to high-level scientific theorizing). The other concerns *domains* of reasoning such as explanatory, instrumental, and moral reasoning. Moreover, Peirce's notion of *theorematic deduction* is taken up and reconstructed as *inverse deduction*. Based on this, *inverse abduction* and *inverse induction* are introduced as complements of the ordinary forms. All in all, the contribution suggests a taxonomy of inferential reasoning, in which different forms of abduction (as well as deduction and induction) can be systematically accommodated. The chapter ends with a discussion on forms of abduction found in the current literature.

8.1	<b>Abduction in the Overall Inferential Context</b> .....	177
8.1.1	Disentangling Abduction and IBE .....	177
8.1.2	The Dynamical Interaction of Abduction, Deduction, and Induction	179
8.1.3	Abduction and Abstraction .....	180
8.2	<b>The Logicity of Abduction, Deduction, and Induction</b> .....	183
8.2.1	Inferential Subprocesses and Abduction as Inferential Reasoning .....	183
8.2.2	The Validity of Abduction, Deduction, and Induction .....	184
8.3	<b>Inverse Inferences</b> .....	185
8.3.1	Theorematic Deduction as Inverse Deduction .....	185
8.3.2	An Example for Theorematic Deduction.	187
8.3.3	Inverse Abduction and Inverse Induction .....	188
8.4	<b>Discussion of Two Important Distinctions Between Types of Abduction</b> .....	189
8.4.1	Creative Versus Selective Abduction .....	189
8.4.2	Factual Versus Theoretical Abduction ....	190
8.4.3	Explanatory Versus Nonexplanatory Abduction .....	192
8.5	<b>Conclusion</b> .....	193
	<b>References</b> .....	193

For Peirce, not the proverbial misfortune comes in threes, but rather does *fortune*, not least with respect to his inferential triad of abduction, deduction, and induction. On the one hand, they are thought to cover the whole process of scientific reasoning from problem statement to the final adoption of a hypothesis [8.1, CP 5.171 (1903)]. On the other hand, he claimed that there are but these three elementary types of inferences so that all kinds of reasoning must belong to either abduction, deduction, or induction [8.2, CP 8.209 (c. 1905)]. Moreover, and this may sound strange, he explains in

the same place that even his earlier classification of inferences dating from 1867 can be understood in the same way, that is, in the sense of the mature Peirce's conception of the three inferences.

Peirce [8.2, CP 8.209 (c. 1905)]:

"I say that these three are the only elementary modes of reasoning there are. I am convinced of it both a priori and a posteriori. The a priori reasoning is contained in my paper in the Proceedings of the American Academy of Arts and Sciences for

April 9, 1867. I will not repeat it. But I will mention that it turns in part upon the fact that induction is, as Aristotle says, the inference of the truth of the major premiss of a syllogism of which the minor premiss is made to be true and the conclusion is found to be true, while abduction is the inference of the truth of the minor premiss of a syllogism of which the major premiss is selected as known already to be true while the conclusion is found to be true. Abduction furnishes all our ideas concerning real things, beyond what are given in perception, but is mere conjecture, without probative force. Deduction is certain but relates only to ideal objects. Induction gives us the only approach to certainty concerning the real that we can have. In forty years diligent study of arguments, I have never found one which did not consist of those elements.”

This is puzzling, if one considers Peirce’s own discussion of his earlier conception in his later work where he states explicitly that [8.2, CP 8.221 (1910)]:

“in almost everything I printed before the beginning of this century I more or less mixed up Hypothesis and Induction (i. e., abduction and induction according to his later terminology, G.M.).”

Thus, if he is not contradicting himself, both statements must be true, however, each in a specific respect.

This is one riddle I will try to solve in this chapter, but it is not the only one. I take it as one specific stumbling stone on the way to a full understanding of the very notion and logic of abduction. In order to achieve a comprehensive account of abduction, however, it is also necessary to accommodate a whole host of different concepts of abduction that have been suggested in recent years. *Magnani*, for instance, not only distinguishes between creative and selective abduction [8.3], but also between sentential and model-based abduction, theoretical and manipulative abduction, explanatory and nonexplanatory abduction [8.4]. The latter distinction is drawn from *Gabbay and Woods* [8.5], who maintain that abduction be extended to cover not merely explanatory, but also nonexplanatory abduction, although they remain diffident qualifying their differentiation “as a loose and contextually flexible distinction” [8.5, p. 115].

Another classification is proposed by *Schurz* [8.6] who distinguishes between *factual abduction*, *law-abduction*, *theoretical-model-abduction*, and *second order existential-abduction*, with the first and last being further divided into subclasses. Building on this classification and extending it, *Hoffmann* [8.7] produces a 3 × 5 matrix containing 15 types. Most importantly, he amends *Schurz*’s main categories by a form focusing

on *theoric transformations* that generate or select a new system of representation. However, the idea of theoric transformations relates to Peirce’s distinction between theorematism and corollary *deduction*, which raises the question of whether theoric transformations really belong to the realm of *abductive* reasoning (note that *Hoffmann* discusses Peirce’s analysis of Desargues’ theorem in [8.8, NEM III/2, 870–871 (1909)]. Here, another Peircean puzzle enters the scene, because he himself has claimed that theorematism deduction “is very plainly allied to retroduction (i. e., abduction, G.M.), from which it only differs as far as I now see in being indisputable” [8.9, MS 754 (1907)].

Thus, while there seem to be many different forms of abduction, it is unclear how many distinctive forms there really are. However, what is much more important is that the scientific community still seems to grapple with the very notion of abduction, that is, what are the central features of abduction as such or of its specific forms. Above, I started citing Peirce with his claim that there be only three basic and distinct kinds of inferences. However, apart from what has already been mentioned above, a persistent problem seems to be to distinguish between abduction and induction, inasmuch as *inference to the best explanation* (henceforth IBE) has to be understood as a form of induction in the Peircean sense. In [8.10], I have tried to disentangle abduction and IBE, and I have not been alone with this view [8.11]. However, *Gabbay and Woods* mention *inference-to-the-best-explanation abductions* [8.5, p. 44], and their schema for abduction [8.5, p. 47] seems to capture both abduction and IBE. *Magnani* [8.3, p. 19], [8.4, pp. 18–22] and *Schurz* [8.6, pp. 201–203] equally subsume IBE to abductive reasoning. I reckon that this has to do with similarities between their notion of *selective abduction* on the one hand and IBE on the other.

In my view, Peirce was right to claim that there are but three kinds of reasoning and that there are clear lines of demarcation between them. Accordingly, I think there is reason to *tighten the Peirce-strings* by integrating different forms of abduction (as well as deduction and induction) within a clear and coherent taxonomy. In Sect. 8.1, I will first point out that abduction and IBE are distinct (Sect. 8.1.1), then show how abduction, deduction, and induction hang together to form a productive inferential cycle from a pragmatist point of view (Sect. 8.1.2), and finally explain how this productivity enables us to construct a hierarchy of conceptual levels (Sect. 8.1.3). Within this context, different forms of abductions can be distinguished in terms of the cognitive levels at which they are located, and in terms of whether *new concepts are invented* or *existing ones applied*.

In Sect. 8.2, I explicate the logic of each of the three inferential types. This is done in two steps. First



(Sect. 8.2.1), the inferences will be analyzed in terms of three characteristic subprocesses that Peirce assumes for inferences in general, that is, (1) *colligation*, (2) *observation*, and (3) *judgment* ([8.12, CP 2.442–244 (c. 1893)], see also *Kapitan* [8.13, p. 479]). Next, the validity of each inference will be discussed in Sect. 8.2.2.

Based on this analysis, Peirce’s notion of *theore-matic reasoning* is explored in Sect. 8.3. In Sect. 8.3.1, theore-matic deduction is explicated as *inverse deduc-tion*, leading from the result of corollarial deduction to the premise of corollarial deduction, that is, the theoret-ical point of view from which the result can be deduced. An instructive example is given in Sect. 8.3.2, and in Sect. 8.3.3 the idea of inverse inferences is extended to inverse abduction and induction. As a result, we end up with three ordinary and three inverse forms of *pure* inferential types (note that Peirce has also introduced *analogy* as a compound inference conjoining abduction and induction [8.14, CP 1.65]; see also [8.15] on this issue).

## 8.1 Abduction in the Overall Inferential Context

### 8.1.1 Disentangling Abduction and IBE

In a recent overview of Peirce’s theory of abduction, Psillos stresses that abduction, deduction, and induction “constitute the three ultimate, basic and independent modes of reasoning. This is a view that runs through the corpus of the Peircean work” [8.16, p. 121]. So, the task of defining and distinguishing these three kinds of inferences might be assumed to be easy. However, reality is different, not least because [8.16, pp. 136–137]

“[t]he picture of abduction that Peirce has painted is quite complex. On the face of it, there may be a question of its coherence. Abduction is an inference by means of which explanatory hypotheses are admitted, but it is not clear what this admission amounts to.”

Why is abduction so hard to grasp? To my mind, the main reason is that it is often confounded with (aspects of) induction or IBE and that Peirce himself has given rise to such confusion. At first, he came up with a syl-logistic account of this inferential triad [8.12, CP 2.623 (1878)], but told us later that “in almost everything I printed before the beginning of this century I more or less mixed up Hypothesis and Induction” [8.2, CP 8.221 (1910)]. As already quoted above, however, he also maintained that in some sense the early Peirce’s concept of *hypothesis* and the mature Peirce’s *abduction* were still equivalent. What’s more, the passage quoted

In Sect. 8.4, I will discuss three important distinc-tions among forms of abductive reasoning: *creative* versus *selective* abduction (Sect. 8.4.1), *factual* versus *theoretical* abduction (Sect. 8.4.2), and *explanatory* ver-sus *nonexplanatory* abduction (Sect. 8.4.3). It turns out that abductions (and other inferences) are to be dis-tinguished in terms of knowledge generation (creative) versus knowledge application (selective) and along two cognitive dimensions: one concerns *levels* of abstrac-tion (from elementary embodied and perceptual levels to high-level scientific theorizing). The other concerns *domains* of reasoning such as explanatory, instrumental, and moral reasoning. In the concluding Sect. 8.5, the main results of my analysis are summarized and routes for further research indicated.

Although I consider my argumentation coherent and in line with Peirce, I do not claim to deliver an exegesis of what Peirce himself might have thought, especially since parts of my inferential taxonomy are clearly not contained in his works.

in the introduction makes explicit that the purpose of abduction is twofold (1) to generate new hypotheses and (2) to select hypotheses for further examination (see also [8.13, p. 477]; [8.17, p. 503]). The same is true for the following passage [8.18, CP 6.525 (1901)]:

“The first starting of a hypothesis and the enter-taining of it, whether as a simple interrogation or with any degree of confidence, is an inferen-tial step which I propose to call abduction. This will include a preference for any one hypothesis over others which would equally explain the facts, so long as this preference is not based upon any previous knowledge bearing upon the truth of the hypotheses, nor on any testing of any of the hypotheses, after having admitted them on probation. I call all such inference by the peculiar name, abduction, because its legitimacy depends upon altogether different principles from those of other kinds of inference.”

Thus, the question remains as to whether abduction is associated with IBE at least in some sense, and if so, whether abduction as such bears this feature or whether there are two different basic kinds of abduction – creative and selective – as *Magnani* [8.3, 4] and *Schurz* [8.6] hold. In the succeeding passages, Peirce writes on the testing of hypotheses and explains his concept of induction, in particular as a means to determine which of a number, or even a multitude, of hypotheses

is the best explanation and ought to be adopted as true or likely to be true [8.18, CP 6.526–536 (1901)]. Within this elaboration, he is careful to make sure that the selective aspect of abduction is something different [8.18, CP 6.528 (1901)]:

“These distinctions (among forms of induction, G.M.) are perfectly clear in principle, which is all that is necessary, although it might sometimes be a nice question to say to which class a given inference belongs. It is to be remarked that, in pure abduction, it can never be justifiable to accept the hypothesis otherwise than as an interrogation. But as long as that condition is observed, no positive falsity is to be feared; and therefore the whole question of what one out of a number of possible hypotheses ought to be entertained becomes purely a question of economy.”

Elsewhere, Peirce points out that the abductive selection of hypotheses is usually guided by criteria like simplicity and breadth [8.19, CP 7.220–222 (c. 1901)], and his main idea is that we do not consider just any possible hypothesis, but those hypotheses that make most sense to us from the outset (i. e., those with the highest subjective prior probabilities; see also [8.20]). However, this is no IBE, since it is merely a question of economy, and Peirce clearly states that no sound judgment could be based on this kind of selection, for it is “the most deceptive thing in the world” [8.12, CP 2.101 (1902)]. The further course of inquiry, and whether an originally neglected possibility will have to be taken up at a later stage, depends entirely on induction. For when the selected hypothesis is finally evaluated in the light of empirical data it has to be judged [8.12, CP 2.759 (1905)]:

“whether the hypothesis should be regarded as proved, or as well on the way toward being proved, or as unworthy of further attention, or whether it ought to receive a definite modification in the light of the new experiments and be inductively reexamined ab ovo, or whether finally, that while not true it probably presents some analogy to the truth, and that the results of the induction may help to suggest a better hypothesis.”

If the hypothesis is *regarded as proved*, then this is IBE. However, it could not possibly be regarded as proved if there were yet another hypothesis around that could not be excluded based on the evidence gathered so far. Hence, selection in the context of abduction and selection in the context of induction are quite different. In the former case, its role is merely practical, not logical; that is, one hypothesis is tried first, and if it testifies to be beyond any doubt, other alternatives would not

have to be considered anymore (however, this implies that by the same token all possible alternatives must, in fact, be refuted). Or if it is to be conceived as *logical*, then the hypothesis to be rejected at this stage has either to be conceived as abductively (here: explanatorily) invalid (see Sect. 8.2.2), or the rejection has to follow from an *inductive* evaluation of the competing hypotheses. From such an inductive evaluation it might follow that the hypothesis currently countenanced is *well on the way of being proved* in the above-quoted sense, in that it is better than a number of other hypotheses, although further testing or further reflection about novel approaches seems appropriate.

Anyhow, it has to be admitted that Peirce is imprecise in this respect. However, in order not to confuse abductive and inductive logic, I would suggest the rigid interpretation just stated. Moreover, I would like to refer to *Aliseda*, who has pointed out very clearly the difficulties of coming to grips with the selection of a *best* or *preferred* inference as an abductive task [8.21, pp. 72–74], even though she herself endorses hypothesis selection as an abductive task [8.21, p. 33].

On this background, let us now consider *Gabbay* and *Woods*’ reconstruction of abductive reasoning [8.5, p. 47]. According to them, it starts with a cognitive target  $T$  (e.g., to explain a certain phenomenon) that cannot be met based on the reasoner’s background knowledge  $K$ , and that the reasoner wants to attain (hence  $T$ ).  $R$  denotes the attainment relation on  $T$ , and  $R^{\text{pres}}$  the presumptive attainment relation on  $T$ . If  $R(K, T)$  is not possible, the reasoner aims at an enhanced successor knowledge base  $K^*$  so that  $R(K^*, T)$  holds.  $H$  denotes a hypothesis, and  $K(H)$  a knowledge base revised by  $H$ . Furthermore, there is  $C(H)$ , which means that it is “justified (or reasonable) to conjecture that  $H$ ” [8.5]. And finally, “ $H^c$  denotes the discharge of  $H$ .  $H$  is discharged when it is forwarded assertively and labelled in ways that reflect its conjectural origins” [8.5]. Based on these definitions they suggest the following schema [8.5, p. 47]:

- |   |                       |
|---|-----------------------|
| 1. $T!$   | [declaration of $T$ ] |
| 2. $\neg(R(K, T))$                                | [fact]                |
| 3. $\neg(R(K^*, T))$                              | [fact]                |
| 4. $R^{\text{pres}}(K(H), T)$                     | [fact]                |
| 5. $H$ meets further conditions $S_1, \dots, S_n$ | [fact]                |
| 6. Therefore, $C(H)$                              | [conclusion]          |
| 7. Therefore, $H^c$                               | [conclusion]          |

As I try to explain also in Sect. 8.3, the reach of abduction ought to be limited to steps 1 through 4, with (4) establishing a valid abductive inference, that is, that  $R^{\text{pres}}(K(H), T)$  holds. This establishes abductive validity in that  $H$  is capable of explaining the surprising facts. And this is precisely what  $R^{\text{pres}}(K(H), T)$  cap-

tures. Steps 5 through 7 ought, to my mind, be attributed to induction, in this case a rather tentative induction that Peirce has labeled “*abductive induction*” [8.18, CP 6.526 (c. 1901)], because it only qualifies a hypothesis  $H$  as better than possible alternative hypotheses, but not in the sense of a full-fledged inductive judgment to the truth of  $H$ . Further conditions  $S_1, \dots, S_n$  may exist in the form of background knowledge pertaining to Peirce’s criteria of simplicity or breadth (see above) or additional empirical evidence in favor of  $H$ . However, these further pieces of information clearly go beyond the abductive task; they may be produced by *deductive* reasoning about what certain hypotheses imply (because how do  $S_1, \dots, S_n$  become conscious?), and are finally considered in *inductive* reasoning. Therefore, I propose to repatriate steps 5 through 7 to the realm of *induction*, and to take very seriously the following statement [8.2, CP 8.218 (c. 1901)]:

“Nothing has so much contributed to present chaotic or erroneous ideas of the logic of science as failure to distinguish the essentially different characters of different elements of scientific reasoning; and one of the worst of these confusions, as well as one of the commonest, consists in regarding abduction and induction taken together (often mixed also with deduction) as a simple argument. Abduction and induction have, to be sure, this common feature, that both lead to the acceptance of a hypothesis because observed facts are such as would necessarily or probably result as consequences of that hypothesis. But for all that, they are the opposite poles of reason [...].”

Recently, *McKaughan* [8.20], *Campos* [8.22], and *Mackonis* [8.23] have argued in favor of a wide notion of IBE, including abduction, although they endorse the sharp distinction others and myself have made. However, in the light of the subtle, but nonetheless important, distinctions I have tried to highlight in this section, I think there is not much use fitting it all in one global concept of IBE.

### 8.1.2 The Dynamical Interaction of Abduction, Deduction, and Induction

By the end of the nineteenth century, Peirce rejected his original syllogistic approach and said “I was too much taken up in considering syllogistic forms [...], which I made more fundamental than they really are” [8.12, CP 2.102 (1902)]. However, even more to the point, Peirce realized that induction “never can originate any idea whatever. Nor can deduction. All the ideas of science come to it by the way of Abduction” [8.1, CP 5.145

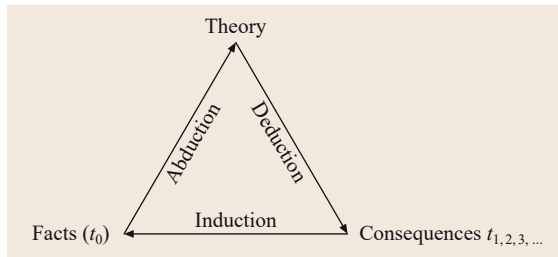
(1903)]. The crucial point here is that induction can only operate with concepts that are already at hand. On top of this, even simple regularities like  $\forall x(Fx \rightarrow Gx)$  do not suggest themselves, but have to be considered by an active mind, before they can be tested and eventually accepted or rejected (see Sect. 8.1.3, relating to Carnap’s disposition predicates). This is why the mature Peirce suggests that abduction is the process by which new concepts, laws, and theories are first conceived, before they are investigated further by deductive and inductive processes [8.1, CP 5.171 (1903)]:

“Abduction is the process of forming an explanatory hypothesis. It is the only logical operation which introduces any new idea; for induction does nothing but determine a value, and deduction merely evolves the necessary consequences of a pure hypothesis. Deduction proves that something *must* be; Induction shows that something *actually* is operative; Abduction merely suggests that something *may be*. Its only justification is that from its suggestion deduction can draw a prediction which can be tested by induction, and that, if we are ever to learn anything or to understand phenomena at all, it must be by abduction that this is to be brought about.”

Abduction is most important in our overall reasoning, because without it we could not possibly acquire any idea of the world, not even elementary perceptions of objects, let alone scientific theories. Hence, “no new truth can come from induction or from deduction” [8.2, CP 8.219 (c. 1901)]. Whereas abduction is very powerful in terms of the generation of fruitful new ideas, however, it is very weak in terms of empirical validity, as Peirce often stresses. He even says that his discovering the true nature of abduction so late was “owing to the extreme weakness of this kind of inference” [8.12, CP 2.102 (1902)]. Empirical validity is gained by deducing necessary consequences from the abduced hypotheses, especially predictions that can be tested empirically, and the inductive evaluation of the experimental results (or other suitable evidence). Figure 8.1 illustrates the dynamical interaction of the three inferential types.

So far, the role of abduction and deduction seems self-evident. However, an explanation should be given for the role of induction in this triad, in particular why it points back to where abduction starts in Fig. 8.1. After all, induction is typically understood as the inference to the truth (or falsity) of the theory in question, and as a consequence it should point back to the theory itself, like for example in *Magnani’s* ST-Model [8.4, p. 16]; also [8.3, p. 23].

However, induction in the Peircean sense is tied back to his pragmatism, which, again, rests on the logic



**Fig. 8.1** The dynamical interaction of abduction, deduction, and induction

of abduction (cf. also [8.15, pp. 207–212]). Peirce [8.1, CP 5.196 (1903)]:

“That is, pragmatism proposes a certain maxim which, if sound, must render needless any further rule as to the admissibility of hypotheses to rank as hypotheses, that is to say, as explanations of phenomena held as hopeful suggestions.”

In other words, all we need in the first place is the confidence that abduction may enable us, by the construction of concepts and hypotheses, to acquire objective knowledge at all. Otherwise, all deductive and inductive examination were futile. However, since abduction is based on experience, induction cannot go beyond that level to infer some kind of “absolute truth” [8.24–27]. Therefore, all that induction can do is to establish habits of expectation and action based on theories.

Peirce [8.1, CP 5.197 (1903)]:

“What, then, is the end of an explanatory hypothesis? Its end is, through subjection to the test of experiment, to lead to the avoidance of all surprise and to the establishment of a habit of positive expectation that shall not be disappointed.”

To be sure, theories in this sense are understood realistically, not as psychological contrivances [8.24, pp. 201–203]. However, finally adopting a theory means to project its content onto all its cases, observed and unobserved, in the past, present, and future. And only in this sense can a theory still be revised or eventually rejected in the future. It is important, therefore, that a circle closes in this very sense of creating habits of expectation and action so that these habits can, in principle, always be broken up again in the future (see also [8.26, pp. 51–57]).

Peirce [8.2, CP 8.270 (1902)] (see also [8.1, CP 5.524 (c. 1905)]):

“The question is what passes in consciousness, especially what emotional and irritational states of feeling, in the course of forming a new belief. The

man has some belief at the outset. This belief is, as to its principal constituent, a habit of expectation. Some experience which this habit leads him to expect turns out differently; and the emotion of *surprise* suddenly appears.”

Thus, when an accepted theory is subsequently applied to relevant cases, it is *not only* being applied, but also reassessed over and over. In this very sense, knowledge acquisition and knowledge application are fundamentally tied together and follow the same inferential principles. That is, every application of previously acquired knowledge has to be understood as:

1. Abducing from a certain situational configuration to a suitable interpretation, then
2. Deducing a certain course of action or expectation, and
3. Inducing whether one’s actions or expectations were confirmed by experience.

Furthermore, if every application of knowledge constitutes a chance to strengthen or weaken the underlying belief, then by the same token, all failures and situation-specificities in the application of knowledge [8.28] and action-guiding principles, in particular moral principles [8.29, 30], can also be addressed and analyzed within this very frame of reference.

### 8.1.3 Abduction and Abstraction

As we have seen in the previous sections, the key feature of abduction is to construct new explanatory concepts to accommodate the surprising facts that give rise to them. “By its very definition abduction leads to a hypothesis which is entirely foreign to the data” [8.31, MS 692 (1901)]. Abductive hypotheses are typically of a higher degree of complexity in the sense that phenomena are not just explained by other phenomena that might have brought them about, but by a *theory* of the phenomena. Such theories constitute higher cognitive levels, and this entails that theoretical entities are not observable in the same way as the phenomena they are meant to explain. Of course, *Hanson’s* principle of theory-ladenness of experience [8.32] states that no experience (at least no description of experiences) is theory-free. Nonetheless, there are levels of cognitive architectures building on each other. This has been clear ever since *Carnap* discovered that not even simple disposition predicates could be reduced to observation sentences [8.33, 34].

When *Schurz* [8.6] differentiates between variants like *fact-abduction*, *law-abduction*, or *theoretical abduction*, he also distinguishes such levels, however, without making this aspect of successive theory-

building fully explicit. I think it would be a fruitful endeavor to reconstruct how conceptual (or theoretical) levels are built onto one another by successive abductions. For instance, when a simple disposition is discovered (that sugar dissolves in water), this constitutes an empirical law, which is itself a concept of a regularity in nature. It can also be used to explain why someone else does not see the sugar in her drink. We could say that it is very well in there, but cannot be seen, because it has dissolved.

What this simple example shows is that even in simple fact-abduction, we do not just infer to the fact, but to the law from which it then follows that sugar might be in a liquid, even if no sugar can be seen. Thus, dispositional laws and simple action schemes (*When the switch is pushed, the light will go on*) constitute an elementary theory level, that is, regularities in terms of observation language. However, these regularities are themselves phenomena that one may wish to explain, especially when one starts wondering, *how* the switch is causally connected with the light. At first, it was established *that* a natural regularity exists. Now, as the regularity is established as a matter of fact, it becomes the object of theoretical reflection and represents the fact to be explained.

This is what Hintikka highlights when he discusses the difference between abduction and IBE. He says that [8.32, p. 509]:

“when a dependence law telling us how the observed variable depends on the controlled one the law does not *explain* the result of the experiment. It is the *result* of the experiment, nature’s answer to the experimental investigator’s question.”

*Earnan McMullin* has made the same point concerning the role of laws in explanation: “Laws are the explananda; they are the questions, not the answers” [8.35, p. 90]. And he continues [8.35, p. 91]:

“To explain a law, one does not simply have recourse to a higher law from which the original law can be deduced. One calls instead upon a *theory*, using this term in a specific and restricted sense. Taking the observed regularity as effect, one seeks by abduction a causal hypothesis which will explain the regularity. To explain why a particular sort of thing acts in a particular way, one postulates an underlying structure of entities, processes, relationships, which would account for such a regularity. What is ampliative about this, what enables one to speak of this as a strong form of understanding, is that if successful, it opens up a domain that was previously unknown, or less known.”

I consider this a strong and important point (see also [8.36, 37] on explanatory hierarchies and explanatory coherence). Laws, in this view, are not the solutions (the explanations) but the problems (the facts to be explained). However, I would not go so far as to deny laws, even simple empirical laws, any explanatory function. It just depends on the point of view and the theoretical level, which is needed and appropriate to solve a particular explanatory problem. If one is looking for causal relationships between events, one is in fact searching for law-like explanations. And this not only applies to children in their early cognitive development. Most adults are content with knowing what keys to press in order to use certain functions of a software; in such cases the question is how a certain result is brought about, and the explanation consists in functional relations between the keys or menu options and the results visible on the screen. The same applies to cookbooks and manuals for technical appliances in which it is explained, how things work or why something I tried did not work. I assume that, for example, *Schurz’s* account of explanation as unification applies not only to scientific theories, but also to such simple forms of explanation [8.10, 38].

Thus, there seems to be an order of theory-levels or levels of abstraction, where the higher ones explain the lower ones, and where abduction is the process that takes the reasoner from a lower level to a higher one. Such a hierarchy of levels may also be the clue to understanding how (intuitive) cognition works below explicit sentential reasoning and how the latter comes about in ontogenetic development.

Peirce famously argued that perceptual judgments are no abductions. However, he seems to have been too strict or narrow-minded in this context (see also [8.3, pp. 42–43], [8.4, pp. 268–276]). While he clearly admits that perceptual judgment “is plainly nothing but the extremest case of Abductive Judgements” [8.1, CP 5.185 (1903)] and that “abductive inference shades into perceptual judgment without any sharp line of demarcation between them” [8.1, CP 5.181 (1903)], he maintains that they are nonetheless distinct, because unlike abductive inferences, perceptual judgments were “absolutely beyond criticism” [8.1, CP 5.181 (1903)]. Peirce points out repeatedly that abduction as an inference requires control and that this misses in perceptual judgment [8.1, CP 5.157, 181, 183, 194 (1903)]. He therefore holds that perceptual judgment is the “starting point or first premiss of all critical and controlled thinking” [8.1, CP 5.181 (1903)], hence something on which abduction is based, but which does not belong to abduction itself.

However, Peirce fails to consider two aspects of perceptual judgments: first, they might be conceivable as

established facts in the sense of a complete triad of abduction, deduction, and induction. This would explain why we are (normally) certain of our perceptions. Second, and more importantly, Peirce fails to consider that abductions are well-controlled, not by conscious thought, but by action. Perceptions can be understood as habits of action, that is, of categorization and behavior in accordance with what we perceive. And finally, the abductive or conjectural part of this process is that with every new perception the individual literally *makes sense* of what enters into the sensory system. Sometimes these *creations* are fallacious or even foolish, but this puts them fully in line with abduction in general.

At least, this is what I suggest at this point, and it would certainly have to be examined in more detail. However, *Magnani's* [8.4] and *Park's* [8.39] reflections on animal, visual, and manipulative abduction point in the very same direction and could be accommodated as basic forms, not only of abduction, but also of inferential functioning in general. Moreover, reconstructing *the formation* of perceptual judgments in this way makes the pragmatist epistemology even more stronger, because there is no specific level of consciousness, where epistemic processes start (see [8.40] for a discussion of this link between abduction and pragmatism). As for perceptual judgments, they are at this basic level not controlled by conscious reflection, but nonetheless controlled in an embodied and enacted manner [8.41]. Epistemology, then, would rather be embedded in life and would not have to resort to any a priori forms of cognition whatsoever. It also constitutes, to my mind, a sound basis for nonreliabilist epistemological externalism.

Without being able to explain my main point here in detail, I just assume that there are levels over levels of understanding and interaction with one's environment. And even though Peirce has never developed a theory of successive abstraction in this overarching sense, he had a clear idea of the basic principle, which he calls *hypostatic abstraction*. To explain his ideas, Peirce relates to Molière's *Malade imaginaire*, where a medical student takes his oral examination in the last scene. He is asked why opium puts people to sleep and he confidently replies that opium had a dormitive virtue whose nature was to lull the senses to sleep [8.1, CP 5.534 (c. 1905)]:

“Quia est in eo  
Virtus dormitiva,  
Cujus est natura  
Sensus assoupire.

Whereupon the chorus bursts out,  
Bene, bene, bene, bene, responde:  
Dignus, dignus est entrare,  
In nostro docto corpore.”

Peirce explains [8.1, CP 5.534 (c. 1905)]:

“Even in this burlesque instance, this operation of hypostatic abstraction is not quite utterly futile. For it does say that there is some peculiarity in the opium to which the sleep must be due; and this is not suggested in merely saying that opium puts people to sleep.”

Elsewhere, he discusses the same idea, but speaks of *subjectal abstraction* as opposed to *precisive abstraction* (see also [8.42]).

Peirce [8.8, NEM III/2, p. 917 (1904)]:

“There are two entirely different things that are often confused from no cause that I can see except that the words *abstract* and *abstraction* are applied to both. One is *αφαίρεσις* leaving something out of account in order to attend to something else. That is *precisive* abstraction. The other consists in making a subject out of a predicate. Instead of saying, Opium puts people to sleep, you say it has dormitive virtue. This is an important proceeding in mathematics. For example, take all *symbolic* methods, in which operations are operated upon. This may be called *subjectal abstraction*.”

By *subjectal abstraction* Peirce means that “a transitive element of thought is made substantive, as in the grammatical change of an adjective into an abstract noun” [8.12, CP 2.364 (1901)]. Even though *dormitive virtue* does not explain why opium puts people to sleep, it states an explanatory problem in the sense that a general law (with opium as cause and putting people to sleep as effect) has to be explained at a higher, more abstract level. In this very sense, *dormitive virtue* goes beyond stating a mere disposition of opium (see also *Schurz's* discussion on this issue [8.6, pp. 219–221]).

Forms of abductive reasoning could, therefore, be distinguished according to levels of abstraction in the sense in which Jean Piaget discusses constructive development and cognitive architectures [8.43]. Among the forms that *Schurz* [8.6] differentiates, some are at the same level, while others belong to different levels. *Factual abduction* and *law abduction* all concern simple empirical laws, the latter establishing them, the former applying them. Higher level abduction is what he calls *theoretical-model abduction*: “The explanandum of a theoretical-model abduction is typically a well-confirmed and reproducible empirical phenomenon expressed by an *empirical law*” [8.6, p. 213].

However, the distinction between empirical laws, on the one hand, and theoretical models, on the other hand, seems to be still rather crude (again, if this is compared to the fine-grained, but highly systematic, distinctions

made by *Piaget* and *Garcia* [8.43]). To date, research on cognitive architectures has primarily focused on the lower end of the cognitive hierarchy, i. e., how relative simple conceptual and action schemata are built and grounded in the brain's modal systems for perception, emotions, and actions [8.41, 44, 45]. However, since ab-

duction is the process that leads to successively more abstract cognitions in the sense of hierarchical complexity, there is a promising route for further research and a systematic differentiation of types of abductions according to the cognitive levels, to which they apply (as for moral cognition see [8.46] as an example).

## 8.2 The Logicity of Abduction, Deduction, and Induction

### 8.2.1 Inferential Subprocesses and Abduction as Inferential Reasoning

As already mentioned above, Peirce regarded abduction as an extremely weak kind of inference. This raises the question of whether it is an inference at all. On top of this, he says that abduction is “nothing but guessing” [8.19, CP 7.219 (1901)] and its results merely “the spontaneous conjectures of instinctive reason” [8.18, CP 6.475 (1908)]. However, abduction is also said to “cover all the operations by which theories and conceptions are engendered” [8.1, CP 5.590 (1903)], and since it takes us to novel concepts and theories, he cannot mean *guesses* in the ordinary sense of picking out something at random from a range of existing objects of choice. However, the question remains whether abduction is an inference or merely an instinct. In a way, both seems to be true [8.47], but for the present purpose it suffices to stress that abduction has an inferential aspect [8.32, 47, 48]. So, let us try to track this inferential aspect of abduction.

In this respect it may be instructive to consider Peirce's thoughts on inference in general. On his view, all inferences are mental acts of reasoning and as such describe a process with a definite beginning and a definite end. Any inference begins with a question that requires an answer in the form of the respective conclusion. Abduction aims at possible explanations, deduction at necessary consequences following from certain premises, and induction aims at determining whether to accept or reject a hypothesis. Whatever the inference, however, the process of answering these questions contains three distinctive steps, which Peirce calls *colligation*, *observation*, and *judgment*.

Peirce [8.12, CP 2.442 (c. 1893)]:

“The first step of inference usually consists in bringing together certain propositions which we believe to be true, but which, supposing the inference to be a new one, we have hitherto not considered together, or not as united in the same way. This step is called *colligation*.”

Peirce [8.12, CP 2.443–444 (c. 1893)]:

“The next step of inference to be considered consists in the contemplation of that complex icon . . . so as to produce a new icon. [. . .] It thus appears that all knowledge comes to us by observation. A part is forced upon us from without and seems to result from Nature's mind; a part comes from the depths of the mind as seen from within [. . .].”

Peirce [8.12, CP 2.444]:

“A few mental experiments – or even a single one [. . .] – satisfy the mind that the one icon would at all times involve the other, that is, suggest it in a special way [. . .] Hence the mind is not only led from believing the premiss to judge the conclusion true, but it further attaches to this judgment another – that *every* proposition *like* the premiss, that is having an icon like it, *would* involve, and compel acceptance of, a proposition related to it as the conclusion then drawn is related to that premiss.”

He concludes that “[t]he three steps of inference are, then, colligation, observation, and the judgment that what we observe in the colligated data follows a rule” [8.12, CP 2.444]. The step of colligation is consistently used and explained and thus seems to be rather clear [8.1, CP 5.163 (1903)], [8.1, CP 5.579 (1898)]. However, Peirce is less precise about the other two. In particular, his differentiation, in this context, between a *plan* and the *steps* of reasoning may cause some confusion [8.1, CP 5.158–166 (1903)]. As for the *plan* he says that [8.1, CP 5.162 (1903)]:

“we construct an icon of our hypothetical state of things and proceed to observe it. This observation leads us to suspect that something is true, which we may or may not be able to formulate with precision, and we proceed to inquire whether it is true or not.”

Thus, we observe what is colligated in the premise in order to produce a result. Even though this observation may be guided by strategies and other background knowledge the result will first come about in a spontaneous act as the reasoner becomes conscious of it.

When discussing observation in the context of abduction, he goes on to a general description of observation that brings out this main feature very plainly [8.1, CP 5.581 (1898)]:

“And then comes an Observation. Not, however, an External observation of the objects as in Induction, nor yet an observation made upon the parts of a diagram, as in Deduction; but for all that just as truly an observation. For what is observation? What is experience? It is the enforced element in the history of our lives. It is that which we are constrained to be conscious of by an occult force residing in an object which we contemplate. The act of observation is the deliberate yielding of ourselves to that *force majeure* – an early surrender at discretion, due to our foreseeing that we must, whatever we do, be borne down by that power, at last.”

Thus, the observed result is forced upon us in a rather uncontrolled manner. We just see it and can't help seeing it. However, in order to come to a conclusion as the last step of inference, we have to evaluate whether the result is valid in terms of the respective inference. This constitutes the judgmental step that finalizes each inference (see [8.49] for this matter).

### 8.2.2 The Validity of Abduction, Deduction, and Induction

Peirce is explicit concerning the validity of an abductive judgment [8.1, CP 5.197 (1903)]:

“What is good abduction? What should an explanatory hypothesis be to be worthy to rank as a hypothesis? Of course, it must explain the facts. But what other conditions ought it to fulfill to be good? The question of the goodness of anything is whether that thing fulfills its end. What, then, is the end of an explanatory hypothesis? Its end is, through subjection to the test of experiment, to lead to the avoidance of all surprise and to the establishment of a habit of positive expectation that shall not be disappointed. Any hypothesis, therefore, may be admissible, in the absence of any special reasons to the contrary, provided it be capable of experimental verification, and only insofar as it is capable of such verification. This is approximately the doctrine of pragmatism.”

Valid *abduction* thus has to satisfy two criteria:

1. It has to *explain the facts*, meaning that the initial surprise be eliminated, and
2. The explanation has to be *capable of experimental verification* in the pragmatist sense.

Any abductively observed result that does not meet these criteria will have to be rejected. If the criteria are met, however, the hypothesis will have to be accepted as a valid abductive conclusion (see also my reflections in Sect. 8.1.1). From this point of view, we can now understand Peirce's famous statement of the abductive inference [8.1, CP 5.189 (1903)]:

“The surprising fact, C, is observed;  
But if A were true, C would be a matter of course,  
Hence, there is reason to suspect that A is true.”

This only relates the *final* subprocess of abduction, the judgmental part. However, it is not to be confused with abduction as an inferential cognitive process as a whole [8.50]. *Kapitan* has famously criticized Peirce's concept of abduction, claiming that it was essentially a deductive argument [8.51]. However, he fails to see that the above statement does not describe the entire process of abductive reasoning. And, as far as this *deductive* aspect of the abductive judgment is concerned, Peirce himself expresses this clearly [8.1, CP 5.146 (1903)], [8.40, p. 168]. However, this does not turn the abductive judgment into a deductive inference, because it is not the primary task to derive C, since C is already known and constitutes the premise, whereas A constitutes the conclusion.

*Kapitan's* later account of abduction [8.51] is largely adequate, but still there is one widespread problem. Like many others, he conflates the abductive judgment with the selective aspect that I have argued above (Sect. 8.1.1) should be excluded. *Hintikka* [8.52, pp. 44–52] grapples with the notion of abductive inference for just the same reason. In my view, this makes it all the more important to drive a deep wedge between accommodating the facts as *the* abductive task, and evaluating abductively valid hypotheses as an inductive task.

The validity of *deduction* seems to be unproblematic. So, let us move straight to *induction*. It has already been pointed out above that induction is the inference that yields factual knowledge, constituting factual truth. However, what is the precise relation between knowledge and truth? The classical notion of knowledge as *justified true belief* requires that a proposition be true in order to be known. However, a main theorem from the point of view of pragmatism is that knowledge is logically prior, that is, knowledge establishes truth rather than requiring it as a condition. After all, this is the basic idea of disquotationalism [8.26, pp. 57–64].

As regards the validity of induction, I adopt an analysis of knowledge and its formation proposed by *Suppe* [8.53] within the framework of a possible-worlds semantics. He suggests a nonreliabilistic externalist approach to knowledge. On this view, we know *p* if it is not causally possible that we perceive the evidence



as it is unless the suggested hypothesis is true. This is indicated by a causal possibility operator  $\diamond$ , where *causal possibility* refers to all logically possible worlds that are consistent with the natural laws of our world (i. e., our current background knowledge regarding natural laws). According to Suppe’s approach, the truth of a proposition results from knowing it, and knowing results from the condition stated in (iv), below, being satisfied. Furthermore, “satisfying (iv) entails the satisfaction of condition (iii)” [8.53, p. 402], since  $R$  and/or  $K$  function as decisive indicators for  $\Phi$ .

*S propositionally knows that  $\theta$  if and only if*

- (i) *S* undergoes a cognitive process  $R$ , or *S* has prior knowledge that  $K$
- (ii) *S*, knowing how to use  $\Phi$  and knowing how to use  $\theta$  with the same propositional intent, as a result of undergoing  $R$  or having prior knowledge that  $K$  entertains the proposition  $\Phi$  with that propositional intent as being factually true or false
- (iii)  $\Phi$  is factually true
- (iv) there exists a conjunction  $C$  of partial world state descriptions and probability spaces such that  $C \ \& \ \sim \diamond \Phi \ (C \ \& \ R \ \& \ K \ \& \ \sim \Phi) \ \& \ \diamond \Phi \ (C \ \& \ \sim \Phi) \ \& \ R \ \& \ \diamond \ (R \ \& \ \sim \Phi)$
- (v) as a result of undergoing  $R$  or  $K$ , *S* believes that  $\Phi$  [8.53, p. 405].

As a result, induction can be conceived in terms of an elaborate eliminative inductivism in the sense of Earman [8.54]. A theory is to be adopted, if all that has been observed so far supports it and that no alternative hypothesis is conceivable (at the current state of knowledge).

The results of my analysis are condensed in Fig. 8.2 (which is reduced to the essential features). Note that the diagram shows steps in the inferential processes. They are not to be misread as syllogisms.

### 8.3 Inverse Inferences

#### 8.3.1 Theorematic Deduction as Inverse Deduction

Writing about mathematical reasoning, Peirce says: “My first real discovery about mathematical procedure was that there are two kinds of necessary reasoning, which I call the Corollarial and the Theorematic” [8.55, NEM IV, p. 49 (1902)]. However, scholars disagree on what is the content of this discovery. For Hintikka, “a valid deductive step is theorematic, if it increases the number of layers of quantifiers in the proposition” [8.56, p. 307]. Ketter claims that Hintikka fails to see “the true importance of Peirce’s corollar-

	Abduction	Deduction	Induction
Colligation	$C$	$H \wedge P$	$\Box((H \wedge P) \rightarrow E) \wedge E$
Observation	$H \rightarrow \diamond C$	$(H \wedge P) \rightarrow E$	$E \wedge \neg \diamond(E \wedge \neg H)$
Judgement	$\diamond H$	$\Box((H \wedge P) \rightarrow E)$	$\Box H$

Fig. 8.2 A formal model of inferential (sub)processes

In abduction we colligate the relevant aspects of a given explanatory problem, that is, what happens to be the case, hence  $C$ .  $C$  is observed with the intention to find a theoretical hypothesis  $H$  that would, if true, explain  $C$ , that is, render the previously surprising phenomenon causally possible. As one hits on an idea, one has to make sure that  $H$  would really explain  $C$ . This is the abductive judgement  $\diamond H$ .

This result gained from abduction is then used as input for the following deduction, together with suitable premises  $P$  available from background knowledge. These are observed so as to generate necessary consequences, in particular empirical hypotheses  $E$ . The judgment  $\Box((H \wedge P) \rightarrow E)$  states that  $E$  follows with necessity from  $(H \wedge P)$ .

Again, the deductive conclusion is input into induction, where it is colligated with the actual experiment, which are then observed. This observation is more than just recording what happens; in fact, such recording would have to be understood as the main part of the inductive colligation. Observation in the context of induction means to look at these results (maybe at the time when they are actually produced) under the aspect of whether they confirm the tested hypothesis and disconfirm its rivals. If the final outcome is positive,  $H$  is accepted as causally necessary, hence  $\Box H$ .

ial/theorematic reasoning distinction” [8.57, p. 409], which, according to Ketter, is that “it makes significant contributions toward showing that mathematics and logic are observational, experimental, hypothesis-confirming sciences” [8.57, p. 409]. Ketter also maintains that the “production of experiments within theorematic reasoning, on Peirce’s view, is done through abduction” [8.57, p. 411]. Referring to this argument, Hoffmann says he had “spent some effort to find in Peirce’s writings hints at such a connection between abduction and theorematic reasoning, but without much success” [8.48, p. 293]. However, Hoffmann acknowledges and discusses obvious similarities between the-

orematic deduction and abduction and comes to the following result: “It is one thing *to prove* a theorem and another to *formulate* it” and continues that “it would make sense to describe the first task as theorematic deduction and the second task as abduction” [8.48, p. 294].

In view of this puzzlement concerning the proper understanding of theorematic deduction and its relation to abduction, my suggestion is not to subdivide theorematic deduction into abductive and deductive aspects, but to reconstruct theorematic deduction as a form of reasoning of its own, albeit similar to abduction in an important respect. As for the similarity between abduction and theorematic deduction, Peirce himself remarks that “[i]t is very plainly allied to retrodution, from which it only differs as far as I now see in being indisputable” [MS 754 (1907), quoted from [8.48, p. 293]]. And the commonality apparently lies in the creative act of introducing a new idea not present in the premises from which one starts [8.58, p. 97].

Peirce [8.55, NEM IV, 42 (1902)]:

“What I call the theorematic reasoning of mathematics consists in so introducing a foreign idea, using it, and finally deducing a conclusion from which it is eliminated. Every such proof rests, however, upon judgments in which the foreign idea is first introduced, and which are simply self-evident. As such, they are exempt from criticism.”

Peirce [8.55, NEM IV, 49 (1902)]:

“The peculiarity of theorematic reasoning is that it considers something not implied at all in the conceptions so far gained, which neither the definition of the object of research nor anything yet known about could of themselves suggest, although they give room for it.”

Again, the *foreign idea* is what alludes to abduction, and once it is gained, *self-evident* judgments can be taken in order to prove a theorem. Elsewhere in the same text, Peirce has made clear that these self-evident judgments are, in fact, corollarial deductions [8.55, NEM IV, 38 (1902)]:

“*Theorematic deduction* is deduction in which it is necessary to experiment in the imagination upon the image of the premiss in order from the result of such experiment to make corollarial deductions to the truth of the conclusion.”

All these explanations by Peirce can be accommodated, if theorematic abduction is conceived of as an inverse deduction that infers from the result of corollar-

ial deduction to the premises from which the result can be deductively derived. The similarity with abduction results from the fact that theorematic deduction takes the reasoner to a theoretical point of view, which is the point in the above diagram on inferential reasoning (Fig. 8.1) where abduction would take her. Thus, abduction and theorematic deduction both aim at the same point (Fig. 8.3).

Within this frame of reference, it also becomes clear why Peirce thinks that theorematic deduction is *ampliative*. He just did not call it *ampliative deduction*, because he feared that this labeling would have been considered as unacceptable [8.55, NEM IV, 1 (1901)]:

“It now appears that there are two kinds of deductive reasoning, which might, perhaps, be called *explicative* and *ampliative*. However, the latter term might be misunderstood; for no mathematical reasoning is what would be commonly understood by *ampliative*, although much of it is not what is commonly understood as *explicative*. It is better to resort to new words to express new ideas. All readers of mathematics must have felt the great difference between *corollaries* and *major theorems*.”

The overall process of theorematic deduction can then be analyzed based on the three inferential subprocesses discussed in Sect. 8.2.1, only that the process runs in the inverse direction, starting from a proposition to be proved, say whether  $p$  or  $\neg p$  is logically true (colligation). This is the premise of theorematic deduction, which is then observed in order to find a conceptual point of view from where to derive either  $p$  or  $\neg p$  (observation). Once a candidate for this is found, it has to be established by a corollarial deduction to  $p$  or  $\neg p$ , which is equivalent to *judgment* in the context of theorematic deduction. This is how I understand Peirce when he says [8.55, NEM IV, p. 38 (1902)] (see also [8.12, CP 2.267 (c. 1903)]; [8.59, CP 4.233 (1902)]):

“*Theorematic deduction* is deduction in which it is necessary to experiment in the imagination upon the

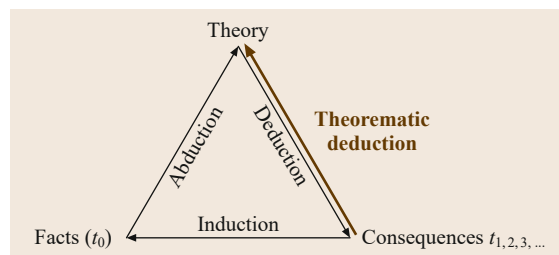
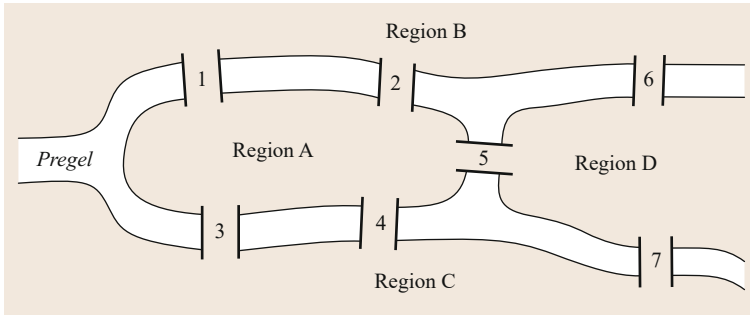


Fig. 8.3 Theorematic deduction in relation to the inferential triad



**Fig. 8.4** The seven bridges of Königsberg

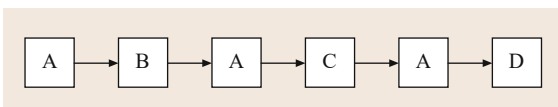
image of the premiss in order from the result of such experiment to make corollarial deductions to the truth of the conclusion.”

Hoffmann, too, stresses that theorematic reasoning (he uses the notion of *theoric transformations*) essentially consists in “looking at facts from a novel point of view” – a phrase taken from [8.60, MS 318] ([8.7, p. 581], [8.48, p. 291], [8.4, p. 181]). And the fact that attaining this novel point of view is first of all the result of *observation* that subsequently has to be subjected to a corollarial deduction as judgment within theorematic deduction may also explain the following passage.

Peirce [8.61, NEM III, p. 869 (1909)]:

“To the Diagram of the truth of the Premisses something else has to be added, which is usually a mere *May-be*, and then the conclusion appears. I call this *Theorematic* reasoning because all the most important theorems are of this nature.”

Ketner [8.57, p. 408] refers to this passage to underpin his view that theorematic deduction is a kind of abduction. However, on my account theorematic deduction is a *May-be*, firstly, in the sense of introducing a theoretical point of view, and secondly, because it is spontaneously generated by *observation* and still has to be submitted to *judgment*. This is my reconstruction of theorematic deduction as inverse deduction. Further refinements might be necessary, in particular analyzing the variants that Levy discusses in [8.58, pp. 98–103]. However, this must be left to a separate analysis. Here, I prefer to provide an instructive example and extend the idea of inverse inferences to include inverse abduction and inverse induction.



**Fig. 8.5** Graph of the state sequence in seven bridges problem

### 8.3.2 An Example for Theorematic Deduction

In addition to Peirce’s examples like Desargues’ theorem (discussed in [8.7, pp. 581–584]), I suggest Leonhard Euler’s solution of the Königsberg bridge problem as a case in point. In Euler’s time, the river Pregel formed the topological shape shown in Fig. 8.4. The question is whether it is possible to pass all seven bridges on a walk while passing each bridge only once.

To solve this problem, Euler used a graph in which the state sequence is shown as transitions from region to region. Figure 8.5 shows how this looks like, if one starts in region A and passes the first five bridges in numerical order. Accordingly, the number of regions in this diagram will always be  $N + 1$ , where  $N$  is the number of all bridges. Moreover, with the five bridges connected to region A, this region is mentioned three times. A so-called *uneven region*, that is, one with an uneven number of bridges, will always appear  $(n + 1)/2$  times in the graph, independently from whether one starts in this very region or in another region. This is different for even regions. If we only consider regions A and B, there are only two bridges. If one starts in A, A is mentioned twice and B only once. If one starts in B, it is the other way round. In general, the region is mentioned  $n/2$  times if one starts outside this region, and  $n/2 + 1$  times if one starts from within.

However, all regions in the seven bridges problem are uneven so that the solution is rather simple. A walk on which one passes each bridge only once encompasses seven transitions between eight states. However, each region must appear  $(n + 1)/2$  times in the diagram, which means three times for region A and two for regions B through D, that is, nine altogether. Hence, the desired walk is impossible.

This example shows that from an abstract topological point of view it is possible to formulate principles from which the impossibility of the specified walk can be deduced. The diagram in Fig. 8.4, together with the question, represents the *colligated* premiss, which

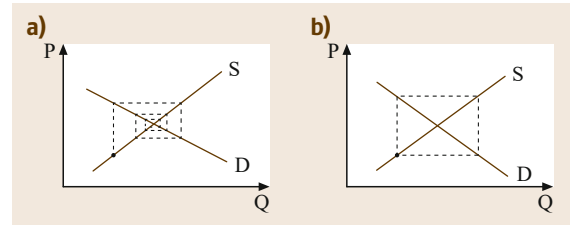
is then observed. The result of this *observation* is the approach represented in the graph in Fig. 8.5 and the further reflections based on it. The final *judgment* consists in deriving the solution, that is, the proof as such.

### 8.3.3 Inverse Abduction and Inverse Induction

Based on the reconstruction of theorematic deduction as an inverse deduction, it follows naturally that there could be two other forms of reasoning: inverse abduction and inverse induction (Fig. 8.6). Moreover, since inverse (theorematic) deduction is similar to abduction in that it aims at the same point in Fig. 8.6, inverse abduction should be similar to induction, and inverse induction should be similar to deduction.

Inverse abduction starts from some theory or abstract concept and searches, for examples, possible instantiations. For instance, the economist *Nicholas Kaldor* [8.62] suggested the cobweb model, which explains how supply (S) and demand (D) develop if time lags are assumed for the reaction of the supply side to a change in demand and vice versa (see Fig. 8.7). If the supply curve is steeper than the demand curve, prices (P) and quantities (Q) will gradually converge to the equilibrium. However, if the slope is the same, supply and demand will fluctuate cyclically.

If it is asked what would be a case in point of such a persistently fluctuating supply and demand, this would require what I call *inverse abduction*. The theoretical model has to be understood on the abstract level, but it is unclear whether there is a concrete case at all to it. An example would be the so-called *pork cycle* that was observed in the 1920s in the United States and in Europe. Kaldor's theoretical model provides a possible explanation for such phenomena, but in this case the argument runs in the opposite direction, from the theory to the case. The similarity to induction consists in the fact that inverse abduction projects a possible explanation onto a case (and



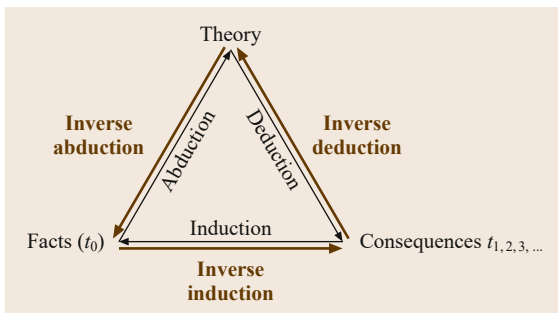
**Fig. 8.7a,b** The cobweb model: (a) successive adjustment of supply and demand; (b) cyclical fluctuations of supply and demand

all other relevant cases one might think of or encounter). The difference is that it is only a provisional projection (if the theory is true), whereas in induction factual truth in the pragmatist sense is established.

The difference can also be stated in this way: inverse abduction starts from the *colligation* of a theoretical model of some sort, which has a meaning but as yet no reference. This theoretical model is *observed* in order to be able to project it onto some case to which it refers (here is the similarity to induction). Finally, it has to be *judged* (abductively) whether the case that suggested itself can really be subsumed to the theoretical model.

Other examples for inverse abduction are riddles that we give children to solve, once they can use concepts independently from concrete references, for example, *What has a face and two hands, but no arms or legs? A clock.* The task is to find something concrete that satisfies this abstract definition. Again, the definition is first colligated, then observed in order to project it onto some concrete object, and the final part consists in the judgment as to whether the definition really applies to the object and whether this inference is thus abductively valid (in this case, as a *possible* circumscription of a clock).

Turning to *inverse induction*, this inference starts from the purported truth (or falsity) of a theory and tries to infer back to a crucial experiment that determines whether the theory would have to be accepted or rejected. This form of reasoning typically applies when two competing approaches stand against each other, in particular when both are well-confirmed by ordinary induction, but are mutually incompatible. One famous example are Bell's inequalities based on the Einstein–Podolsky–Rosen (EPR) Gedankenexperiment. The EPR assumptions entail the fulfilment of Bell's inequalities, quantum theory entails their violation. Hence, the inequalities did not prove anything in themselves, but were the basis for a decisive empirical test, eventually carried out by Alain Aspect, which



**Fig. 8.6** Inverse inferences

established the validity of quantum theory and refuted Einstein in his attempt to save classical physics.

Inverse induction is similar to deduction because it essentially involves deductive steps to derive the decisive experimental conditions. However, as opposed to Peirce's thematic deduction, it does not prove anything, and unlike corollary deduction it does not just derive what follows from a certain theory, but starts from competing theories and the intention to determine which one is true and which one false. From this *col-ligated* premise, the theories are *observed* in order to find the decisive experimental conditions, and the final *judgment* does not concern the deductive validity but whether the test would really be decisive.

Another, much simpler, example is the so-called *Wason* selection task [8.63], one of the most investigated experimental paradigms. This task consists in determining which two of four cards one has to turn over in order to know whether a certain rule is true or false. There are cards with a yellow or a red back, and on their front-sides they have a number, even or odd. Now, there are four cards showing (1) a 3, (2) an 8, (3) a red back, and (4) a yellow back. The rule says that if a card shows an even number on one face, then its op-

posite face is red. So, which two cards have to be turned over to see whether this rule is violated? The solution is the two cards showing the 8 and the yellow back.

Not even 10% get this right (at least in this rather formal context). The reason may be that they fail to see that they need to use modus ponens (*even number*  $\rightarrow$  *red back*) and modus tollens ( $\neg$  (*red back*)  $\rightarrow$   $\neg$  (*even number*)). However, the task is not just to use modus ponens and modus tollens correctly, and therefore it is not just about deduction, as is usually thought. Rather is it the most important part of this reasoning task to *find out* (through *observation*) that these deductive rules allow you to determine which two cards have to be turned over. Moreover, strictly speaking, there are also two competing hypotheses involved: the rule and its negation.

To sum up, all inverse inferences contain elements of its predecessor in the ordinary order, and these elements are important in the *observational* subprocesses. Here, inverse abduction relates to induction, inverse deduction to abduction, and inverse induction to deduction. However, the final judgments are abductive in inverse abduction, deductive in inverse deduction, and inductive in inverse induction.

## 8.4 Discussion of Two Important Distinctions Between Types of Abduction

### 8.4.1 Creative Versus Selective Abduction

In this section, some of the important and/or controversial distinctions between abductions shall be discussed starting with the one between creative and selective abduction (see also [8.64]). Most scholars endorse the view that abduction has to fulfil these two purposes, but *Magnani* [8.3, 4] and *Schurz* [8.6], in particular, discuss them even as separate forms of abduction. However, I oppose this view and think that (1) there is no separate form of abduction that is selective, and (2) that the ways in which abduction might be rightly called *selective* are specific forms of one and the same basic form of abduction. As I have pointed out already in Sects. 8.1.1 and 8.2.1 all other candidates for selective abduction would have to be reinterpreted as forms of induction.

*Selective abduction* is an elusive concept, not only because it is easily confused with IBE, but also because different authors understand it in different ways. In particular, I see a difference in the usage by *Magnani* and *Schurz*, on the one hand, and by *Kapitan* and *Hintikka* on the other hand. As for the latter, *Hintikka* [8.17, p. 503] cites *Kapitan*, who claims that “[t]he purpose of *scientific* abduction is both (i) to gen-

erate new hypotheses and (ii) to select hypotheses for further examination” [8.13, p. 477]. They state this with particular reference to [8.18, CP 6.525], where *creation* and *selection* do not appear as two different kinds of abduction, but as two aspects of one and the same notion of abduction [8.18, CP 6.525 (c. 1901)]:

“The first starting of a hypothesis and the entertaining of it, whether as a simple interrogation or with any degree of confidence, is an inferential step which I propose to call *abduction*. This will include a preference for any one hypothesis over others which would equally explain the facts, so long as this preference is not based upon any previous knowledge bearing upon the truth of the hypotheses, nor on any testing of any of the hypotheses, after having admitted them on probation.”

The last part of the passage ensures that the selective aspect is not confused with induction. Furthermore, Peirce makes clear that selection does not mean separating stupid ideas from sensible ones, because they all have to *explain the facts*, that is, have to be valid abductions. Thus, selection does not refer to the abduc-

tive judgment by which stupid ideas are sorted out. The hypotheses among which to select have already passed this test. However, elsewhere Peirce makes clear that “the whole question of what one out of a number of possible hypotheses ought to be entertained becomes purely a question of economy” [8.18, CP 6.528 (1901)]. Hence, this aspect of selection concerns abduction only from a practical point of view, not from a logical one, as I have argued above in Sect. 8.1.1.

Turning to *Magnani* and *Schurz*, the latter writes [8.6, p. 202]:

“Following *Magnani* (2001, p. 20) I call abductions which introduce new concepts or models *creative*, in contrast to *selective* abductions whose task is to choose the best candidate among a given multitude of possible explanations.”

This sounds as if IBE were included in this notion of selective abduction. However, *Magnani* is careful to distinguish between these, when he discusses what he calls the “two main epistemological meanings of the word abduction” [8.4, p. 10], that is, *creative* and *selective* abduction on the one hand, and IBE on the other hand. Then, he goes on to differentiate between *creative* and *selective* abduction [8.4, p. 10]; see also [8.3, p. 19]:

“An illustration from the field of medical knowledge is represented by the discovery of a new disease and the manifestations of causes which can be considered as the result of a *creative* abductive inference. Therefore, *creative* abduction deals with the whole field of the growth of scientific knowledge. This is irrelevant in medical diagnosis where instead the task is to *select* from an encyclopedia of prestored diagnostic entities.”

As it turns out, *selective abduction* in *Magnani*’s sense is nothing else than the application of previously established knowledge. In this sense, some suitable background knowledge is activated or *selected* vis-à-vis a certain problem. As I understand it, medical diagnosis is only one example; such *selective* abductions seem to be part of everybody’s daily routines. I have discussed abduction as knowledge application above in Sect. 8.1.2, so there is nothing more to add here. On this account, *selective abduction* is to be reconstructed as the abductive step of knowledge application, in particular in the sense that:

1. Specific (explanatory) concepts or theories are activated (selected) from one’s background knowledge, triggered by the initial problem at hand.
2. Accepted as the result of abductive judgment (whereas other spontaneously generated ideas may be rejected as abductively invalid).

3. And, if there are more than one abductively valid ideas, ranked in order of a priori plausibility, however, only for economical reasons.

To be sure, the latter aspect is clearly the least central one, since it is merely of practical importance. And it should be noted that *Magnani* does not attribute it to *selective abduction* when he writes: “Once hypotheses have been selected, they need to be ranked [...] so as to plan the evaluation phase by first testing a certain preferred hypothesis” [8.3, p. 73]. As also *Peirce* warns in [8.18, CP 6.525, see above], it should by no means be confused with inductive reasoning.

This reconstruction of *selective abduction* as the abductive step in knowledge application allows us, finally, to solve the riddle highlighted in the introduction. It concerns what *Peirce* calls *a priori reasoning* in the passage quoted there, and which he associates with his earlier, syllogistic, concept of abduction (i. e., hypothetical reasoning). When *Peirce* explains that this kind of [8.2, CP 8.209 (c. 1905)]:

“abduction is the inference of the truth of the minor premiss of a syllogism of which the major premiss is selected as known already to be true while the conclusion is found to be true,”

1. The major premiss to be *selected* is the theory to which one abduces (e.g.,  $\forall x(Fx \rightarrow Gx)$ ).
2. Based on the conclusion (of the syllogism), *Ga*, which is found to be true and which needs to be explained.
3. And *Fa* results from the assumption that the occurrence of *Ga* is a case of  $\forall x(Fx \rightarrow Gx)$ .

With respect to (3), the only question remaining is whether the abduction runs from *Ga* to  $\forall x(Fx \rightarrow Gx)$ , as I have suggested, or from *Ga* to *Fa*, as *Schurz* [8.6] might perhaps argue based on his notion of *factual abduction*. This is discussed in the following section.

### 8.4.2 Factual Versus Theoretical Abduction

This is how *Schurz* formalizes the basic form of *factual abduction* [8.6, p. 206]:

“*Known Law*: If *Cx*, then *Ex*  
*Known Evidence*: *Ea* has occurred  


---

---

*Abduced Conjecture*: *Ca* could be the reason.”

Let us take an example that *Aliseda* uses in [8.21]. I wake up in the morning in a hotel, look out of the window, and see that the lawn is wet (*w*). Wondering about why it is wet, I think that it might have rained (*r*) or that the sprinklers were on (*s*) last night. Hence, there

are two possible causes,  $r$  and  $s$ . However, the question is whether I abduce to  $r$  and  $s$  or to  $r \rightarrow w$  and  $s \rightarrow w$ , respectively. In my view, both is true in a certain way, which becomes clear if we distinguish inferential subprocesses.

Of course, as we look out of the window and wonder about  $w$  (*colligation*), either  $r$  or  $s$  or both spring to our minds (*observation*). However, since we are looking for an explanation of  $w$ , we are not interested in  $r$  or  $s$  as such, but whether  $w$  because of  $r$  ( $r \rightarrow w$ ) or whether  $w$  because of  $s$  ( $s \rightarrow w$ ). In other words, the law must be implicit in observing the fact, because the fact only makes sense as part of the law. What's more, a spontaneous idea is no valid abduction (not yet). In order to abduce that  $r$  or that  $s$  we have to perform a judgment (explicitly or implicitly) of the type of Schurz's schema. Thus, Schurz's schema fleshes out the abductive judgment in the case of factual abduction. And even though  $r$  or  $s$  may be our spontaneous ideas they are engendered not as such, but as the antecedents of  $r \rightarrow w$  and  $s \rightarrow w$ , respectively.

This may all appear self-evident. However, since factual abduction is basically abduction to known laws and theories (rather than to facts pure and simple), we can unify Schurz's subforms of factual abduction, namely *observable-fact abduction*, *first-order existential abduction*, and *unobservable-fact abduction* [8.6, pp. 27–210]. Moreover, it reveals that Schurz's distinction between factual abduction, on the one hand, and law abduction, on the other hand, does not refer to entirely different forms of abductive inference. The only difference is that law abduction relates to the *creative* abduction of new laws, whereas factual abduction relates to *selective* abduction as the abductive step of the application of known laws. Schurz sanctions this view when he writes [8.6, p. 207]:

“In the setting of factual abduction, the problem consists in the *combinatorial explosion* of the search space of possible causes in the presence of a rich background store of laws but in the *absence* of a rich factual knowledge. Thus, factual abductions are primarily *selective* in the sense of Magnani.”

However, I see yet another problem with this description. It assumes that there is a multitude of possible hypotheses from which one or a few plausible ones have to be chosen. In the very same sense he explains that [8.6, p. 204]:

“in abduction problems we are often confronted with thousands of possible explanatory conjectures (or conclusions) – everyone in the village might be the murderer.”

To my mind, this misrepresents (factual) abduction. For on the one hand, if we take each of the village's inhabitants as a hypothetical candidate for the murderer, and intend to boil down their number by some kind of inference, this would have to be *induction*. On the other hand, if the problem really is to reduce the search space, then we are not dealing with a multitude of conjectures as abductive solutions to some abductive problem (finding the murderer), but we are dealing with a *problem*. The fact that there is a multitude of possibilities changes the situation. The task is not simply to select one of those *hypotheses*, but to come up with a *theory* that explains the murder and identifies particular individuals as suspects.

The deeper truth is that instead of merely selecting we move to higher level of reasoning, just in the sense that I have described in Sect. 8.1.3. The very first level, in the example of the murderer, is that one understands that the very concept of a murder implies that the victim has been killed by someone. Given that there are certain objective restrictions, not every human being can possibly have committed the crime, but just the set of the villagers. The next step is to move to the level of narratives in the sense of a coherent description of what might have happened. However, there might be still too many possibilities, or also none. Yet another step could consist in applying theoretical knowledge as professional profilers do.

As already expounded in Sect. 8.1.3, my suggestion is to reconstruct different forms of abduction in the dimension of theoretical abstraction. Since factual abduction comes out as applied law or theory abduction, there is no fundamental difference between factual and theoretical abduction. However, what should be distinguished systematically are cognitive levels in reasoning, down from elementary cognitive levels captured by forms like *visual* (or *iconic*) and *manipulative* abduction [8.3, 4], and up to high-level abductions like *theoretical model abduction*, *common cause abduction* [8.6], or *trans-paradigmatic abduction* [8.65]. Magnani, Schurz, Hoffmann, and others have done pioneering work explicating abductive inferences at both ends concrete versus abstract cognition, a dimension which I prefer to call hierarchical complexity. However, the precise structures of hierarchical complexity have yet to be revealed (cf. Sect. 8.1.3, above).

One also has to be careful to distinguish forms that do not fit entirely in this order. This seems to apply, for example, to Schurz's notions of (extrapolational) *micropart abduction* and *analogical abduction* [8.6, pp. 216–219]. The former consists, for example, in extrapolating from the behavior of observable macroobjects to assume that unobservable microparts like atoms behave in the same (or a similar) way. However, this is

equivalent to an analogical inference from macro to micro, and as such both do not indicate a certain level of abstraction or complexity, but, following also Peirce, are to be reconstructed as compound inferences (including an abductive and an inductive step to hit the abductive target), as I have tried to reveal in [8.15]. Moreover, Schurz's concept of *hypothetical (common) cause abduction* [8.6, pp. 219–222], where he draws to the *dormitive virtue* example (see Sect. 8.1.3), is, to my mind, no valid form of abduction, since this kind of reasoning establishes a problem (*Why does opium put people to sleep? or What does its dormitive virtue consist of?*), not the solution. It yields the premise of an abductive inference, but not more.

### 8.4.3 Explanatory Versus Nonexplanatory Abduction

However, Schurz [8.6] points to yet another interesting form of abduction, when he discusses “statistical factor analysis” as a kind of “probabilistic common cause abduction” [8.6, pp. 228–231]. He believes that [8.6, p. 228]:

“factor analysis is a certain generalization of hypothetical common cause abduction, although sometimes it may better be interpreted in a purely instrumentalistic way.”

As I have argued a few lines above, hypothetical common cause abduction is no abduction. However, I fully endorse Schurz's interpretation of factor analysis (to be sure, he thinks of *exploratory factor analysis*, not *confirmatory factor analysis*, which also exists). Exploratory factor analysis is a method to reveal correlative structures among numerical representations of empirical items and thus give us a clue as to possible common causes for certain types of effects (the dependent variables). However, factor analyses do not explain anything. *Factors*, once extracted, have to be interpreted, and this is where they are used as hints toward a possible explanation. Hence, their value consists in being instrumental to find interesting patterns in a dataset, but they do not explain anything as such.

Throughout this chapter, I have focused on (forms of) explanatory abduction as the basic purposive context, because this is the received understanding of abduction in general and because most concepts of abductions fall into this category. However, Gabbay and Woods [8.5] have made the point that there are types of reasoning that do not have an explanatory purpose. In particular, they point to abductions that do not aim at a plausible explanation, because they, in fact, “advance propositionally implausible hypotheses” [8.5, p. 115]. The purpose of such abductions cannot be ex-

planatory, but they can serve to fulfil some other kind of purpose.

As an example, Gabbay and Woods discuss Newton's action-at-a-distance theorem, which was never conceived of as an explanatory hypothesis by Newton, since he thought that such an action was causally impossible [8.5, p. 116]. From that point of view, it is clear that in the explanatory context an *action at a distance* poses a problem, that is, that of explaining gravitation, not a solution (like Schurz's hypothetical cause abduction discussed above). However, Gabbay and Woods point out that [8.5, p. 118–119]:

“[t]he action-at-a-distance equation serves Newton's theory in a wholly instrumental sense. It allows the gravitational theory to predict observations that it would not otherwise be able to predict.”

Thus, there are hypotheses that are not set forth in order to explain something, but to serve some practical purpose. Newton used the action-at-a-distance equation as a tool to predict phenomena. Psychologists have used factor analysis as a tool to find basic personality traits (pioneered by [8.66]).

However, technological sciences in general – mechanical, electronic, medical engineering and the like – do not aim at explanations. They aim at practical, though principled, solutions to practical problems. They may be built on explanatory theories, but what they develop does not have to be *true*; it has to be *effective*. Sometimes, technologies have been invented, before the mechanisms that they employed were sufficiently understood (as for instance in the case of x-rays). Moreover, although technological theories are typically based on explanatory theories, the latter ones are *input* in this context and appear in the *colligation* of abductive inferences to technological theories.

For instance, laser technology employs physics in many ways, but the technology itself is *abducted* from these background theories. Before the laser was invented, Charles Townes and Arthur Schawlow developed the *maser* (microwave amplification by stimulated emission of radiation) in 1954 to amplify radio signals. A few years later, the first optical laser was invented. The technological aim was to *produce* focused light, not to explain anything. And even though it was unclear, at the outset, what practical purposes the technology could be used for, it was effective in producing what it was invented for. On top of this, searches for practical applications – of which we know many today – can be easily accommodated as inverse abductions (as suggested in Sect. 8.3.3), from technological theories (here: laser technology) to concrete practical problems which might be solved, either in principle or better than without the technology. Based on technologies, the practical



problems and further background knowledge, the functioning of machines and appliances can be deductively derived, and the machines or prototypes so constructed are then evaluated in terms of effectiveness and efficiency.

Hence, there seems to exist (at least) a second kind of cognitive architecture parallel to the explanatory architecture (and, accordingly, Magnani [8.4, p. 71] is right to claim that Gabbay and Woods' [8.5] notion of instrumental abduction is orthogonal to the forms he distinguishes). On the one hand, explanatory concepts and theories aim at true accounts, and truth is the evaluative criterion for induction. On the other hand, there

are technological theories, which are inductively evaluated in terms of effectiveness and – as far as economic aspects are concerned – efficiency.

However, technological theories seem to be just one domain of reasoning among others to complement explanatory reasoning. At least moral concepts and ethical theories could be a third domain [8.26, pp. 90–101], [8.30], and they are evaluated neither in terms of truth nor effectiveness, but in terms of *justice*. I can only allude to these domains, here, and a separate paper will be necessary to expound these ideas. However, what seems obvious is that there are distinct realms of abduction and of reasoning in general.

## 8.5 Conclusion

To sum up, I have argued (as Peirce did) that there are precisely three basic kinds of inferences: abduction, deduction, and induction. I have distinguished three inferential subprocesses and introduced three inverse types of inference, based on the analysis of inferential subprocesses. My claim is that all kinds of real reasoning ought to be reducible to one of these three basic forms, its inverse forms, or a particular subprocesses within one inferential type. However, I also mentioned analogical reasoning as a special compound form of inferential reasoning and referred the reader to my [8.15].

Moreover, I have tried to point out that apart from these fundamental kinds of reasoning, inferences can be distinguished along two dimensions. One is the dimension of hierarchical complexity so that concepts and theories are built upon one another across cognitive levels, from elementary perception and action to high-

level scientific theories. The other dimension, discussed in the previous section, is that of *domains*. By a *domain* I do not mean, in this context, issues of content to which one and the same theory is applied, but domains of reasoning. In this respect I distinguished explanatory, technological, and moral/ethical concepts and theories.

This framework opens up a taxonomical system that might be able to accommodate the various forms of reasoning in general, and of abduction in particular, that have been suggested so far. I have discussed a few of them, but by far not all. However, my hope is that this taxonomy allows us to account for all a multitude of varieties of abduction, deduction, and induction, while recognizing them in their particular place and function in an overall system and help us to a distinctive understanding of similarities and differences between these individual forms.

## References

- 8.1 C.S. Peirce: *Pragmatism and Pragmaticism*, Collected Papers of Charles Sanders Peirce, Vol. 5, ed. by C. Hartshorne, P. Weiss (Harvard Univ. Press, Cambridge 1934)
- 8.2 C.S. Peirce: *Reviews, Correspondence, and Bibliography*, Collected Papers of Charles Sanders Peirce, Vol. 8, ed. by A.W. Burks (Harvard Univ. Press, Cambridge 1958)
- 8.3 L. Magnani: *Abduction, Reason, and Science: Processes of Discovery and Explanation* (New York, Kluwer 2001)
- 8.4 L. Magnani: *Abductive Cognition: The Epistemological and Eco-Cognitive Dimensions of Hypothetical Reasoning* (Springer, Berlin 2009)
- 8.5 D.M. Gabbay, J. Woods: *The Reach of Abduction – Insight and Trial*, A Practical Logic of Cognitive Systems Ser., Vol. 2 (Elsevier, Amsterdam 2005)
- 8.6 G. Schurz: Patterns of abduction, *Synthese* **164**, 201–234 (2008)
- 8.7 M.H.G. Hoffmann: Theoric transformations and a new classification of abductive inferences, *Trans. C. S. Peirce Soc.* **46**, 570–590 (2010)
- 8.8 C.S. Peirce: *Mathematical Miscellanea*, The New Elements of Mathematics by Charles S. Peirce, Vol. III/2, ed. by C. Eisele (Mouton, The Hague, 1976)
- 8.9 C.S. Peirce: MS 754 (1907). In: *Annotated Catalogue of the Papers of Charles S. Peirce*, ed. by R.S. Robin (Univ. Massachusetts Press, Amherst 1967), available online (3 July 2016): <http://www.iupui.edu/~peirce/robin/robin.htm>
- 8.10 G. Minnameier: Peirce–Suit of truth: Why inference to the best explanation and abduction ought not to be confused, *Erkenntnis* **60**, 75–105 (2004)

- 8.11 S. Paavola: Hansonian and harmanian abduction as models of discovery, *Int. Stud. Philos. Sci.* **20**, 93–108 (2006)
- 8.12 C.S. Peirce: *Elements of Logic*, Collected Papers of Charles Sanders Peirce, Vol. 2, ed. by C. Hartshorne, P. Weiss (Harvard Univ. Press, Cambridge 1932)
- 8.13 T. Kapitan: Peirce and the structure of abductive inference. In: *Studies in the Logic of Charles Sanders Peirce*, ed. by N. Houser, D.D. Roberts, J.V. Evra (Indiana Univ. Press, Bloomington 1997) pp. 477–496
- 8.14 C.S. Peirce: *Principles of Philosophy*, Collected Papers of Charles Sanders Peirce, Vol. 1, ed. by C. Hartshorne, P. Weiss (Harvard Univ. Press, Cambridge 1931)
- 8.15 G. Minnameier: Abduction, induction, and analogy – On the compound character of analogical inferences. In: *Model-Based Reasoning in Science and Technology: Abduction, Logic, and Computational Discovery*, ed. by W. Carnielli, L. Magnani, C. Pizzi (Springer, Heidelberg 2010) pp. 107–119
- 8.16 S. Psillos: An explorer upon untrodden ground: Peirce on abduction. In: *Handbook of the History of Logic*, Vol. 10, ed. by D.M. Gabbay, S. Hartmann, J. Woods (Elsevier, Amsterdam 2011) pp. 117–151
- 8.17 J. Hintikka: What is abduction? The fundamental problem of contemporary epistemology, *Trans. C. S. Peirce Soc.* **34**, 503–533 (1998)
- 8.18 C.S. Peirce: *Scientific Metaphysics*, Collected Papers of Charles Sanders Peirce, Vol. 6, ed. by C. Hartshorne, P. Weiss (Harvard Univ. Press, Cambridge 1935)
- 8.19 C.S. Peirce: *Science and Philosophy*, Collected Papers of Charles Sanders Peirce, Vol. 7, ed. by A.W. Burks (Harvard Univ. Press, Cambridge 1958)
- 8.20 D.J. McKaughan: From ugly duckling to swan: C. S. Peirce, abduction, and the pursuit of scientific theories, *Trans. C. S. Peirce Soc.* **44**, 446–468 (2008)
- 8.21 A. Aliseda: *Abductive Reasoning – Logical Investigations into Discovery and Explanation* (Springer, Dordrecht 2006)
- 8.22 D.G. Campos: On the distinction between Peirce's abduction and Lipton's inference to the best explanation, *Synthese* **180**, 419–442 (2011)
- 8.23 A. Mackonis: Inference to the best explanation, coherence and other explanatory virtues, *Synthese* **190**, 975–995 (2013)
- 8.24 C. Hookway: *The Pragmatic Maxim: Essays on Peirce and Pragmatism* (Oxford Univ. Press, Oxford 2012)
- 8.25 C. Hookway: Truth, reality, and convergence. In: *The Cambridge Companion to Peirce*, ed. by C. Misak (Cambridge Univ. Press, Cambridge 2004) pp. 127–149
- 8.26 C. Misak: *Truth, Politics, Morality: Pragmatism and Deliberation* (Routledge, London 2000)
- 8.27 I. Levi: Beware of syllogism: Statistical reasoning and conjecturing according to peirce. In: *The Cambridge Companion to Peirce*, ed. by C. Misak (Cambridge Univ. Press, Cambridge 2004) pp. 257–286
- 8.28 G. Minnameier: What's wrong with it? – Kinds and inferential mechanics of reasoning errors. In: *Learning from Errors*, ed. by J. Seifried, E. Wuttke (Verlag Barbara Budrich, Opladen 2012) pp. 13–29
- 8.29 G. Minnameier: Deontic and responsibility judgments: An inferential analysis. In: *Handbook of Moral Motivation: Theories, Models, Applications*, ed. by F. Oser, K. Heinrichs, T. Lovat (Sense, Rotterdam 2013) pp. 69–82
- 8.30 G. Minnameier: A cognitive approach to the 'Happy Victimiser', *J. Moral Educ.* **41**, 491–508 (2012)
- 8.31 C.S. Peirce: MS 692 (1901). In: *Annotated Catalogue of the Papers of Charles S. Peirce*, ed. by R.S. Robin (Univ. Massachusetts Press, Amherst 1967), available online (3 July 2016): <http://www.iupui.edu/~peirce/robin/robin.htm>
- 8.32 N.R. Hanson: *Patterns of Discovery* (Univ. of Cambridge Press, Cambridge 1958)
- 8.33 R. Carnap: Testability and meaning, *Philos. Sci.* **3**, 419–471 (1936)
- 8.34 R. Carnap: Testability and meaning, *Philos. Sci.* **4**, 1–40 (1937)
- 8.35 E. McMullin: *The Inference that Makes Science* (Marquette Univ. Press, Milwaukee 1992)
- 8.36 P. Thagard: Coherence, truth, and the development of scientific knowledge, *Philos. Sci.* **74**, 28–47 (2007)
- 8.37 T.A.F. Kuipers: Laws, theories, and research programs. In: *Handbook of the Philosophy of Science: General Philosophy of Science – Focal Issues*, ed. by T.A.F. Kuipers (Elsevier, Amsterdam 2007) pp. 1–95
- 8.38 G. Schurz: Explanation as unification, *Synthese* **120**, 95–114 (1999)
- 8.39 W. Park: How to learn abduction from animals? – From avicenna to magnani. In: *Model-Based Reasoning in Science and Technology – Theoretical and Cognitive Issues*, ed. by L. Magnani (Springer, Berlin 2014) pp. 207–220
- 8.40 C. El Khachab: The logical goodness of abduction in C.S. Peirce's thought, *Trans. C. S. Peirce Soc.* **49**, 157–177 (2013)
- 8.41 L.W. Barsalou: Grounded cognition, *Annu. Rev. Psychol.* **59**, 617–645 (2008)
- 8.42 J.J. Zeman: Peirce on abstraction. In: *The Relevance of Charles Peirce*, ed. by E. Freeman (The Hegeler Institute, La Salle, IL 1983) pp. 293–311
- 8.43 J. Piaget, R. Garcia: *Psychogenesis and the History of Science* (Columbia Univ. Press, New York 1989)
- 8.44 L.W. Barsalou: The human conceptual system. In: *The Cambridge Handbook of Psycholinguistics*, ed. by M. Spivey, K. McRae, M. Joannis (Cambridge Univ. Press, New York 2012) pp. 239–258
- 8.45 P. Thagard: Cognitive architectures. In: *The Cambridge Handbook of Cognitive Science*, ed. by K. Frankish, W. Ramsay (Cambridge Univ. Press, Cambridge 2012) pp. 50–70
- 8.46 G. Minnameier: A new stairway to moral heaven – A systematic reconstruction of stages of moral thinking based on a Piagetian logic of cognitive development, *J. Moral Educ.* **30**, 317–337 (2001)
- 8.47 S. Paavola: Peircean abduction: Instinct or inference, *Semiotica* **153**, 131–154 (2005)
- 8.48 M.H.G. Hoffmann: Problems with Peirce's concept of abduction, *Found. Sci.* **4**, 271–305 (1999)
- 8.49 F. Poggiani: What makes a reasoning sound? C.S. Peirce's normative foundation of logic, *Trans. C. S. Peirce Soc.* **48**, 31–50 (2012)
- 8.50 K.T. Fann: *Peirce's Theory of Abduction* (Martinus Nijhoff, The Hague 1970)

- 8.51 T. Kapitan: Peirce and the autonomy of abductive reasoning, *Erkenntnis* **37**, 1–26 (1992)
- 8.52 J. Hintikka: *Socratic Epistemology: Explorations of Knowledge-Seeking by Questioning* (Cambridge Univ. Press, Cambridge 2007)
- 8.53 F. Suppe: Science without induction. In: *The Cosmos of Science: Essays of Exploration*, ed. by J. Earman, J.D. Norton (Univ. Pittsburgh Press, Pittsburgh 1997) pp. 386–429
- 8.54 J. Earman: *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory* (MIT Press, Cambridge 1992)
- 8.55 C.S. Peirce: *Mathematical Philosophy*, The New Elements of Mathematics by Charles S. Peirce, Vol. IV, ed. by C. Eisele (Mouton, The Hague, 1976)
- 8.56 J. Hintikka: C. S. Peirce's first real discovery and its contemporary relevance, *Monist* **63**, 304–315 (1980)
- 8.57 K.L. Ketner: How Hintikka misunderstood Peirce's account of theorematic reasoning, *Trans. C. S. Peirce Soc.* **21**, 407–418 (1985)
- 8.58 S.H. Levy: Peirce's theoremical/corollary distinction and the interconnections between mathematics and logic. In: *Studies in the Logic of Charles Sanders Peirce*, ed. by N. Houser, D.D. Roberts, J.V. Evra (Indiana Univ. Press, Bloomington 1997) pp. 85–110
- 8.59 C.S. Peirce: *The Simplest Mathematics*, Collected Papers of Charles Sanders Peirce, Vol. 4, ed. by C. Hartshorne, P. Weiss (Harvard Univ. Press, Cambridge 1933)
- 8.60 C.S. Peirce: MS 318 (1907). In: *Annotated Catalogue of the Papers of Charles S. Peirce*, ed. by R.S. Robin (Univ. Massachusetts Press, Amherst 1967), available online (3 July 2016): <http://www.iupui.edu/~peirce/robin/robin.htm>
- 8.61 C.S. Peirce: *Mathematical Miscellanea*, The New Elements of Mathematics by Charles S. Peirce, Vol. III/1, ed. by C. Eisele (Mouton, The Hague, 1976)
- 8.62 N. Kaldor: A classificatory note on the determination of equilibrium, *Rev. Econ. Stud.* **1**, 122–136 (1934)
- 8.63 P.C. Wason: Self-contradictions. In: *Thinking: Readings in Cognitive Science*, ed. by P.N. Johnson-Laird, P.C. Wason (Cambridge Univ. Press, Cambridge 1977) pp. 114–128
- 8.64 G. Minnameier: Abduction, selection, and selective abduction. In: *Model-Based Reasoning in Science and Technology: Logical, Epistemological and Cognitive Issues*, ed. by L. Magnani, C. Casadio (Springer, Berlin, Heidelberg 2016)
- 8.65 F.V. Hendricks, J. Faye: Abducting explanation. In: *Model-Based Reasoning in Scientific Discovery*, ed. by L. Magnani, N.J. Nersessian, P. Thagard (Kluwer, New York 1999) pp. 271–294
- 8.66 R.B. Cattell: *The Description and Measurement of Personality* (World Book, New York 1946)

# Magnani's M

## 9. Magnani's Manipulative Abduction

Woosuk Park

Despite the extensive research in logic, cognitive science, artificial intelligence, semiotics, and philosophy of science, there is no sure proof that we have better or deeper understanding of abduction than its modern founder, Charles S. Peirce. In this sense, one of the most important developments in recent studies on abduction is Lorenzo Magnani's discovery of manipulative abduction. In this paper, I shall examine in what ways Magnani goes with and beyond Peirce in his views on manipulative abduction. After briefly introducing his distinction between theoretical and manipulative abduction (Sect. 9.1), I shall discuss how and why Magnani counts diagrammatic reasoning in geometry as the prime example of manipulative abduction (Sect. 9.2). Though we can witness an increasing interest in the role of abduction and manipulation in what Peirce calls theorematic reasoning, Magnani is unique in equating theorematic reasoning itself as abduction. Then, I shall discuss what he counts as some common characteristics of manipulative abductions (Sect. 9.3), and how and why Magnani views manipulative abduction as a form of practical reasoning (Sect. 9.4). Ultimately, I shall argue that it is manipulative abduction that enables Magnani to extend abduction to all directions to develop the eco-cognitive model of abduction. For this purpose, fallacies and animal abduction will be used as examples (Sect. 9.5).

9.1	<b>Magnani's Distinction Between Theoretical and Manipulative Abduction</b> .....	197
9.2	<b>Manipulative Abduction in Diagrammatic Reasoning</b> .....	198
9.2.1	Abductive and Manipulative Aspects of Diagrammatic Reasoning.....	198
9.2.2	Magnani on Manipulative Abduction in Diagrammatic Reasoning.....	201
9.3	<b>When Does Manipulative Abduction Take Place?</b> .....	203
9.4	<b>Manipulative Abduction as a Form of Practical Reasoning</b> .....	204
9.5	<b>The Ubiquity of Manipulative Abduction</b> .....	206
9.5.1	Manipulative Abduction in Fallacies.....	206
9.5.2	Manipulative Abduction in Animals.....	207
9.6	<b>Concluding Remarks</b> .....	212
	<b>References</b> .....	212

### 9.1 Magnani's Distinction Between Theoretical and Manipulative Abduction

It is certainly worthwhile to take a view of Magnani's multiple distinctions of abduction:

1. Selective/creative
2. Theoretical/manipulative
3. Sentential/model based.

Above all, our focal interest lies in understanding the relationships between these three distinctions. The

most revealing seems to be the following text [9.1, p. 11]:

“What I call *theoretical abduction* certainly illustrates much of what is important in creative abductive reasoning, in humans and in computational programs, especially the objective of selecting and creating a set of hypotheses (diagnoses, causes, hy-

potheses) that are able to dispense good (preferred) explanations of data (observations), but fails to account for many cases of explanations occurring in science and in everyday reasoning when the exploitation of environment is crucial. [...] I maintain that there are two kinds of theoretical abduction, *sentential*, related to logic and to verbal/symbolic inferences, and *model based*, related to the exploitation of internalized models of diagrams, pictures, etc.”

This text is important because it presents a tentative definition of theoretical abduction and the subdivision of theoretical abduction into sentential and model-based abductions. The passage above says that theoretical abduction can be a kind of creative reasoning (not only selective, as in diagnosis).

Insofar as Magnani views theoretical abduction also as a kind of creative abduction, and again insofar as he tries to distinguish between theoretical and manipulative abductions, he must also view manipulative abduction as a kind of creative abduction. *Magnani* introduces the concept of manipulative abduction as follows [9.1, p. 12] (cf. [9.2, pp. 15,16,43]):

“The concept of *manipulative abduction* captures a large part of scientific thinking where the role of action is central, and where the features of this action are implicit and hard to be elicited: Action can provide otherwise unavailable information that enables the agent to solve problems by starting and by performing a suitable abductive process of generation or selection of hypotheses.”

For my present purpose, the following text is more informative [9.1, p. 39] (cf. [9.2, p. 53]):

“*Manipulative abduction* [9.2] – contrasted with theoretical abduction – happens when we are thinking through doing and not only, in a pragmatic sense, about doing. [...] Manipulative abduction

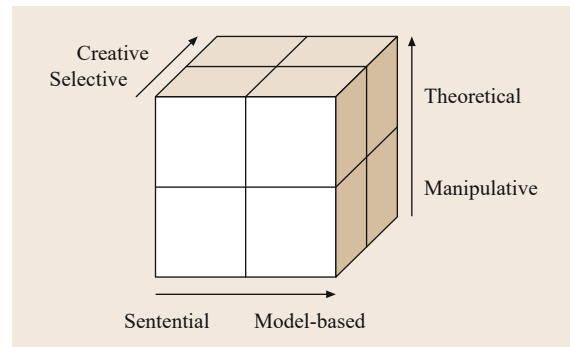


Fig. 9.1 Magnani’s classification of abduction: Cubic model

refers to an extra-theoretical behavior that aims at creating communicable accounts of new experiences to integrate them into previously existing systems of experimental and linguistic (theoretical) practices.”

As is clear from this quote, manipulative abduction is contrasted with theoretical abduction by Magnani.

Further, in *Magnani*’s writings it is also stressed that manipulative abduction is occurring taking advantage of those model-based (e.g., iconic) aspects that are embedded in external models [9.1, p. 58]:

“We have seen that manipulative abduction is a kind of abduction, usually model based and so intrinsically *iconic*, that exploits external models endowed with delegated (and often implicit) cognitive and semiotic roles and attributes.”

This line of thought clearly indicates a possibility that Magnani’s multiple distinctions of abduction may work in such a way that each distinction represents a different dimension in our understanding of abduction. What I have in mind might be presented crudely as in the cubic model of Magnani’s classification of abduction (Fig. 9.1).

## 9.2 Manipulative Abduction in Diagrammatic Reasoning

### 9.2.1 Abductive and Manipulative Aspects of Diagrammatic Reasoning

In Peirce’s distinction between corollarial and theorematic reasoning, there are some typical texts in Peirce’s writings frequently invoked by the commentators. For example, Hoffmann cites [9.3, NEM 4:38] in [9.4, p. 290] (cf. [9.5, CP 2.267], [9.6, CP 7.204] and [9.3, NEM 4:288]):

“Corollarial deduction is where it is only necessary to imagine any case in which the premises are true in order to perceive immediately that the conclusion holds in that case [...] Theorematic deduction is deduction in which it is necessary to experiment in the imagination upon the image of the premiss in order from the result of such experiment to make corollarial deductions to the truth of the conclusions.”

Marietti cites the following texts as “the classic description given by Peirce of the two types of deduction” [9.7, CP 2.267]:

“A corollarial deduction is one that represents the conditions of the conclusion in a diagram and finds from the observation of this diagram, as it is, the truth of the conclusion. A Theorematic Deduction is one which, having represented the conditions of the conclusion in a diagram, performs an ingenious experiment upon the diagram, and by the observation of the diagram, so modified, ascertains the truth of the conclusion.”

Marietti [9.7, p. 120, NEM 4:42]:

“What I call the *theorematic* reasoning of mathematics consists in so introducing a foreign idea, using it, and finally deducing a conclusion from which it is eliminated.”

CP 4.233 seems, however, the most extensive text pertinent to Peirce's distinction between corollarial/theorematic distinction in broader perspective. I discussed this text rather extensively in [9.8].

In their treatment of these typical Peircean texts, recently many commentators have discussed diagrammatic reasoning in connection with abduction. For example, Hoffmann [9.4, p. 411] writes (cf. [9.9, p. 337], [9.10, p. 69], [9.11, p. 465] and [9.4, pp. 292–293]):

“The creativity of theorematic reasoning and the role of observation in it support the interpretation that there must be, for Peirce, a connection between this form of deductive reasoning and abduction. Both, at least, seem to fulfill the same task. What theorematic deduction is for mathematics, abduction seems to be for scientific discoveries in general. Thus, Ketner (1985) maintained ‘that production of experiments within theorematic reasoning, on Peirce's view, is done through abduction’.”

It is interesting to note that, while Ketner clearly invokes abduction in the “experiments within theorematic reasoning”, Hoffmann himself is merely making an analogy between theorematic deduction in mathematics and abduction in science. Campos [9.12] is also somewhat similar to Hoffmann's case [9.12, p. 135]:

“In mathematical reasoning, the imagination creates experimental diagrams that function as signs that are then perceived, interpreted, judged, often transformed, re-imagined, re-interpreted, and so on, in a continuous process. Experimental hypotheses are imaginative suggestions that become subject to logical scrutiny as possible keys to the solution of a theorematic deduction. Once conceived, the

experimental diagrams act like objects for observation; they now resist the mind, as it were, and must be evaluated as solutions to the mathematical problem. In this respect, this process is akin to abduction in the natural sciences.”

Marietti [9.7] seems to go one step further in this regard, for she explicitly mentions the necessity of abduction in theorematic demonstration in mathematics [9.7, p. 124]:

“In my view, it is possible to identify an operation in thought – it takes place essentially above the diagram and involves precisely the perceptual relations organized by it – which forms the core of the abductive inference that makes the demonstration synthetic, theorematic, creative. This is the point: There has to be an abductive inference in every informative demonstration given that a connecting thread running through Peirce's thoughts on the logic of science is that ‘[a]ll the ideas of science come to it by the way of abduction’ (CP 5.145). In a theorematic demonstration – that is in a demonstration introducing new knowledge into our mathematical system – it is necessary to carry out an abductive passage.”

However, Marietti leaves it unclear how it is possible to have “synthetic, theorematic, creative” mathematical demonstration. So, one might say that Marietti is here detecting the indispensability of abduction in theorematic reasoning.

Stjernfelt [9.13] is clearly much more informative as to how abduction takes place in theorematic reasoning [9.13, p. 276]:

“An important issue here – both related to the *addition of new elements or foreign ideas* and to the *experiment* aspects – is the relation between theorematic reasoning and abduction. A finished piece of theorematic reasoning, of course, is deductive – the conclusion follows with necessity from the the premises. But in the course of conducting the experiment, an abductive phase appears when investigating which experimental procedure, among many, to follow; *which* new elements or foreign ideas to introduce. This may require repeated, trial-and-error abductive guessing, until the final structure of the proof is found – maybe after years or centuries. Exactly the fact that neither premises nor theorems need to contain any mentioning of the experiment or the introduction of new elements makes the abductive character of experimentation clear. Of course, once the right step has been found, abductive searching may cease and the deductive character of the final proof stands out.”

Here, Stjernfelt makes it exactly clear when and where abduction intervenes in theorematic reasoning. When conducting the experiments with diagrams, “an abductive phase appears”. He is also quite explicit about what are abduced at that phase: i.e., “which experimental procedure to follow”, and “*which* new elements or foreign ideas to introduce”. His view seems attractive in that at least it rather persuasively appeasing the apparent conflict between theorematic deduction and abduction. Though “a finished piece of theorematic reasoning, of course, is deductive”, abductive phase appears in doing experiments with diagrams. At the same time, he also explains why it has been difficult to notice abductive phase in theorematic reasoning, for “once the right step has been found, abductive searching may cease and the deductive character of the final proof stands out”. He is perceptive enough to point out that [9.13, p. 276]:

“Exactly the fact that neither premises nor theorems need to contain any mentioning of the experiment or the introduction of new elements makes the abductive character of experimentation clear.”

A few recent commentators are also to be congratulated for their discussion of manipulation in diagrammatic reasoning. After pointing out that “for Peirce mathematics is a science of observation and experimentation upon diagrams akin to the physical sciences”, for example, Campos draws our attention to the following text [9.12, p. 129] (cf. [9.14, CP 4.530]):

“One can make experiments upon uniform diagrams, and when one does so, one must keep a bright lookout for unintended and unexpected changes thereby brought about in the relations of different significant parts of the diagram to one another. Such operations upon diagrams, whether external or imaginary, take the place of the experiments upon real things that one performs in chemical and physical research. Chemists have ere now, I need not say, described experimentation as the putting of questions to Nature. Just so, experiments upon diagrams are questions put to the Nature of the relations concerned.”

In case of experiments on diagrams, it may be hard not to notice the manipulative character, for in experimenting on diagrams by doing certain operations we are doing nothing but manipulations. Indeed, Campos is able to locate a text, in which Peirce actually uses the language of *manipulation* extensively [9.15, CP 3.363, emphasis is mine]:

“As for algebra, the very idea of the art is that it presents formulae which can be *manipulated*, and

that by observing the effects of such *manipulation* we find properties not to be otherwise discerned. In such *manipulation*, we are guided by precious discoveries which are embodied in general formulae. These are the patterns which we have the right to imitate in our procedure, and are the icons par excellence of algebra [Emphasis is mine].”

Among the commentators of Peirce, Marietti seems to be the one who highlights *manipulation on diagrams* in more detail. She is rather explicit in presenting *manipulation on diagrams* as the core of mathematical proofs, and thereby mathematics itself [9.16, p. 166]:

“Manipulation and observation of diagrammatic signs characterize Peirce’s idea of mathematical reasoning. The more that such reasoning leads to relevant conclusions, the more manipulation and observation play a key role in it.”

Marietti [9.7, p. 112]:

“It is well known that Peirce conceived mathematics as a semiotic activity using particular kinds of signs, which he called diagrams. In his view, the work of a mathematician consists entirely in observing and manipulating these diagrammatic signs.”

Marietti’s commentaries are quite recommendable in that they enable us to understand not only that manipulation on diagrams is everywhere in mathematics, but also why it has such an important role in mathematics. For example, she writes [9.16, p. 151]:

“In order to experiment on a diagram, in fact, we must face a single instance of it, that is to say, a spatial and temporal object, be it actually drawn on a blackboard or rather only scribed in our mind. To look for a suitable strategy of demonstration (to look for the *foreign idea* required in order to get to the conclusion) means to manipulate a concrete, individual diagram according to established rules, until the new cognition, whatever it might be, appears on it. Manipulation, action, concrete experimentation: this is what a mathematical proof consists in.”

Here, Marietti painstakingly elaborates what is involved in diagrammatic reasoning. Experiment on diagrams is nothing but manipulation on them, which requires “to manipulate a concrete, individual diagram”. Indeed, both equations, that is, *experimenting is manipulating*, and *manipulation is manipulation on individual diagrams* are emphasized by Marietti, as is clear from the following [9.7, p. 121]:

“It is even more understandable, however, in the case of the most fertile type of deduction, theore-

matic deduction, in which it is not only a question of observing the relations that emerge but also to *carry out* some operations on this single model, on the individual *token*. What is to be done is to manipulate this sign, to implement a series of ingenious experiments – hence creative and not mechanical – aimed at finding the correct modification so as to cause to emerge the new relations that we then observe. Such concrete work presupposes the concrete quality of a material sign.”

*Marietti* further hints at what is involved in *manipulation on individual diagrams*. *Sun-Joo Shin* seems to be another important contributor to the research on Peircean diagrammatic reasoning, who emphasizes the aspect of individuality in diagrams [9.17, 18]. [9.16, p. 153]:

“The second step of the demonstration introduces that spatiotemporal level which is indispensable in view of the manipulation of the diagram. The *ekthesis* consists in the individualization of the initial proposition, the *protasis*, which is expressed in general terms.”

*Marietti* [9.16, p. 154]:

“Demonstrating means manipulating individual diagrams. In order to experiment on a diagram, as we said above, we must face a single instance of it. Deduction cannot work on general signs alone.”

*Marietti* [9.16, p. 155]:

“The individual diagrammatic sign really acts upon us. We manipulate it, and it reacts on us concretely showing some new relations that impose themselves upon our understanding, escaping any doubt.”

In sum, some commentators of Peirce's distinction between corollarial and theorematic reasoning indeed detect abductive and manipulative aspects of diagrammatic reasoning in geometry. None of them, however, invokes *manipulative abduction* in diagrammatic reasoning.

### 9.2.2 Magnani on Manipulative Abduction in Diagrammatic Reasoning

Magnani already dealt with the role of model based and manipulative abductions in geometrical reasoning in several places [9.1, 2, 19, 20]. Since there have been many attempts to understand Peirce's philosophy of mathematics focusing on his distinction between corollarial and theorematic reasoning, as we saw earlier, Magnani's results in model based and manipulative abduction can be easily combined with previous results

in diagrammatic reasoning in geometry. For example, Peircean corollarial reasoning would be model based, theoretical, and visual abduction. On the other hand, Peircean theorematic reasoning would be model based, manipulative, and visual abduction [9.1, pp. 117–118, 176–178].

In assimilating Magnani's views of manipulative abduction to Peirce's theory of diagrammatic reasoning in geometry, the most difficult point would be that Peircean corollarial and theorematic reasonings are usually counted as deductions. *Magnani* claims that “theorematic deduction can be easily interpreted in terms of manipulative abduction” [9.1, p. 178]. However, it is not clear what he has in mind. Probably, further hints can be secured from the following quote from *Magnani* [9.1, p. 181]:

“As I have already indicated Peirce further distinguished a *corollarial* and a *theoric* part within *theorematic reasoning*, and connected theoric aspects to abduction [Hoffmann, 1999, p. 293]: *Théoric reasoning* [...] is very plainly allied to what is normally called abduction [Peirce, 1966, 754, ISP, p. 8].”

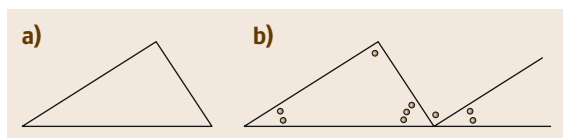
As Hoffmann points out, however, there can be different possible interpretations of what Peirce says “either he identified abduction/retroduction and theoric reasoning here or he claimed that there is abduction in mathematics beyond theoric deduction” [9.4, p. 293].

Though extremely interesting and significant, it is not my present concern to answer whether we can safely understand Peirce's theoric reasoning as abduction, as *Magnani* claims. What is at stake is rather how and why *Magnani* views diagrammatic reasoning in geometry as the prime example of manipulative abduction. If so, by assuming that we now understand the basics of *Magnani*'s notion of manipulative abduction, we need to raise the following further questions. Why does he appeal to diagrammatic reasoning in geometry whenever he has to give an example of manipulative abduction? What exactly does *Magnani* mean by manipulative abduction in geometrical reasoning? What exactly do mathematicians do when they experiment on diagrams? What kind of things could be manipulated in mathematicians' manipulative abduction? What is it for to do manipulative abduction?

*Magnani* uses Fig. 9.2 as an example of cognitive manipulating in diagrammatic demonstration. According to him, this example, taken from the field of elementary geometry, shows how [9.1, p. 176]:

“a simple manipulation of the triangle in Fig. 9.2a gives rise to an external configuration – Fig. 9.2b – that carries relevant semiotic information about the





**Fig. 9.2a,b** Diagrammatic demonstration that the sum of the internal angles of any triangle is 180. (a) Triangle (b) diagrammatic manipulation/construction (after [9.1, p. 176])

internal angles of a triangle *anchoring* new meanings”

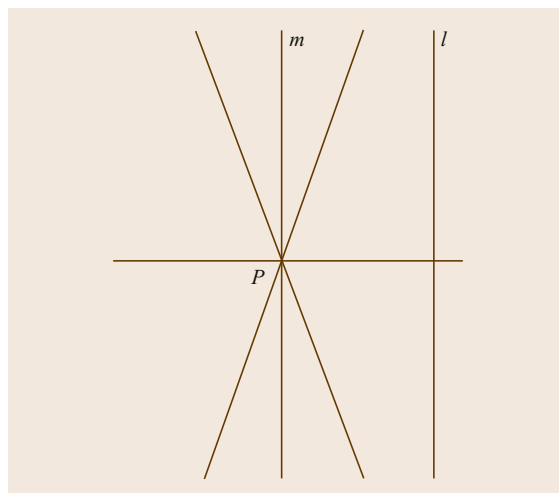
What Magnani counts as cognitive manipulating here is “the entire process through which an agent arrives at a physical action”. And he claims that the very process can be understood by means of the concept of manipulative abduction. He continues [9.1]:

“In this perspective manipulative abduction is a specific case of cognitive manipulating in which an agent, when faced with an external situations from which it is hard or impossible to extract new meaningful features of an object, elects or creates an action that structures the environment in such a way that it gives new information which would be otherwise unavailable and which is used specifically to infer explanatory hypotheses.”

The reason why diagrammatic reasoning in geometry is instrumental for Magnani to introduce *manipulative abduction* is not hard to understand. From his point of view, it is important to first note that “mathematical diagrams play various roles in a typical abductive way”. Secondly, they are external representations that provide both explanatory and nonexplanatory abductive results [9.1, pp. 118–119]. The first point can be elaborated by the following [9.1, p. 118]:

“Following the approach in cognitive science related to the studies in distributed cognition, I contend that in the construction of mathematical concepts many external representations are exploited, both in terms of diagrams and of symbols. I have been interested in my research in diagrams which play an *optical* role – microscopes (that look at the infinitesimally small details), telescopes (that look at infinity), windows (that look at a particular situation), a *mirror* role (to externalize rough mental models), and an *unveiling* role (to help create new and interesting mathematical concepts, theories, and structures).”

Also, the second point is further explained by Magnani as follows.



**Fig. 9.3** Euclidean parallel line (after [9.1, p. 120])

Two of them are central [9.1, pp. 118–119]:

- They provide an intuitive and mathematical *explanation* able to help the understanding of concepts difficult to grasp or that appear obscure and/or epistemologically unjustified. I will present in the following section some mirror diagrams which provided new mental representations of the concept of parallel lines.
- They help abductively *create* new previously unknown concepts that are *nonexplanatory*, as illustrated in the case of the discovery of the non-Euclidean geometry.

As we can infer from the passage just quoted, the discovery of non-Euclidean geometry provides Magnani an ideal springboard to elaborate his views on manipulative abduction in diagrammatic reasoning. In fact, *Magnani* [9.1] uses Lobachevsky’s discovery of non-Euclidean geometry as an example, in which manipulative abduction played a crucial role. After briefly narrating what happened to the parallel postulate of Euclidean geometry throughout the history, he explains Lobachevsky’s strategy to face the problem situation as follows [9.1, p. 123]:

“Lobachevsky’s strategy for resolving the anomaly of the fifth postulate was first of all to manipulate the symbols, second to rebuild the principles, and then to derive new proofs and provide a new mathematical apparatus; of course his analysis depended on some of the previous mathematical attempts to demonstrate the fifth postulate. The failure of the demonstrations – of the fifth postulate from the other four – that was present to the attention of Lobachevsky, lead him to believe that the difficulties that had to be overcome were due to causes

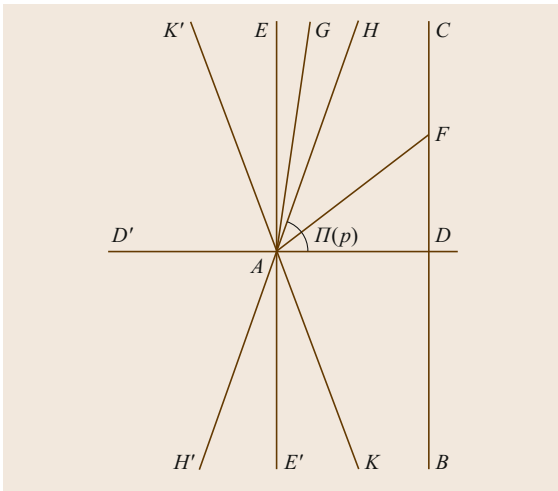


Fig. 9.4 Non-Euclidean parallel lines (after [9.1, p. 131])

traceable at the level of the first principles of geometry.”

According to Magnani, we can detect the anomaly of the fifth postulate from Fig. 9.3, for, unlike other four postulates, “we cannot verify empirically whether two lines meet, since we can draw only segments, not lines” [9.1, p. 121].

In contrast to the diagram of Fig. 9.3, as Magnani explains, the diagram of Fig. 9.4 introduces a new definition of parallelism [9.1, p. 13]:

“All straight lines which in a plane go out from a point can, with reference to a given straight line in the same plane, be divided in two classes – into *cutting* and *not-cutting*. The boundary lines of the one and the other class of those lines will be called *parallel to the given lines*.”

See Magnani [9.1, p. 131] for further details:

“The external representation is easily constructed like in of [Lobachevsky, Figure 2.11, Prop. 16, p. 13], where the angle  $HAD$  between the parallel  $HA$  and the perpendicular  $AD$  is called the angle of parallelism, designated by  $\Pi(p)$  for  $AD = p$ . If  $\Pi(p)$  is  $< 12\pi$ , then upon the other side of  $AD$ , making the same angle  $DAK = \Pi(p)$  will lie also a line  $AK$ , parallel to the prolongation  $DB$  of the line  $DC$ , so that under this assumption we must also make a distinction of sides in parallelisms”

Magnani explicitly claims that Lobachevsky’s inference here “to rebuild the first principles of geometry is prevalently a kind of *manipulative* and *model-based* abduction” [9.1, 127]. As Magnani aptly points out, “Lobachevsky’s target is to perform a geometrical abductive process able to create the new and very abstract concept of non-Euclidean parallel lines”, and what is remarkable is that “the whole epistemic process is mediated by interesting manipulations of external mirror diagrams” [9.1, p. 123].

### 9.3 When Does Manipulative Abduction Take Place?

Earlier we saw how Magnani introduces the distinction between theoretical and manipulative abduction, and how he counts diagrammatical reasoning in geometry as an example of manipulative abduction. In order to capture the essence of manipulative abduction, however, we may need a bit more systematic desiderata. In other words, one might want an individually necessary but jointly sufficient set of conditions for manipulative abduction. Does Magnani present some such things in addition to his frequently invoked claim that “*manipulative* abduction happens when we are thinking *through* doing and not only, in a pragmatic sense, about doing?” [9.1, p. 46].

The answer seems to be positive, for, at least Magnani presents what he calls “some common features of the tacit templates of manipulative abduction [...] that enable us to manipulate things and experiments in science are related to” [9.1, pp. 47–48]:

“1. sensibility toward the aspects of the phenomenon which can be regarded as *curious* or *ano-*

*malous*; manipulations have to be able to introduce potential inconsistencies in the received knowledge [...]; 2. preliminary sensibility toward the *dynamical* character of the phenomenon, and not to entities and their properties, common aim of manipulations is to practically reorder the dynamic sequence of events into a static spatial one that should promote a subsequent bird’s eye view [...]; 3. referral to experimental manipulations that exploit *artificial apparatus* to free new possibly stable and repeatable sources of information about hidden knowledge and constraints [...]; 4. various contingent ways of epistemic acting: *looking* from different perspectives, checking the different information available, comparing subsequent events, *choosing*, *discarding*, *imaging* further manipulations [...].”

Further, Magnani finds it interesting that manipulative abductions are present in mathematics in the sense that geometrical constructions indeed satisfy all these requirements [9.1, p. 49].

However, it would be prudent not to conclude so fast, for it is not clear which comes first. In a lengthy paragraph just quoted selectively, Magnani uses the well-known history of electromagnetism in the nineteenth century in order to make sense of many requirements for manipulative abduction. Magnani uses Oersted's report of his experiment about electromagnetism as an example of 1, for it described some anomalous aspects. Also, he uses Davy's setup of using an artificial tower of needles as an example of 3. *Magnani* seems indebted here to *Gooding's* views of the roles of construals in science [9.1, pp. 48–51] and [9.21, pp. 29–69]. But, in order for the episodes in that history to be examples of manipulative abduction, should not there be first pre-established set of requirements for manipulative abduction? Magnani seems delighted to find geometrical constructions satisfy all these requirements. But he does not first discover manipulative abduction in geometrical constructions, and sort out the commonalities of manipulative abductions, and not the other way around? There is some suspicion as to possible circularity in Magnani's way of thinking. Furthermore, not to mention the fact that it is by no means clear what relations hold between these requirements, there is room for doubt whether each of the requirements is evidently an individually necessary condition for manipulative abduction.

Some such worries make us wonder whether Magnani simply enumerates some interesting traits that appear in cases of manipulative abduction. Magnani

seems still collecting interesting cases that deserve to be characterized as manipulative abductions, and identifying their interesting traits. In other words, he has just drawn our attention to when and by what manipulative abductions occur. What is interesting is, if we are on the right track, such a search to find cases of manipulative abduction and their typical characteristics might result in a disjunctive property, whose extension could be rather huge. Though it might be interesting and meaningful to pursue such a property, it is definitely not necessary and sufficient for manipulative abduction. Contrary to what one might believe Magnani is doing, he may be after entirely different target. Roughly speaking, Magnani seems to aim at demonstrating the ubiquity or pervasiveness of manipulative abduction. What I have in mind could become clearer by “the example discussed above”. It is rather impressive that Magnani is able to uncover all the various iconic roles in geometrical diagrams, such as optical, mirror, unveiling roles. These different iconic roles are related to different types of representations. Further, there is no end to the synthesis or multiplication of these representations, for [9.1, p. 49]:

“[t]he various procedures for manipulating objects, instruments and experiences will be in their turn reinterpreted in terms of procedures for manipulating concepts, models, propositions, and formalisms”

## 9.4 Manipulative Abduction as a Form of Practical Reasoning

In order to sharpen the notion of manipulative abduction in contrast to theoretical abduction, we need some defining characteristics that appropriately capture the essence of manipulative abduction. But we by no means have exhausted the possible cases of manipulative abduction. We seem to rather find the ubiquity or pervasiveness of manipulative abduction. So, we seem to face a dilemma or at least an essential tension: If we stop somewhere satisfied with a set of conditions that fits with some typical common characteristics of manipulative abductions, we might run the risk of ignoring or neglecting some interesting cases of manipulative abduction; On the other hand, if we indefinitely continue our search for all different types of manipulative abduction, and try to subsume all of them under the general rubric of manipulative abduction, it could become difficult to find any that is definitely not a case of manipulative abduction.

I believe that Magnani is also well aware of such a dilemma or an essential tension in his investigations

into manipulative abduction. He seems to make a decision to do an inductive search for important cases of manipulative abduction first, thereby postponing to give a rigorous definition or a necessary and sufficient conditions for manipulative abduction. My hunch may be supported by Magnani's two interesting moves. First, in Chapter 7, “Abduction in human and logical agents: Hasty generalizers, hybrid abducers, fallacies” of his book on abductive cognition, he introduces an illuminating new perspective on manipulative abduction, according to which we can view manipulative abduction “as a form of practical reasoning” [9.1, p. 362; see also p. 384]:

“What has been called manipulative abduction in the previous chapters will be re-interpreted as a form of practical reasoning, a better understanding of which can furnish a description of human beings as hybrid reasoners in so far they are users of ideal (logical) and computational agents.”

There are several interesting points to note regarding Magnani's new perspective of manipulative abduction as a form of practical reasoning. For convenience's sake, let us distinguish between scientific contexts and the contexts of ordinary life. As is clear from our discussion earlier, manipulative abduction plays important roles in scientific contexts. But with this new perspective, we may deepen our understanding of some of the characteristics of manipulative abduction in science. According to Magnani, for example, the first three among the new characteristics of manipulative abduction are also found in geometrical constructions. It may not be irrelevant to invoke, in this regard, *John Woods'* apt characterization of Magnani [9.22, p.240]:

“At the centre of Magnani's investigations is the reasoning of the *practical agent*, of the individual agent operating *on the ground*, that is, in the circumstances of real life. In all its contexts, from the most abstractly mathematical to the most muckily empirical, Magnani emphasizes the cognitive nature of abduction.”

If Woods is right, then it is not a small matter that manipulative abduction can be interpreted as a form of practical reasoning. Magnani is shifting our focus from more theoretical and abstract aspects of science to more practical and experimental aspects of science by his emphasis on manipulative abduction. Now, if we turn to the world of everyday reasoning with the new perspective of manipulative abduction as a form of practical reasoning, we may realize how indispensable manipulative abduction is for individual agents in virtually every moments and situations in real life. Again, *Woods'* succinct summary of the pages 363–384 of *Magnani* [9.1] is to the point [9.22]:

“In an original thrust, he identifies the practical agent as a cognitive system whose resources are comparatively scant and who sets his cognitive targets with due regard (and respect) for these resource-limitations.”

Magnani's second move is also impressive. In addition to the common characteristics of manipulative abduction discussed earlier, Magnani identifies some other common characteristics of manipulative abduction from the perspective of manipulative abduction as a form of practical reasoning [9.1, pp. 51–52]:

“5. Action elaborates a *simplification* of the reasoning task and a redistribution of effort across time [9.23], when we need to manipulate concrete things in order to understand structures which are otherwise too abstract [9.24], or when we are in presence of *redundant* and unmanageable information;

6. Action can be useful in presence of incomplete or inconsistent information – not only from the *perceptual* point of view – or of a diminished capacity to act upon the world: it is used to get more data to restore coherence and to improve deficient knowledge;
7. Action enables us to build *external artifactual models* of task mechanisms instead of the corresponding internal ones, that are adequate to adapt the environment to the agent's needs: Experimental manipulations exploit *artificial apparatus* to free new possible stable and repeatable sources of information about hidden knowledge and constraints.
8. Action as a *control of sense data* illustrates how we can change the position of our body (and/or of the external objects) and how to exploit various kinds of prostheses (Galileo's telescope, technological instruments and interfaces) to get various new kinds of stimulation: action provides some tactile and visual information (e.g., in surgery), otherwise unavailable.”

As individual agents with scant resources to manage and survive in complicated and unfriendly environments, we can understand without much difficulty what Magnani is talking about in this quote. Furthermore, as I shall show by some examples in the next section, Magnani's recent research, including not only those works directly concern abduction (such as *Magnani* [9.1]) but also virtually all other works apparently dealing with other subject matters (such as *Magnani* [9.25–27]) can be interpreted as examining the roles of manipulative abduction in all the different areas. Before turning to examples, let us briefly examine what new aspects of manipulative abduction are introduced by the characteristics 5 through 8 in addition to characteristics 1 through 4. The most salient point would be that *action-based* character of these characteristics are emphasized by Magnani. In some sense, there seems to be one–one correspondence between the new characteristics 5 through 8 and the old characteristics 5 through 8. For example, in both 3 and 7, *artificial apparatus* is invoked. The only difference is that in the latter externality of artifactual models is emphasized. Likewise, in both 4 and 8, some contingent ways of acting are invoked. The only difference is that, unlike epistemic acting in the former (e.g., looking from different perspective), real acting is done in the latter (e.g., control of sense-data by changing the position of the body). In other words, Magnani seems to be insinuating that, when we interpret manipulative abduction as a form of practical reasoning, thereby exploiting some action, we can find the new characteristics that were in some sense already pregnant in the old characteristics.

The following quote from *Magnani* [9.1] seems perfectly supports my understanding by clearly combining his two moves [9.1, p. 397]:

“Human beings spontaneously (and also animals, like already Peirce maintained) perform more or less rudimentary abductive and inductive reasoning. Starting from the low-level inferential performances of the kid’s hasty generalization that is a strategic success and a cognitive failure human beings arrive to the externalization of *theoretical* inductive and abductive agents as *ideal agents*, logical and

computational. It is in this way that *merely successful strategies* are replaced with successful strategies that also tell the *more precise truth* about things. Human informal nondemonstrative inferential processes of abduction (and of induction) are more and more externalized and objectified: These external representations can be usefully rerepresented in our brains (if this is useful and possible), and they can originate new improved organic (mentally internal) ways of inferring or suitably exploited in a hybrid manipulative interplay, as I have said above.”

## 9.5 The Ubiquity of Manipulative Abduction

In order to fathom Magnani’s mind in expanding the scope of manipulative abduction by treating it as a form of practical reasoning, the distinction between linguistic and prelinguistic agent could be at least equally fruitful as the distinction between scientific and everyday context. As he wants to find manipulative abduction in everyday as well as scientific contexts, he also wants to find it in prelinguistic as well as linguistic agents. How does Magnani expand manipulative abduction in science to manipulative abduction in everyday context? How does Magnani expand manipulative abduction in linguistic agents to manipulative abduction in prelinguistic agents? A nice strategic point to get an overview of both lines of expanding the scope of manipulative abduction may be secured in Chapter 7 of Magnani [9.1] entitled “Abduction in human and logical agents: Hasty generalizers, hybrid abducers, fallacies”. On the one hand, this chapter provides us with a clear example of nonexplanatory abduction, thereby representing the shift from the scientific to the practical. On the other hand, it presents us a foil for the search of manipulative abduction in nonhuman, prelinguistic agents. In both respects, interestingly, Magnani seems to be strongly influenced by and responding to *Gabbay* and *Woods* [9.28] and *Woods* [9.29].

### 9.5.1 Manipulative Abduction in Fallacies

As pointed out by *Park* [9.30], we can witness the recent surge of interest in classifying different patterns or types of abduction. Many philosophers, including *Thagard*, *Magnani*, *Gabbay* and *Woods*, *Schurz*, and *Hoffmann*, have suggested their own classifications emphasizing different aspects of abduction [9.1, 2, 28, 31–33]. Such a development is remarkable, in view of the fact that until quite recently the focus of the research on Peircean abduction was basically to identify its log-

ical form [9.34]. Among these contributions, *Gabbay* and *Woods* [9.28] seem most instrumental in expanding our purview by introducing nonexplanatory abductions. This is important, for as is clear from *Schurz* [9.32] and *Hoffmann* [9.33], attempts at classifying abduction is still largely focusing on the problem of classifying explanatory abduction in science. Not to mention nonexplanatory abduction in science, such as instrumental abduction, *Gabbay* and *Woods* [9.28] covers abduction in nonscientific context, such as legal abduction. *Magnani* welcomes *Gabbay* and *Woods*’ distinction between explanatory and nonexplanatory abduction as follows [9.1, p. 71]:

“In my previous book on abduction [9.2] I made some examples of abductive reasoning that basically are nonexplanatory and/or instrumentalist without clearly acknowledging it. The contribution of *Gabbay* and *Woods* to the analysis of abduction has the logical and epistemological merit of having clarified these basic aspects of abduction, until now disregarded in the literature. Their distinction between explanatory, nonexplanatory and instrumental abduction is orthogonal to mine in terms of the theoretical and manipulative (including the subclasses of sentential and model based) and further allows us to explore fundamental features of abductive cognition.”

*Magnani* is also strongly influenced by *Woods*’ extensive and revolutionary study of fallacies. Above all, *Magnani* is fully sympathetic with *Woods*’ project of the naturalization of logic, the official core topic of which is the one of logical fallacies [9.35, p. 20]. What is needed here is just to understand how *Magnani* adopts and appropriates *Woods*’ views of fallacies for his own eco-cognitive project. *Woods* contends, and *Magnani* confirms that fallacy has been counted as “a mistake in

reasoning, a mistake which occurs with some frequency in real arguments and which is characteristically deceptive" [9.1, p. 404] and [9.29]. However, Magnani points out that [9.1, pp. 404–405]:

“when they are used by actual reasoners, *beings like us*, that is in an eco-logical and not merely logical – ideal and abstract – way, they are *no longer* necessarily fallacies.”

Magnani agrees with Woods' conviction that from Aristotle onward logic has irremediably mismanaged the fallacies project. And he concurs with Woods' belief that naturalization of logic is appropriate to the task of “an account of fallacious reasoning – and of its detection, avoidance, and repair” [9.35, 36]. What Woods calls *EAUI-conception* of fallacies is the traditional perspective of fallacies that “fallacies are *Errors of Reasoning*, Attractive, Universal, and Incorrigible” [9.36, p. 135], [9.35, p. 21]. Now, Magnani reports Woods' views of fallacies as follows [9.35, p. 22]:

“According to Woods' last and more recent observations the traditional fallacies – hasty generalization included – do not really instantiate the traditional concept of fallacy (the EAUI-conception). In this perspective it is not that it is sometimes strategically justified to commit fallacies (a perfectly sound principle, by the way), but rather that in the case of the *Gang of Eighteen* traditional fallacies they simply are not fallacies. The distinction is subtle, and I can add that I agree with it in the following sense: The traditional conception of fallacies adopts – so to say – an *aristocratic* (ideal) perspective on human thinking that disregards its profound eco-cognitive character. Errors, in an eco-cognitive perspective, certainly are not the exclusive fruit of the so-called fallacies, and in this wide sense, a fallacy is an error – in Woods' words – ‘that virtually everyone is disposed to commit with a frequency that, while comparatively low, is nontrivially greater than the frequency of their errors in general’.”

By the term *Gang of Eighteen*, Woods refers to following typical fallacies [9.36, p. 5]:

“*ad baculum*, *ad hominem*, *ad populum*, *ad verecundiam*, *ad ignorantiam*, *ad misericordiam*, affirming the consequent, denying the antecedent, begging the question, gambler's fallacy, *post hoc, ergo propter hoc*, composition and division (of which *secundum quid* is a special case), faulty analogy, and *ignoratio elenchi* (of which straw man is a special case).”

Magnani's eco-cognitive perspective draws a rather sharp distinction between strategic and cognitive ra-

tionality, and he explicitly claims that “many of the traditional fallacies – hasty generalization for example – call for an equivocal treatment” [9.35, p. 21] and [9.1]. What he means by “an equivocal treatment” is that the so-called fallacies [9.1]:

“are sometimes cognitive mistakes and strategic successes, and in at least some of those cases, it is more rational to proceed strategically, even at the cost of cognitive error.”

Magnani also claims that his general agreement with Woods' views of fallacies can be further strongly motivated by his emphasis on what he calls the general *military* nature of language, that is:

1. Human language possesses a *pregnance-mirroring* function.
2. In this sense we can say that vocal and written language is a tool exactly like a knife.
3. The so-called fallacies, are certainly linked to that efficacious *military intelligence*, which relates to the problem of the role of language in the so-called *coalition enforcement*, which characterizes all the various kinds of groups and collectives of humans [9.35, p. 22].

Indeed Magnani contends that in this perspective language is basically rooted in a kind of *military intelligence*, a term coined by the mathematician *René Thom* [9.37], the creator of the so-called catastrophe theory. See Chap. 8 of *Magnani* [9.1] for more in-depth study of military intelligence and the notion of coalition enforcement.

Certainly most people would believe that communication is the primary function of language. Also, when broadly understood, communication might include the manipulation of other human beings by language. One possible danger is that whenever we talk about communicative function as the primary function of language, we tend to ignore or neglect the manipulative function of language. Clever and sometimes malicious uses of fallacies in order to manipulate other human beings, definitely there is military intelligence involved. In a word, we may say that manipulative abduction is crucial in understanding the role of fallacies in military intelligence.

## 9.5.2 Manipulative Abduction in Animals

I emphasized the central importance of animal abduction in Magnani's thought in a series of papers [9.30, 38, 39]. This section draws extensively from these, especially [9.38]. Unlike these previous articles, this time I want to highlight the role of manipulative abduction in animal cognition. One of the most pressing issues in

understanding abduction is whether it is an instinct or an inference. For many commentators find it paradoxical “that new ideas and hypotheses are products of an instinct (or an insight), and products of an inference at the same time” [9.40, p. 131]. Paavola refers to [9.41–46]. As Paavola points out, we seem to face a dilemma: “If abduction relies on instinct, it is not a form of reasoning, and if it is a form of reasoning, it does not rely on instinct” [9.40, p. 131]. Fortunately, Lorenzo Magnani’s recent discussion of animal abduction sheds light on both instinctual and inferential character of Peircean abduction (Magnani [9.1, especially Chapter 5], “Animal abduction: From mindless organisms to artifactual mediators”, which was originally published in Magnani *an Li* [9.25]). Contrary to many commentators, who find conflicts between abduction as instinct and abduction as inference, he claims that they simply co-exist.

In order to overcome the conflict between abduction as an instinct and abduction as an inference, it is not enough to draw attention to some relevant texts from Peirce and to provide insightful interpretation of them. Magnani needs to indicate exactly where he is going beyond Peirce, thereby pointing out wherein lies the limitation of Peirce’s views on abduction. It is of course an important matter for Magnani himself whether he is going *beyond* Peirce or not [9.1, p. 221]. Magnani finds such a clear example from Peirce’s different treatments of practical reasoning and scientific thinking [9.1, pp. 278–279]:

“Elsewhere Peirce seems to maintain that instinct is not really relevant in scientific reasoning but that it is typical of just the reasoning of practical men about every day affairs. So as to say, we can perform instinctive abduction (that is not controlled, not *reasoned*) in practical reasoning, but this is not typical of scientific thinking.”

Here Magnani quotes extensively from Peirce’s Carnegie application of 1902 (MS L75) (cf. Arisbe Website [9.47].) We should note that Magnani is fully aware of the fact that we can find many instances where Peirce allowed abductive instinct to humans even in scientific reasoning. For example, hypothesis selection is a largely instinctual endowment of human beings which Peirce thinks is given by God or related to a kind of Galilean *lume naturale* [9.1, p. 277] and [9.6, CP 7.220]:

“It is a primary hypothesis underlying all abduction that the human mind is akin to the truth in the sense that in a finite number of guesses it will light upon the correct hypothesis.”

Magnani counts commentators like [9.4, 40, 48] as maintaining that “instinct [...] *does not* operate at the

level of conscious inferences like for example in the case of scientific reasoning” [9.1, p. 279]. And he implicitly blames their assumption of instinct “as a kind of mysterious, not analyzed, guessing power” for such a claim [9.1]. Indeed Magnani distances himself from those commentators and Peirce himself as follows [9.1]:

“I think a better interpretation is the following that I am proposing here: Certainly instinct, which I consider a simple and not a mysterious endowment of human beings, is at the basis of both *practical* and scientific reasoning, in turn instinct shows the obvious origin of both in natural evolution.”

But on what ground does Magnani claim superiority of his interpretation? How could he be so sure that instinct is at the basis of both practical and scientific reasoning? Even though Magnani does not formulate an argument that proves his claim once and for all, there seem to be enough clues for fathoming his mind. In addition to his reliance on the naturalistic ground for abductive instinct in humans, Magnani is also attracted to the so-called synechism of Peirce. Further, he seems encouraged by two intriguing points from Peirce: (1) that “thought is not necessarily connected with brain”, and (2) that “instincts themselves can undergo modifications through evolution” [9.1, p. 278]. For the first point, Magnani actually quotes from Peirce [9.14, CP 4.551]:

“Thought is not necessarily connected with brain. It appears in the work of bees, of crystals, and throughout the purely physical world; and one can no more deny that it is really there, than that the colours, the shapes, etc. of objects are really there.”

On the other hand, for the second point, he again quotes from Peirce: [instincts are] “inherited habits, or in a more accurate language, inherited dispositions” [9.1, p. 278] and [9.5, CP 2.170].

In other words, Magnani seems to assimilate abduction as an instinct and abduction as an inference form both directions. This interpretation of Magnani’s strategy seems to be supported strongly by his explicit announcement [9.1, p. 267]:

“I can conclude that instinct versus inference represents a conflict we can overcome simply by observing that the work of abduction is partly explicable as a biological phenomenon and partly as a more or less *logical* operation related to *plastic* cognitive endowments of all organisms.”

To those who would allow abductive instinct to nonhuman animals but not to humans, he tries to emphasize the instinctual elements in human abductive reasoning. On the other hand, to those who would allow

abduction as inference to humans but not to nonhuman animals, he suggests to broaden the concept of inference, and thereby that of thinking. For the former project, Magnani cites hypothesis generation in scientific reasoning as a weighty evidence for abductive instinct in humans: From this Peircean perspective, hypothesis generation is a largely instinctual and non-linguistic endowment of human beings and, of course, also of animals. It is clear that for Peirce abduction is rooted in the instinct and that many basically instinctually rooted cognitive performances, like emotions, provide examples of abduction available to both human and nonhuman animals [9.1, p. 286]. Here, of course, Magnani's claim about *hypothesis generation* as instinctual must be still controversial. Someone may object that it should be able to work out what might explain a phenomenon. For further discussion of this complicated issue, please see [9.1, pp. 18–19]. In this regard, Magnani distinguishes between “(1) abduction that only generates plausible hypothesis (*selective or creative*)” and “(2) abduction considered as *inference to the best explanation*, that also evaluates hypotheses by induction” [9.1, p. 18, Magnani's emphasis]. And, he makes it explicit that the first meaning of abduction is what he accepts in his epistemological model. Though inconclusive, Magnani's claim about *hypothesis generation* as instinctual is more defensible under the first meaning of abduction. Even after having noted these supportive points, however, it is still unclear how abduction could be rooted in instinctual-rooted cognitive performances like emotion.

As for the latter project, Magnani wants to secure inferential character of animal abduction from sign activity and semiotic processes found in nonhuman animals. He frequently appeals to Peirce [9.49, CP 5.283]:

“all thinking is in signs, and signs can be icons, indices or symbols. Moreover, all inferences are a form of sign activity, where the word sign includes feeling, image, conception, and other representation.”

Here is a lengthy quote from Magnani that makes this point crystal clear [9.1, p. 288] and [9.5, 14, CP 5.283]:

“Many forms of thinking, such as imagistic, emphatic, trial and error, and analogical reasoning, and cognitive activities performed through complex bodily skills, appear to be basically model based and manipulative. They are usually described in terms of living beings that adjust themselves to the environment rather than in terms of beings that acquire information from the environment. In this sense these kinds of thinking would produce

responses that do not seem to involve sentential aspects but rather merely *noninferential* ways of cognition. If we adopt the semiotic perspective above, which does not reduce the term *inference* to its sentential level, but which includes the whole arena of sign activity – in the light of Peircean tradition – these kinds of thinking promptly appear full, inferential forms of thought. Let me recall that Peirce stated that all thinking is in signs, and signs can be icons, indices, or symbols, and, moreover, all inference is a form of sign activity, where the word sign includes *feeling, image, conception, and other representation.*”

Magnani is well aware of the fact animals have been widely considered as mindless organisms for a long time. So, based on the cornerstone laid by Peirce, this semiotic perspective needs further extension. But how is it possible? According to Magnani, that is possible thanks to the recent results in cognitive science and ethology about animals, and of developmental psychology and cognitive archeology of humans and infants. [9.1, p. 283]:

“Philosophy itself has for a long time disregarded the ways of thinking and knowing of animals, traditionally considered *mindless* organisms. Peircean insight regarding the role of abduction in animals was a good starting point, but only more recent results in the fields of cognitive science and ethology about animals, and of developmental psychology and cognitive archeology about humans and infants, have provided the actual intellectual awareness of the importance of the comparative studies.”

Magnani not only points out that inferences are not necessarily structured like a language, but also there are animal-like aspects in human thinking and feeling. [9.1, p. 283]:

“Sometimes philosophy has anthropocentrically condemned itself to partial results when reflecting upon human cognition because it lacked in appreciation of the more *animal-like* aspects of thinking and feeling, which are certainly in operation and are greatly important in human behavior.”

Encouraged by the discovery of “the ways of thinking in which the *sign activity* is of a nonlinguistic sort” [9.1, p. 189] in lower animals, Magnani claims that “a higher degree of abductive abilities has to be acknowledged” to them [9.1, pp. 290,291]:

“Chicken form separate representations faced with different events and they are affected by prior experiences (of food, for example). They are mainly



due to internally developed plastic capabilities to react to the environment, and can be thought of as the fruit of learning. In general this plasticity is often accompanied by the suitable reification of external artificial *pseudo representations* (for example landmarks, alarm calls, urine marks and roars, etc.) which artificially modify the environment, and/or by the referral to externalities already endowed with delegated cognitive values, made by the animals themselves or provided by humans.”

In fact, Magnani goes even farther in his ascription of pseudo thought to nonhuman animals in his discussion of *affordances*, *multimodal abduction*, *cognitive niches*, and *animal artifactual mediators*. It is exactly where we can find what Magnani believes to be a clear evidence that manipulative abduction plays a crucial role in animal sign activities.

As Magnani points out, we are now in a much better position than Peirce to understand the way of thinking and knowing of animals, thanks to [9.1, p. 283]:

“more recent results in the fields of cognitive science and ethology about animals, and of developmental psychology and cognitive archeology about humans and infants.”

But Magnani reminds us of the fact that Darwin already paved a way toward the appreciation of cognitive faculties of animals [9.1, pp. 284-285]:

“It is important to note that Darwin also paid great attention to those external structures built by worms and engineered for utility, comfort, and security. I will describe later on in this chapter the cognitive role of artifacts in both human and nonhuman animals. Artifacts can be illustrated as *cognitive mediators* [9.2] which are the building blocks that bring into existence what it is now called a *cognitive niche*: Darwin maintains that ‘We thus see that burrows are not mere excavations, but may rather be compared with tunnels lined with cement’ [9.50, p. 112]. Like humans, worms build external artifacts endowed with precise roles and functions, which strongly affect their lives in various ways, and of course their opportunity to know the environment.”

I would like to discuss a strong or active sense of learning abduction from animals. I do interpret Magnani’s ideas on perceiving affordance in human and nonhuman animals as an answer to the problem of how to learn abduction from animals in this sense. As far as the problem of perceiving affordances is concerned, we do not have to confess our inferiority to nonhuman animals. It is we humans who have perceived affordances in some highly creative ways. However, we cannot eas-

ily claim our superiority over nonhuman animals either. It is roaches not humans that turn out to demonstrate better ability for survival, which may imply superiority in perceiving affordances. In a word, I think we may safely and more profitably forget the issue of inferiority or superiority. Let it suffice to say that we humans, unlike nonhuman animals, seem to have very unique abductive instinct displayed by our perceiving affordances.

Magnani would be happy with my interpretation, for he himself claims that “cognitive niche construction can be considered as one of the most distinctive traits of human cognition” [9.2, p. 331]. According to Magnani, both human and nonhuman animals are chance seekers, and thereby ecological engineers. They “do not simply live their environment, but actively shape and change it looking for suitable chances” [9.1, p. 319]. Further, “in doing so, they construct cognitive niches” [9.1]. Then, in chance seeking ecological engineering in general, and in cognitive niche construction in particular, what exactly does differentiate humans from nonhuman animals?

In order to answer this question, we need to understand in what respects Magnani extends or goes beyond Gibson’s notion of affordance. In principle, it should not be too difficult, because Magnani himself indicates explicitly or implicitly some such respects of his own innovation. Magnani takes Gibson’s notion of affordance “as what the environment offers, provides, or furnishes” as his point of departure. He also notes that Gibson’s further definitions of affordance as [9.1, p. 333]:

- “1. Opportunities for action.
2. The values and meanings of things which can be directly perceived.
3. Ecological facts.
4. Implying the mutuality of perceiver and environment.”

may contribute to avoiding possible misunderstanding. Given this Gibsonian ecological perspective, Magnani appropriates some further extensions or modifications by recent scholars in order to establish his own extended framework for the notion of affordance. It is simply beyond my ability to do justice to all elements of Magnani’s extended framework for affordances. Let me just note one issue in which Magnani shows enormous interest, Gibsonian direct perception.

Magnani takes Donald Norman’s ambitious project of reconciling constructivist and ecological approaches to perception seriously [9.1, 6.4.3, p. 343]. Above all, Magnani notes that Norman “modifies the original Gibsonian notion of affordance also involving

mental/internal processing" [9.1, p. 337] based on a text, where *Norman* writes [9.51, p. 14]:

"I believe that affordances result from the mental interpretation of things, based on our past knowledge and experience applied to our perception of the things about us."

If Norman is right, we may safely infer, as Magnani does, that pace Gibson, "affordances depend on the organism's experience, learning, and full cognitive abilities" [9.1, p. 337]. Both Norman and *Magnani* are evidencing these ideas by formidable array of recent results in cognitive experimental psychology and neuroscience [9.1, p. 341].

Now, given this extended framework for that extends and modifies some aspects of the original Gibsonian notion of affordances, what exactly is Magnani's contribution? In some sense, this is an unnecessary stupid question, for everybody already knows the correct answer. By his expertise on abduction, and in particular his Peircean thesis of perception as abduction, Magnani contributes enormously to deepen our understanding of some truly big issues, such as how to reconcile constructivist and ecological theories of perception. So, my question aspires to understand more specifically how the Peirce–Magnani view of perception as abduction contributes in that regard. Let us suppose that the original Gibsonian notion of affordance has been extended and modified à la Norman. Would Magnani claim that such an extension or modification is impossible without abductive activities of organisms? Or, would he claim that such an extension or modification is still incomplete without abduction?

Be that as it may, the big picture *Magnani* presents is this [9.1, p. 348]:

"Organisms have at their disposal a standard endowment of affordances (for instance through their hardwired sensory system), but at the same time they can extend and modify the scope of what can afford them through the suitable cognitive abductive skills."

If we probe the question as to what exactly are involved in organisms' employment of cognitive abductive skills, *Magnani* would respond roughly as the following lines [9.1, p. 346]:

"in sum organism already have affordances available because of their instinctive gifts, but also they can dynamically abductively *extract* natural affordances through *affecting* and modifying perception

(which becomes semiencapsulated). Finally, organisms can also *create* affordances by building artifacts and cognitive niches."

There are several points that become clear from this quote, I think. First, in addition to the original Gibsonian framework for affordances, there is room for organisms to participate in perceiving affordances (in the broad sense). Secondly, abductive skills are performed by organisms in perceiving affordances. Thirdly, in such abductively perceiving affordances, perception and action are inseparably intertwined. Finally, organisms can even create affordance by abduction. Except for the first point, I think, all these seem to be due to Magnani.

At the beginning of this section, I introduced Paavola's dilemma: "If abduction relies on instinct, it is not a form of reasoning, and if it is a form of reasoning, it does not rely on instinct". Though I welcomed basically Magnani's way out of this dilemma, that is, they simply co-exist, it was not clear exactly what that solution means. After having improved our understanding of manipulative abduction in animals, now we may understand it better. Magnani claims that, from a semiotic point of view, the idea that there is a conflict between views of abduction in terms of heuristic strategies or in terms of instinct (insight, perception) [9.4, 40, 52], appears old fashioned. And he elaborates his claim that the two aspects simply coexist, by adding that that is so *at the level of the real organic agent* (Emphasis is mine). Depending upon the cognitive/semiotic perspective we adopt, he claims [9.1, pp. 281–282]:

- "1. We can see it as a practical agent that mainly takes advantage of its implicit endowments in terms of guessing right, wired by evolution, where of course instinct or hardwired programs are central.
2. We can see it as the user of explicit and more or less abstract semiotic devices internally stored or externally available – or hybrid – where heuristic plastic strategies (in some organism they are *conscious*) exploiting relevance and plausibility criteria – various and contextual – for guessing hypotheses are exploited."

Now we may underwrite the fact that the two aspects simply co-exist at the level of real organic agents, who are *manipulative abducers*, for it would be hard to find a better example of manipulative abduction than creating affordances. In other words, manipulative abduction is the key factor in Magnani's thought about animal abduction.

## 9.6 Concluding Remarks

Until quite recently abduction has been studied most extensively in scientific context. Even if we consider abduction in artificial intelligence and cognitive science, still the majority of the researches have been governed by theoretical concerns. With Magnani's discovery of manipulative abduction, everything changes. Insofar as manipulative abduction can be construed as a form of practical reasoning, the possibility of expanding the scope of abduction is wide open. As has been seen earlier, Magnani indeed applies manipulative abduction to an impressive array of areas, only a few of which have been examined in this chapter. We con-

firmed that manipulative abduction plays an important role in the use of fallacies in everyday life. Possibly more impressive would be the establishment of manipulative abduction in animals as a research field. Of course, there are many issues unsettled. Above all, there is a desperate need to sharpen our conception of manipulative abduction by deepening our understanding of the characteristics of manipulative abduction. In both expanding the scope and sharpening the conception of manipulative abduction, Magnani's pioneering achievement is simply overwhelming. We may expect rapid and fruitful progress in both directions in the near future.

## References

- 9.1 L. Magnani: *Abductive Cognition. The Epistemological and Eco-Cognitive Dimensions of Hypothetical Reasoning* (Springer, Berlin 2009)
- 9.2 L. Magnani: *Abduction, Reason, and Science: Processes of Discovery and Explanation* (Kluwer, New York 2001)
- 9.3 C.S. Peirce: *The New Elements of Mathematics*, Vol. 4, ed. by C. Eisele (Humanities Press, New York 1976)
- 9.4 M.H.G. Hoffmann: Problems with Peirce's concept of abduction, *Found. Sci.* **4**(3), 271–305 (1999)
- 9.5 C.S. Peirce: *Elements of Logic*, Collected Papers of Charles Sanders Peirce, Vol. 2. (Harvard Univ. Press, Cambridge 1932), ed. by C. Hartshorne, P. Weiss
- 9.6 C.S. Peirce: *Science and Philosophy*, Collected Papers of Charles Sanders Peirce, Vol. 7 (Harvard Univ. Press, Cambridge 1958), ed. by A.W. Burks
- 9.7 S. Marietti: Semiotics and deduction: Perceptual representations of mathematical processes. In: *Semiotics and Philosophy in Charles Sanders Peirce*, ed. by R. Fabbrichesi, S. Marietti (Cambridge Scholars Press, Cambridge 2006) pp. 112–127
- 9.8 W. Park: On classifying abduction, *J. Appl. Logic* **13**(3), 215–238 (2015)
- 9.9 C. Eisele: Mathematical methodology in the thought of C.S. Peirce, *His. Math.* **9**, 333–341 (1982)
- 9.10 R.A. Tursman: *Peirce's Theory of Scientific Discovery. A System of Logic Conceived as Semiotic* (Indiana Univ. Press, Bloomington 1987)
- 9.11 E.J. Crombie: What is deduction? In: *Studies in the Logic of Charles Sanders Peirce*, ed. by N. Houser, D.D. Roberts, J. Van Evra (Indiana Univ. Press, Bloomington and Indianapolis 1997) pp. 460–476
- 9.12 D.G. Campos: The Interpretation and hypothesis-making in mathematics: A Peircean account. In: *New Essays on Peirce's Mathematical Philosophy*, ed. by M.E. Moore (Open Court, Chicago and La Salle 2010) pp. 123–145
- 9.13 F. Stjernfelt: *Natural Propositions: The Actuality of Peirce's Doctrine of Dicisigns* (Docent, Boston 2014)
- 9.14 C.S. Peirce: *The Simplest Mathematics*, Collected Papers of Charles Sanders Peirce, Vol. 4 (Harvard Univ. Press, Cambridge 1933), ed. by C. Hartshorne, P. Weiss
- 9.15 C.S. Peirce: *Exact Logic (Published Papers)*, Collected Papers of Charles Sanders Peirce, Vol. 3 (Harvard Univ. Press, Cambridge 1933), ed. by C. Hartshorne, P. Weiss
- 9.16 S. Marietti: Observing signs. In: *New Essays on Peirce's Mathematical Philosophy*, ed. by M.E. Moore (Open Court, Chicago and La Salle 2010) pp. 147–167
- 9.17 S.-J. Shin: Peirce's two ways of abstraction. In: *New Essays on Peirce's Mathematical Philosophy*, ed. by M.E. Moore (Open Court, Chicago and La Salle 2010) pp. 41–58
- 9.18 S.-J. Shin: The forgotten individual: Diagrammatic reasoning in mathematics, *Synthese* **186**, 149–168 (2012)
- 9.19 L. Magnani: Thinking through drawing: Diagram constructions as epistemic mediators in geometrical discovery, *Knowl. Eng. Rev.* **28**(3), 303–326 (2013)
- 9.20 L. Magnani, R. Dossena: Perceiving the infinite and the infinitesimal world: Unveiling and optical diagrams in mathematics, *Found. Sci.* **10**, 7–23 (2005)
- 9.21 D. Gooding: *Experiment and the Making of Meaning: Human Agency in Scientific Observation and Experiment* (Kluwer, Dordrecht 1990)
- 9.22 J. Woods: Recent developments in abductive logic, *Stud. History Phil. Sci.* **42**(1), 240–244 (2011)
- 9.23 E. Hutchins: *Cognition in the Wild* (MIT Press, Cambridge 1995)
- 9.24 Piaget: *Adaption and Intelligence* (Univ. Chicago Press, Chicago 1974)
- 9.25 L. Magnani: Animal abduction: From mindless organisms to artifactual mediators. In: *Model-Based Reasoning in Science, Technology, and Medicine*, ed. by L. Magnani, P. Li (Springer, Berlin 2007) pp. 3–37
- 9.26 L. Magnani: *Morality in a Technological World: Knowledge as Duty* (Cambridge Univ. Press, Cambridge 2007)
- 9.27 L. Magnani: *Understanding Violence. The Intertwining of Morality, Religion and Violence: A Philosophical Stance* (Springer, Berlin 2011)
- 9.28 D. Gabbay, J. Woods: *The Reach of Abduction: Insight and Trial*, A Practical Logic of Cognitive Systems, Vol. 2 (North-Holland, Amsterdam 2005)

- 9.29 J. Woods: *Errors of Reasoning: Naturalizing the Logic of Inference* (College Publications, London 2013)
- 9.30 W. Park: How to learn abduction from animals? From Avicenna to Magnani. In: *Model-Based Reasoning in Science and Technology Theoretical and Cognitive Issues*, ed. by L. Magnani (Springer, Heidelberg/Berlin 2014) pp. 53–74
- 9.31 P. Thagard: *Computational Philosophy of Science* (MIT Press, Cambridge 1988)
- 9.32 G. Schurz: Patterns of abduction, *Synthese* **164**, 201–234 (2008)
- 9.33 M.H.G. Hoffmann: Theoric transformations and a new classification of abductive inferences, *Trans. C.S. Peirce Soc.* **46**(4), 570–590 (2011)
- 9.34 T. Kapitan: Peirce and the structure of abductive inference. In: *Studies in the Logic of Charles Sanders Peirce*, ed. by N. Houser, D.D. Roberts, J. Van Evra (Indiana Univ. Press, Bloomington Indianapolis 1997) pp. 477–496
- 9.35 L. Magnani: Naturalizing logic: Errors of reasoning vindicated: Logic reapproaches cognitive science, *J. Appl. Logic* **13**, 13–36 (2015)
- 9.36 J. Woods: *Errors of Reasoning: Naturalizing the Logic of Inference* (College Publications, London 2013)
- 9.37 R. Thom: *Esquisse D'une s'Emiophysique*. In: *Semio Physics: A Sketch* (Addison Wesley, Redwood City 1990) Inter Editions, Paris (1988), Translated by V. Meyer
- 9.38 W. Park: Abduction and estimation in animals, *Found. Sci.* **17**, 321–337 (2012)
- 9.39 W. Park: On animal cognition: Before and after the beast–machine controversy. In: *Philosophy and Cognitive Science. Western and Eastern Studies*, ed. by L. Magnani, P. Li (Springer, Heidelberg/Berlin 2012) pp. 53–74, *Sapere* 2
- 9.40 S. Paavola: Peircean abduction: Instinct or inference?, *Semiotica* **153**(1–4), 131–154 (2005)
- 9.41 K.T. Fann: *Peirce's Theory of Abduction* (Martinus Nijhoff, Hague 1970)
- 9.42 D.R. Anderson: *Creativity and the Philosophy of C.S. Peirce* (Martinus Nijhoff Publishers, Dordrecht 1987)
- 9.43 R.J. Roth: Anderson on Peirce's concept of abduction: Further reflections, *Trans. C.S. Peirce Soc.* **24**(1), 131–139 (1988)
- 9.44 B. Brogaard: Peirce on abduction and rational control, *Trans. C.S. Peirce Soc.* **35**(1), 129–155 (1999)
- 9.45 R.B. Burton: The problem of control in abduction, *Trans. C.S. Peirce Soc.* **36**(1), 149–156 (2000)
- 9.46 H.G. Frankfurt: Peirce's account of inquiry, *J. Philos.* **55**, 588–592 (1985)
- 9.47 C.S. Peirce: Carnegie application of 1902 (MS L75), <http://members.door.net/arisbe/> (1902)
- 9.48 N. Rescher: *Peirce's Philosophy of Science* (Univ. Notre Dame Press, Notre Dame 1978)
- 9.49 C.S. Peirce: *Pragmatism and Pragmaticism*, *Collected Papers of Charles Sanders Peirce*, Vol. 5 (Harvard Univ. Press, Cambridge 1934), ed. by C. Hartshorne, P. Weiss
- 9.50 C. Darwin: *The Formation of Vegetable Mould, Through the Action of Worms With Observations on Their Habits* (Univ. Chicago Press, Chicago 1985), originally published in 1881
- 9.51 D.A. Norman: *The Psychology of Everyday Things* (Basic Books, New York 1988)
- 9.52 S. Paavola: Abduction through grammar, critic and methodeutic, *Trans. C.S. Peirce Soc.* **40**(2), 245–270 (2004)

---

# The Logic of Hypothetical Reasoning, Abduction, and Models

## Part C

### Part C The Logic of Hypothetical Reasoning, Abduction, and Models

Ed. by Atocha Aliseda

- 10 The Logic of Abduction: An Introduction**  
Atocha Aliseda, Ciudad de México, Mexico
- 11 Qualitative Inductive Generalization and Confirmation**  
Mathieu Beirlaen, Bochum, Germany
- 12 Modeling Hypothetical Reasoning by Formal Logics**  
Tjerk Gauderis, Gent, Belgium
- 13 Abductive Reasoning in Dynamic Epistemic Logic**  
Angel Nepomuceno-Fernández, Sevilla, Spain  
Fernando Soler-Toscano, Sevilla, Spain  
Fernando R. Velázquez-Quesada, Sevilla, Spain
- 14 Argumentation and Abduction in Dialogical Logic**  
Cristina Barés Gómez, Sevilla, Spain  
Matthieu Fontaine, Ciudad de México, Mexico
- 15 Formal (In)consistency, Abduction and Modalities**  
Juliana Bueno-Soler, Limeira, Brazil  
Walter Carnielli, Campinas, Brazil  
Marcelo E. Coniglio, Campinas, Brazil  
Abilio Rodrigues Filho, Pampulha, Belo Horizonte, Brazil

This section, *The Logic of Hypothetical Reasoning, Abduction, and Models*, shall be concerned with reviewing some formal models for scientific inquiry. Scientific inquiry is a human enterprise to which we cannot deny a big success. It has been our intellectual instrument for achieving great endeavors, such as the arrival on the moon, the possibility of internet communication, the discovery of infections and the invention of vaccines, and many more. The inferential processes involved in scientific inquiry are an essential aspect to analyze when carrying out an enterprise like the one in this *Springer Handbook of Model-Based Science*. However, the field and the material are so vast that it would be impossible to review everything.

Therefore, the main concern within scientific reasoning will be *ampliative reasoning*, those inferential processes in which the conclusion *expands* the given information. This kind of reasoning manifests itself in inferences such as *induction* and *abduction*, and it is opposed to *deduction*, in which conclusions are certain but add nothing new to the given. A salient aspect of ampliative reasoning is the tentative epistemic status of the conclusions produced, something which makes them *defeasible*. That is, given additional information, it may no longer be warranted to draw a previously valid conclusion.

More particularly, the focus in this section will be precisely on the tentative status of the conclusion produced, of its being *hypothetical*. A hypothetical statement is, at the very best, potential knowledge. It is neither true nor false, but holds a hypothetical epistemic status, one that may be settled later as true (when the hypothesis is corroborated) or as false (when it is falsified). Hypothetical reasoning is understood here as a type of reasoning *for explanations*.

One case of hypothetical reasoning is *enumerative induction*, also known as *inductive generalization*, in which the inferential process that is at stake is one which obtains a universal statement (*all ravens are black*) from a set of individual ones (*the first raven is black, ..., the n-th raven is black*). A generalization from instances is a case of ampliative reasoning because it expands what is stated in the instances by advancing a defeasible prediction (*the next raven will be black*). That is, the generalization may fail when a further instance falsifies the conclusion. As is well known, inductive generalization is based on the assumption of the *uniformity of nature*; the world is uniform and, therefore, it seems safe to draw generalizations out of instances, although they may fail at some point.

Another case of hypothetical reasoning that shall be reviewed in depth is that of abduction. Broadly speak-

ing, abduction is a reasoning process *from a single observation to (plausible) explanations*. This characterization, which largely follows the original formulation of Charles Peirce (to be described in [Chap. 10](#)), still leaves ample room for several interpretations. To begin with, when talking about abduction, or any inferential process for that matter, we may refer to a finished product, in this case the *abductive explanation*, or to an activity, the *abductive process*. These two are closely related, for the abductive process produces an abductive explanation, but they are not the same. Moreover, for a given fact to be explained, there are often several abductive explanations to choose from, but only one that counts as the *best* one. Thus, abduction is connected to both hypotheses construction and hypotheses selection. Some authors consider these processes as two separate steps, construction dealing with the generation of plausible explanations according to some criteria of what counts as such, and selection with applying some preference criteria to select the *best* one among the plausible ones. Another issue to be settled in regard to abduction is the distinction from its closest neighbor, induction.

In this very broad map of hypothetical reasoning, two approaches to abduction are salient, one which interprets it as *argument* versus as *inference to the best explanation* (each case, in turn, may be seen either as a product or as a process). This is a familiar distinction in the philosophy of science, where abduction is closely connected with issues of scientific explanation. In a more in-depth view of logical abduction, three characterizations may be identified, namely as *logical inference*, as a *computational process*, and as a *process for epistemic change*. Each one of these views highlights one relevant aspect of abductive reasoning: its logical structure (under an interpretation as a product), its underlying computational mechanism (under an interpretation as a process), and its role in the dynamics of belief revision. Indeed, there are several ways to characterize this reasoning type, and it may be more appropriate to characterize *abductive patterns* rather than trying to define it as a single concept. These approaches and logical characterizations of abduction will be spelled out in detail in the introductory chapter, *The Logic Abduction: An Introduction*, together with an attempt to provide a proper distinction between induction and abduction.

In [Chap. 11](#) by Mathieu Beirlaen, *Qualitative Inductive Generalization and Confirmation*, the author offers a number of *adaptive logics* for inductive generalization, each of which is analyzed as a criterion of confirmation and confronted with Hempel's satisfaction criterion and the hypothetico-deductive model

of confirmation. The adaptive criteria proposed in this paper offer an interesting alternative perspective on (qualitative) confirmation theory in the philosophy of science.

Adaptive logics are a relatively new proof-theoretical framework designed to model ampliative reasoning and *dynamic* information. For the case of inductive generalization, the defeasibility of the conclusion is dealt with in two respects. On the one hand, each of these proposed logics uses a criterion to assert a generalization as a statement within the proof. On the other hand, each of these logics implements a *strategy*, a specific way by which a generalization is refuted and, therefore, *marked* in the proof, so that it is no longer considered as part of the derivation (until it is unmarked due to new information).

In **Chap. 12** by Tjerk Gauderis, *Modeling Hypothetical Reasoning by Formal logics*, the author offers an interesting discussion in regard to the feasibility of the project of modeling hypothetical reasoning by means of formal logics, exploring the assumptions one has to hold in order to accept or reject this endeavor. The author then puts forward four *patterns* of hypothetical reasoning, showing that not a single one can be easily modeled by formal means. *Abduction of a singular fact* is the one pattern that has received most attention in the logical literature; the author gives a review and a detailed description of two adaptive logics devised for this particular pattern, showing that although it is the simplest pattern of all four, there are already some challenges to model it formally.

These two chapters share the logical framework of adaptive logics as a formal model for scientific inquiry, one for inductive generalization, the other one for single fact abduction. Of the four patterns of hypothetical reasoning put forward by Gauderis, the second one, *abduction of a generalization*, is indeed a case of inductive generalization.

According to the previously mentioned classification of hypothetical reasoning, these two chapters fall into the argumentative approach of the type of reasoning modeled (inductive or abductive). However, the chapter by Gauderis highlights a distinction between *practical abduction* and *theoretical abduction*, which – to a certain extent – corresponds to the argumentative versus inference to the best explanation dichotomy of abduction found in the philosophical literature.

The next three chapters of this section belong to the epistemic approach to abduction of taking an agent-oriented stance, one in which the agent's perspective is

at the center of the formal modeling. However, each of these chapters is actually a combination of at least two abductive characterizations.

In **Chap. 13** by Angel Nepomuceno Fernández, Fernando Soler Toscano and Fernando R. Velázquez Quesada, *Abductive Reasoning in Dynamic Epistemic Logic*, the authors rely on the *dynamic epistemic logic* framework, which is largely based on a semantic perspective of modal logic and is an ideal tool to represent an agent's state of knowledge (and belief) together with the dynamics of epistemic change. Operations to upgrade, update and revise *plausibility models* are put forward to dynamically change both the content and the ordering of these models. Original characterizations of what is an abductive problem (solution) are put forward, not with respect to a background theory, as is the case in the classical approach to abduction (see introductory chapter), but rather with respect to an agent's information at a given state. Moreover, plausibility models provide an ordering among epistemic possibilities and accordingly, the *best abductive explanation* turns out to be the most plausible one. This chapter exhibits the inference to the best explanation approach and a combination of the inferential and epistemic characterizations of abduction.

In **Chap. 14** by Cristina Barés Gómez and Matthieu Fontaine, *Argumentation and Abduction in Dialogical Logic*, the authors offer an interesting discussion in favor of a reconciliation between argumentation theory and formal logic; one in which their selected logical framework, *dialogical logic*, is the formal model for scientific inquiry. More particularly, reasoning is modeled via a dialectical interaction in a game-like scenario between the proponent of a thesis and an opponent to it. The authors endorse the view of Dov Gabbay and John Woods, according to which abduction is a response to an *ignorance problem*. An agent has an ignorance problem with respect to a cognitive target when she lacks the knowledge to attain such a target, and abduction is but one type of solution to this kind of problem. The authors of this chapter propose an extension of the dialogical framework to account for abduction and put forward the notion of a *concession problem*, in order to do so. The chapter follows the argumentative approach to abduction, but extends this view with a dialectical interaction and combines aspects of the inferential and epistemic characterizations.

In the last chapter of this section, **Chap. 15** by Juliana Bueno Soler, Walter Carnielli, Marcelo Coniglio, and Abilio Rodrigues, *Formal (In)consistency, Abduction and Modalities*, the authors take a broader view of

scientific inquiry and deal with the problem of inconsistent information, as when there is conflicting evidence for a fact to be explained. In accordance with the focus on hypothetical reasoning taken in this section, the onset of conflicting information, as presented in this chapter, is just another case of a *tentative conclusion*, one which is taken in a sense *weaker* than true, with a provisional status and pending *further investigation*. Authors of this chapter are interested in reviewing the case when no *obvious* explanation is at hand, especially when contradictory information is involved, and a *meaningful* explanation can still be constructed (one that can not be produced in a classical setting). They develop their own formal framework, based on the classical tableaux systems. In respect to abduction, the authors apply a paraconsistent logic to deal with it, and even go further to draw connections between modalities and consistency. Their approach is focused on the process of hypothesis generation, making it attractive for computational implementation (not developed in the paper) and identified by the authors themselves as a case of *creative abduction*, one which contrasts with

*explicative abduction*, according to the distinction put forward by Lorenzo Magnani. The chapter mainly follows the argumentative approach to abduction and is a combination of the computational and epistemic characterizations.

By virtue of their being formal models of scientific reasoning, the chapters to follow are technical; each one of them offers an original contribution to the field, but at the same time, they all provide the intuition and rationale behind the notions presented. The diversity of formal frameworks displayed in this section shows the wide variety of formal tools for hypothetical reasoning modeling. These tools have proved useful to philosophers, logicians, and computer scientists alike and may also be so to anyone who would like to make use of the potential of formal tools to model scientific inquiry at large. This section, *The Logic of Hypothetical Reasoning, Abduction, and Models*, offers a thorough introduction and is an overview to some formal models for hypothetical reasoning found in the philosophy of science and logical literature.



# 10. The Logic of Abduction: An Introduction

Atocha Aliseda

In this chapter, the focus will be on formal models of hypothetical reasoning, in particular on those concerned with *abductive reasoning*.

In Sect. 10.1, the chapter offers a brief history of the notion of abduction, starting with an attempt to distinguish it from its closest neighbor, induction. Charles Peirce's original conception of abduction is then presented and followed by an overview of abduction in the cognitive sciences, together with some paradigmatic examples of the kind that will be dealt with in the chapters to follow. Sect. 10.2 presents two main approaches to abduction in philosophy, as *argument* and as *inference to the best explanation* (IBE), something which sets the ground to put forward a general logical taxonomy for abduction. Sect. 10.3 goes deeper into three logic-based *classical* characterizations of abduction found in the literature, namely as *logical inference*, as a *computational process*, and as a *process for epistemic change*.

Hypothetical reasoning is understood here as a type of reasoning to explanations. This type of

10.1	<b>Some History</b> .....	219
10.1.1	Induction and Abduction .....	219
10.1.2	The Founding Father: C.S. Peirce .....	220
10.1.3	The Cognitive Sciences .....	221
10.1.4	Some Examples .....	221
10.2	<b>Logical Abduction</b> .....	222
10.2.1	Argument .....	222
10.2.2	Inference to the Best Explanation .....	223
10.2.3	A Taxonomy .....	223
10.3	<b>Three Characterizations</b> .....	225
10.3.1	Inferential .....	225
10.3.2	Computational .....	226
10.3.3	Epistemic Change .....	227
10.4	<b>Conclusions</b> .....	228
	<b>References</b> .....	229

reasoning covers abductive as well as inductive inferences. As for the latter, in this handbook part, the concern will be limited to *enumerative induction* and will leave its full presentation to the corresponding chapter (Chap. 11).

## 10.1 Some History

### 10.1.1 Induction and Abduction

In this section, an attempt has been made to provide a clear distinction between inductive and abductive reasoning. As it turns out, diverse terminologies are being used.

On the one hand, for some, beyond deductive logic, there is only place for inductive logic. For others, abduction is another focus, and it is important, at least, to clarify its relationship to induction. For C.S. Peirce, to whom abduction owes its name, *deduction*, *induction* and *abduction* formed a natural triangle – but the literature in general shows many overlaps, and even confusions.

An example of the former view is given by *Russell* [10.1], when he claims:

“There are two sorts of logic: Deductive and inductive. A deductive inference, if it is logically correct, gives as much certainty to the conclusion as the premises, while an inductive inference, even when it obeys all the rules of logic, only makes the conclusion probable even when the premises are deemed certain.”

More recently, there is also a predominant view that identifies induction with ampliative reasoning. For *Paul Thagard*, who coined the field of *computational philos-*

ophy of science, induction is understood in the broad sense of any kind of inference that expands knowledge in the face of uncertainty [10.2].

Since the time of *John Stuart Mill* (1806–1873), the technical name given to all kinds of nondeductive reasoning has been *induction*, but several *methods for discovery and demonstration of causal relationships* [10.3] were recognized. These included generalizing from a sample to a general property, and reasoning from data to a causal hypothesis (the latter further divided into methods of agreement, difference, residues, and concomitant variation). A more refined and modern terminology is *enumerative induction* and *explanatory induction*, of which *inductive generalization*, *inductive projection*, *statistical syllogism*, and *concept formation* are some instances.

Another term for nondeductive reasoning is *statistical reasoning*, introducing a probabilistic flavor, in which explanations are not certain but only probable. Statistical reasoning exhibits the same diversity as abduction. First of all, just as the latter is strongly identified with *backward deduction* (as it will be shown later in this chapter), the former finds its *reverse notion* in probability (For those readers interested in quantitative approaches: the problem in probability is: Given an stochastic model, what can we say about the outcomes? The problem in statistics is the reverse: Given a set of outcomes, what can we say about the model?). Both abduction and statistical reasoning are closely linked with notions like confirmation (the testing of hypotheses) and likelihood (a measure for alternative hypotheses). The former will be reviewed later in this part of the handbook (Chap. 11).

On the other hand, some authors put forward abduction as the main category and take induction as one of its instances. Abduction as *IBE* is considered by *Harman* [10.4] as the basic form of nondeductive inference, which includes (enumerative) induction as a special case (this approach will be presented later in this chapter).

This confusion in terminology returns in artificial intelligence (AI). *Induction* is used for the process of learning from examples – but also for creating a theory to explain the observed facts [10.5], thus making abduction an instance of induction. Abduction is usually restricted to producing abductive explanations in the form of facts (predicates of some sort, as those used in computational implementations of abduction, to be later introduced). When explanations are rules, it is then regarded as part of induction. Indeed, the relationship between abduction and induction has been a distinguished topic of several workshops in AI mainstream conferences (European Conference on Artificial Intel-

ligence (ECAI) and International Joint Conference on Artificial intelligence (IJCAI)) as well as that of edited books [10.6].

With the sole purpose of providing a methodological distinction between abduction and induction, in this chapter, abduction will be understood as reasoning from a single observation to its explanations, and induction as *enumerative induction*, a reasoning kind from samples to general statements. Given these tentative characterizations, those aspects that distinguish them will be highlighted. While induction explains a set of observations, abduction explains a single one. Induction makes a prediction for further observations, abduction does not (directly) account for later observations. While induction needs no background theory per se, abduction relies on a background theory to construct and test its abductive explanations.

As for their similarities, induction and abduction are both ampliative and defeasible modes of reasoning. More precisely, they are *nonmonotonic* types of inference. A consequence  $\Rightarrow$  is labeled as nonmonotonic whenever  $T \Rightarrow b$  does not guarantee  $T, a \Rightarrow b$ . That is, the addition of a new premise (*a*) may invalidate a previous valid argument. In the terminology of philosophers of science, nonmonotonic inferences are not *erosion proof* [10.7]. Moreover, qua direction, both run in the opposite direction to standard deduction; they both run from evidence to explanation and the status of the produced explanation is hypothetical.

To clear up terminological conflicts, one might want to coin new terminology altogether. Some may argue for a new term of *explanatory reasoning* as done in [10.8] or even better as *hypothetical reasoning* trying to describe its fundamental aspects without having to decide if they are instances of either abduction or induction. In this broader perspective, it is also possible to capture explanation for more than one instance or for generalizations and introduce further fine-structure. Indeed, a classification in terms of *patterns of hypothetical reasoning* may be very appropriate, as will be found later on in this part of the handbook (Chap. 12. A key reference in the literature in terms of *patterns of abduction* is found in [10.9]).

### 10.1.2 The Founding Father: C.S. Peirce

The literature on abduction is so vast that makes impossible to undertake a complete survey here. But any history of abduction cannot fail to mention the founding father: Charles Sanders Peirce (1839–1914).

Peirce is the founder of American pragmatism and the first philosopher to give to abduction a logical form. However, his notion of abduction is a difficult one to un-

ravel. On the one hand, it is entangled with many other aspects of his philosophy, and on the other hand, several different conceptions of abduction evolved in his thought. The notions of logical inference and of validity that Peirce puts forward go beyond our present understanding of what logic is about. They are linked to his epistemology, a dynamic view of thought as logical inquiry, and correspond to a deep philosophical concern, that of studying the nature of synthetic reasoning. In what follows, a few general aspects of his later theory of abduction will be pointed out, to later concentrate on some of its more logical aspects (for a more elaborate analysis of the evolution of Peirce's abduction as well as its connection to his epistemology [10.10–12]).

For Peirce, three aspects determine whether a hypothesis is promising: it must be *explanatory*, *testable*, and *economic*. A hypothesis is an explanation if it accounts for the facts. Its status is that of a suggestion until it is verified, which explains the need for the second criterion.

Finally, the motivation for the economic criterion is twofold: A response to the practical problem of having innumerable explanatory hypotheses to test, as well as the need for a criterion to select the best explanation among the testable ones.

For the explanatory aspect, *Peirce* gave the following often-quoted logical formulation [10.13, CP 5.189]:

“The surprising fact, C, is observed.  
But if A were true, C would be a matter of course.  
Hence, there is reason to suspect that A is true.”

This formulation has played a fundamental role in Peirce scholarship, and it has been the point of departure of many classic studies on abductive reasoning in all fields that make up the cognitive sciences, mainly those in which the approach is argumentative-based. Nevertheless, these accounts have paid little attention to the elements of this formulation and practically none to what Peirce said elsewhere in his writings. This situation may be due to the fact that his philosophy is very complex and not easy to be implemented in the computational realm. The notions of logical inference and of validity that Peirce puts forward go beyond logical formulations but at the same time some of his ideas find a natural place in recent proposals, such as that found in *theories of belief revision* (to be reviewed later in this chapter).

The approach to abductive reasoning, in this handbook part, reflects this Peircean diversity in part, taking abduction as a style of logical reasoning that occurs at different levels and contexts and comes in several degrees.

### 10.1.3 The Cognitive Sciences

Research on abduction in AI dates back to the 1970s of the twentieth century [10.14], but it is only fairly recently that it has attracted great interest, in areas like logic programming, knowledge assimilation, and diagnosis, to name a few. Some publications, collective and individual alike, are found in [10.15–19], to name a few. In all these places, the discussion about the different aspects of abduction has been conceptually challenging but also shows a (terminological) confusion with its close neighbor, induction (similar to what has already been pointed out previously). Abduction has also been a distinguished topic of model-based reasoning (MBR) conferences (those linked to the editorial project of this handbook).

The importance of abduction has been recognized by leading researchers in nearly all fields that make up the cognitive sciences: philosophy, computer science, cognitive psychology, and linguistics. For *Jaakko Hintikka*, abduction is *the fundamental problem of contemporary epistemology*, in which *abductive inferences must be construed as answers to the inquirer's explicit or (usually) tacit question put to some definite source of answers (information)* [10.11, p.519]. For *Herbert Simon*, the nature of the *retroductive process* (Peirce's original term for abduction) is *the main subject of the theory of problem solving in both its positive and normative versions* [10.20, p.151]. For *Paul Thagard*, several kinds of abduction play a key role as heuristic strategies in the program PI (processes of induction), a working system devoted to explain – in computational terms – some of the main problems in philosophy of science, such as scientific discovery, explanation, and evaluation [10.2]. Finally, for *Noam Chomsky*, abduction plays a key role in language acquisition; for the child *abduces* the rules of grammar guided by her innate knowledge of language universals [10.21].

### 10.1.4 Some Examples

There are a variety of approaches that claim to capture the true nature of the notion of abduction. One reason for this diversity lies in the fact that abductive reasoning occurs in a multitude of contexts and aims to cover from the simplest selection of already existing hypotheses in a context of common sense reasoning to the generation of new concepts in science. Here are some examples illustrating this variety (examples are taken from [10.8]):

1. *Common sense: Explaining observations with simple facts.* All you know is that the lawn gets wet either when it rains, or when the sprinklers are on. You wake up in the morning and notice that the lawn

is wet. Therefore you hypothesize that it rained during the night or that the sprinklers had been on.

2. *Common sense: When something does not work.* You come into your house late at night, and notice that the light in your room, which is always left on, is off. It has been raining very heavily, and so you think some power line went down, but the lights in the rest of the house work fine. Then, you wonder if you left both heaters on, something which usually causes the breakers to cut off, so you check them: but they are OK. Finally, a simpler explanation crosses your mind. Maybe the light bulb of your lamp which you last saw working well, is worn out, and needs replacing.

These examples belong to a practical setting found in our day-to-day common sense reasoning. The first one is the paradigmatic example of abduction in AI and will be analyzed in full detail later in this part of the handbook, in Chap. 15. An extension of the second example will be presented in Chap. 13.

Some other instances of abductive reasoning, but in this case oriented to model the cognitive competence of health practitioners and of working scientists are the following:

3. *Statistical reasoning: Medical diagnosis.* Jane Jones recovered quite rapidly from a streptococci infection after she was given a dose of penicillin. Almost all streptococcus infections clear up quickly upon administration of penicillin, unless they are penicillin resistant, in which case the probability of quick recovery is rather small. The doctor knew that Jane's infection is of the penicillin-resistant type, and is puzzled by her recovery. Jane Jones then confesses that her grandmother had given her Bel-

ladonna, a homeopathic medicine that stimulates the immune system by strengthening the physiological resources of the patient to fight infectious diseases (This is an adaptation of Hempel's illustration of his inductive-statistical model of explanation as shown in [10.7]). The part about homeopathy is taken from [10.8])

4. *Scientific reasoning: Kepler's discovery.* It has been claimed that Johannes Kepler's great discovery that the orbit of the planets is elliptical rather than circular was a prime piece of abductive reasoning [10.13, CP 2.623]. What initially led to this discovery was his observation that the longitudes of Mars did not fit circular orbits, but before even dreaming that the best explanation involved ellipses instead of circles, he tried several other forms. Moreover, Kepler had to make some other assumptions about the planetary system, without which his discovery does not work. His heliocentric view allowed him to think that the sun, so near to the center of the planetary system, and so large, must somehow cause the planets to move as they do. In addition to this strong conjecture, he also had to generalize his findings for Mars to all planets, by assuming that the same physical conditions obtained throughout the solar system.

The third example is concerned with the construction of a diagnosis, in which based on a series of observations (symptoms and signs) and of causal relations linking those observations to pathologies, health professionals build their diagnoses in order to determine an illness. But abduction also occurs in theoretical scientific contexts, such as the one described in the fourth example, in which anomalous observations give rise to new ideas that force to revise knowledge found in existing theories.

## 10.2 Logical Abduction

In contemporary philosophy of science, logical approaches to abduction may be traced back to *Hempel's* models [10.8, 22–24] and are therefore related to a reasoning style connected to theories of explanation and empirical progress [10.25]. More recently, logical abduction found a place in computationally oriented theories of belief change in AI as well as in the field of *nonmonotonic logics* [10.18, 26–30].

In what follows, two approaches to abduction, as an argument and as inference to the best explanation, will be described. Later on, a general taxonomy for its logical analysis will be proposed.

### 10.2.1 Argument

In AI circles, Peirce's formulation has been generally interpreted as the following logical argument-schema

$$\begin{array}{c} C \\ \hline A \rightarrow C \\ A \end{array}$$

where the status of *A* is tentative (it does not follow as a logical consequence from the premises).

However intuitive, this interpretation certainly captures neither the fact that  $C$  is surprising nor the additional criteria Peirce proposed. Moreover, the interpretation of the second premise should not be committed to material implication (For a causal interpretation of this conditional [10.31]). But other interpretations are possible; any nonstandard form of logical entailment or even a computational process in which  $A$  is the input and  $C$  the output, are all feasible interpretations for *if  $C$  were true,  $A$  would be a matter of course*.

The additional Peircean requirements of testability and economy are not recognized as such in AI, but to some extent are nevertheless incorporated. The latter criterion is implemented as a further selection process to produce the *best explanation*, since there might be several formulae that satisfy the above formulation but are nevertheless inappropriate as explanations. Testability as understood by Peirce is an extra-logical empirical criterion.

### 10.2.2 Inference to the Best Explanation

Abduction as IBE was proposed by Gilbert Harman as the basic form of nondeductive inference, one which included enumerative induction as one of its instances. According to him [10.4, p.89]:

“Uses of the inference to the best explanation are manifold. When a detective puts the evidence together and decides that it must have been the butler, he is reasoning that no other explanation which accounts for all the facts is plausible enough or simple enough to be accepted.”

This idea may be put into an argumentative form as follows [10.32]:

“ $D$  is a collection of data  
(facts, observations, givens)  
 $H$  explains  $D$   
(would, if true, explain  $D$ )  
No other hypothesis explains  $D$  as well as  $H$  does  
Therefore,  $H$  is probably correct”

Given a fact to be explained, there are often several possible abductive explanations, but (hopefully) only one that counts as the best one. Pending subsequent testing, in the previous common sense example of light failure (2) several abductive explanations account for the unexpected darkness of the room (power line down, breakers cut off, bulb worn out). But only one may be considered as best explaining the event, namely the *true* one, the one that really happened. But other preference criteria may be appropriate, too, especially when there is no direct test available.

Under this approach, abduction may be regarded as a single process by which a single best explanation is constructed. And the focus is on finding selection criteria which allow to characterize a hypothesis as the best one (A key references for the interpretation of abduction as inference to the best explanation is found in [10.33]).

Thus, abduction is connected to both hypotheses construction and hypotheses selection. Some authors consider these processes as two separate steps: construction dealing with what counts as a possible abductive explanation, and selection with applying some preference criterion over possible abductive explanations to select the best one.

As it turns out, the notion of a *best abductive explanation* necessarily involves contextual aspects, varying from application to application. There is at least a new parameter of preference ranking here. There exists both a philosophical tradition on the logic of preference, and logical systems in AI for handling preferences that may be used to single out best explanations [10.34, 35]. A proposal to tackle this approach in the framework of *dynamic epistemic logic*, in which (an extension of) the light failure example is analyzed in full detail, will be presented later in this handbook part (Chap. 13).

### 10.2.3 A Taxonomy

What has been presented so far may be summarized as follows. Abduction is a process whose products are specific abductive explanations, with a certain inferential structure, making an (abductive) explanatory argument. As for the *logical form* of abduction – referring to the inference corresponding to the abductive process that takes a background theory ( $\Theta$ ) and a given observation ( $\varphi$ ) as inputs, and produces an abductive explanation ( $\alpha$ ) as its output – the proposal here is that at a very general level, the logical structure of abduction may be viewed as a threefold relation

$$\Theta, \varphi \Rightarrow \alpha .$$

Other parameters are possible here, such as a preference ranking, but these would rather concern the further selection process. This characterization aims to capture the direction (from evidence to abductive explanation) of this type of reasoning. In the end, however, the goal is to characterize an (abductive) explanatory argument, in its deductive forward fashion, that is, an inference from theory ( $\Theta$ ) and abductive explanation ( $\alpha$ ) to evidence ( $\varphi$ ) as follows

$$\Theta, \alpha \Rightarrow \varphi .$$

Against this background, three main parameters that determine types of explanatory arguments are put forward:

1. An *inferential parameter* ( $\Rightarrow$ ) sets some suitable logical relationship among explananda, background theory, and explanandum.
2. Next, *abductive triggers* determine what kind of abductive process is to be performed:  $\varphi$  may be a novel phenomenon, or it may be in conflict with the theory  $\Theta$ .
3. Finally, *abductive outcomes* ( $\alpha$ ) are the various products (abductive explanations) of an abductive process: facts, rules, or even new theories.

### Abductive Parameters

**Varieties of Inference.** In the above schema, the notion of explanatory inference  $\Rightarrow$  is not fixed. It can be classical derivability  $\vdash$  or semantic entailment  $\models$ , but it does not have to be. Instead, it is regarded as a parameter that can be set independently. It ranges over such diverse values as probable inference ( $\Theta, \alpha \Rightarrow_{\text{probable}} \varphi$ ), in which the explanans render the explanandum only highly probable, or as the inferential mechanism of logic programming ( $\Theta, \alpha \Rightarrow_{\text{LP}} \varphi$ ). Further interpretations include dynamic inference ( $\Theta, \alpha \Rightarrow_{\text{dynamic}} \varphi$ ), replacing truth by information change potential along the lines of belief update or belief revision (Later in the chapter, both the abductive mechanism in logical programming as well as the belief revision framework will be described further. The dynamic interpretation will be reviewed in Chap. 13). To be sure, the point here is that abduction is not one specific nonstandard logical inference mechanism, but rather a way of using any one of these.

**Two Triggers.** According to Peirce, as his logical formulation dictates, abductive reasoning is triggered by a *surprising phenomenon*. The notion of surprise, however, is a relative one, for a fact  $\varphi$  is surprising only with respect to some background theory  $\Theta$  providing expectations. What is surprising to someone (that the lights go on as I enter the copier room) might not be surprising to someone else. One way to interpret a surprising fact is as one in need of an explanation. From a logical point of view, this assumes that the fact is not already explained by the background theory  $\Theta$ :  $\Theta \not\Rightarrow \varphi$ .

Moreover, one may also consider the status of the negation of  $\varphi$ . Does the theory explain the negation of

observation instead ( $\Theta \Rightarrow \neg\varphi$ )? Thus, two triggers for abduction are identified: *novelty* and *anomaly*

**Definition 10.1 (Abductive Novelty:  $\Theta \not\Rightarrow \varphi$ ,  $\Theta \not\Rightarrow \neg\varphi$ )**

$\varphi$  is novel. It cannot be explained ( $\Theta \not\Rightarrow \varphi$ ), but it is consistent with the theory ( $\Theta \not\Rightarrow \neg\varphi$ )

**Definition 10.2 (Abductive Anomaly:  $\Theta \not\Rightarrow \varphi$ ,  $\Theta \Rightarrow \neg\varphi$ )**

$\varphi$  is anomalous. The theory explains rather its negation ( $\Theta \Rightarrow \neg\varphi$ ).

As it will be shown later on, in the computational literature on abduction, novelty is the condition for an *abductive problem*. Following Peirce, incorporating anomaly as a second basic type is put forward (See [10.8] for the proposal).

Of course, nonsurprising facts (where  $\Theta \Rightarrow \varphi$ ) should not be candidates for abductive explanations. Even so, one might speculate if facts which are merely probable on the basis of  $\Theta$  might still need an abductive explanation of some sort to further cement their status.

**Different Outcomes.** Abductive explanations themselves come in various forms: facts, rules, or even theories. Sometimes one simple fact suffices to explain a surprising phenomenon. In other cases, a rule establishing a causal connection might serve as an abductive explanation. And many cases of abduction in science provide new theories to explain surprising facts. These different options may sometimes exist for the same observation, depending on how complex one wants to take it. Later in this handbook part, patterns of hypothetical reasoning will be classified according to abductive outcomes types (Chap. 12).

Moreover, as well known in the history of science, genuine abductive explanations sometimes introduce new concepts, over and above the given vocabulary. For instance, the eventual explanation of planetary motion (example section in this chapter), was not given by Kepler, but by Newton, who introduced a new notion of *force* – and then derived elliptic motion via the law of gravity. Abduction via new concepts – broadly conceived – will be outside the scope of our analysis in this part of the handbook (however, see Chap. 12 for a special case of *conceptual abduction*).

## 10.3 Three Characterizations

This section offers a description of three logical characterizations of abduction found in the literature. Nowadays, there are plenty of logic-based papers on abduction, most of which fit into one of the following three characterizations:

1. Abduction as logical inference
2. Abduction as computation
3. Abduction as epistemic change.

While the first one aims at characterizing abduction as *backward deduction plus additional conditions* and (generally) has classical logical consequence as an underlying inference, the second one focuses on providing specific algorithms that produce abductive explanations, and it is therefore as varied as computational platforms allow for. The last one puts forward abductive operations to revise, expand and update and has close links both with theories of belief revision and of update in AI. These approaches may be labeled as the *classical characterizations of logical abduction*. These characterizations are classical in at least two respects. In the first place, they are classical because they emerged originally as logical models for abduction and capture at least one relevant aspect of abductive reasoning: Its logical structure, its underlying computational mechanism, and its role in the dynamics of belief revision. In another respect, they are classical – in so far as presented in this chapter – because they exhibit a very classical way for doing logic, computation, or formalizations of belief revision theories.

Interesting proposals in all three characterizations and of its combinations are to be found in special issues of the *Logic Journal of the IGPL* (most notably, in special issues on abduction, such as [10.36–38]). Some of the chapters to follow are good examples of these combinations as well.

### 10.3.1 Inferential

The *classical* characterization of abduction as logical inference is mainly a deductive classical logical account in which a background theory ( $\Theta$ ), together with an abductive explanation ( $\psi$ ), constitutes the explananda and do entail the explanandum ( $\varphi$ ). It puts forward the following logical schema.

Given a theory  $\Theta$  (a set of formulae) and a formula  $\varphi$  (an atomic formula),  $\psi$  is an abductive explanation if:

1.  $\Theta \cup \psi \models \varphi$
2.  $\psi$  is consistent with  $\Theta$
3.  $\psi \not\models \varphi$
4.  $\psi$  is *minimal*.

When  $\models$  is interpreted as classical logical consequence, conditions 1 and 2 go hand in hand and are clearly mandatory. The first one dictates the entailment condition while the second one imposes the abductive explanation to be consistent with the background theory, for the *principle of explosion* is valid in classical logic (Chap. 15). As for condition 3, it is necessary in order to avoid self-explanations or, more generally, explanations that are independent from the background theory. Condition 4 aims at capturing either a criterion of best explanation by which *minimal* may be interpreted as selecting the *weakest explanation* (e.g., not equal to  $\Theta \rightarrow \varphi$ ) or a *preferred explanation* (which requires a predefined preference ordering). Additionally, there is usually another requirement restricting the logical vocabulary as well as the syntactic form of the explanation  $\psi$ , such as being an atomic formula from the vocabulary in the logical language. This schema suggests a classification into *abductive inferential styles* (as done in [10.8]), one in which *plain, consistent, explanatory, minimal, and preferential* abductive styles correspond to above conditions.

A (pre) condition to the above schema – not always made explicit – is that the theory does not already entail neither the explanandum nor its negation ( $\Theta \not\models \varphi$  and  $\Theta \not\models \neg\varphi$ ). In many proposals, this condition constitutes an *abductive problem* (such as in Chap. 15). Given the previous taxonomy for abduction by which two abductive triggers were distinguished, the following definition of an *abductive problem* is put forward, one in which *novel* and *anomalous* problems are distinguished.

#### Definition 10.3 (Abductive problem)

Let  $\Theta$  and  $\varphi$  be a theory and a formula, respectively, in some language  $\mathcal{L}$ . Let  $\models$  be a consequence relation on  $\mathcal{L}$ :

- The pair  $(\Theta, \varphi)$  constitutes a (novel) abductive problem when neither  $\varphi$  nor  $\neg\varphi$  are consequences of  $\Theta$ . That is, when

$$\Theta \not\models \varphi \quad \text{and} \quad \Theta \not\models \neg\varphi .$$

- The pair  $(\Theta, \varphi)$  constitutes an *anomalous abductive problem* when  $\varphi$  is not a consequence of  $\Theta$ , but  $\neg\varphi$  is. That is, when

$$\Theta \not\models \varphi \quad \text{and} \quad \Theta \models \neg\varphi .$$

It is typically assumed that the theory  $\Theta$  is a set of formulas closed under logical consequence, and that  $\models$  is a truth-preserving consequence relation.

Given this definition, a (*novel*) *abductive solution* would then be any formula  $\psi$  which follows the logical schema above. It remains to be seen however, what would be an *anomalous abductive solution* for an *anomalous abductive problem*. To this end, several proposals exist in the literature (e.g., [10.39]), many of which acknowledge the possibility of theory revision, for some formula may be retracted from  $\Theta$  in order to maintain the consistency of the revised theory (in Chap. 13 a proposal is made in this direction in the framework of dynamic epistemic theories, one in which a three-way distinction of abductive problems is offered).

Going back to the logical schema above, it should be stressed that some authors are not committed to classical logical entailment but rely instead on some other form of nonclassical consequence, something that often has as a consequence that the conditions that make up the logical schema need not be imposed explicitly or are replaced instead by weaker ones. For example, in Chap. 15 in which abduction is modeled in *paraconsistent logics*, the consistency condition is replaced by a weaker one stating nontriviality ( $\Theta$  and  $\psi$  are nontrivial when there exists a  $B$  such that  $\Theta, \psi \not\vdash B$ ).

To be sure, while there are clear-cut characterizations of what is an *abductive problem (solution)*, there are several logical ways to make it precise. It should be clear by now that there is a significant way in which abduction is interpreted as a logical inference in its own right. This characterization does integrate the view of an otherwise varied group of scholars; *John Woods* has labeled it as the *AKM model*, according to the initial letters of its proponents surnames [10.40, p.305]:

“Thus, for example, for ‘A’ we have *Aliseda* [10.8]; for ‘K’ we have *Kowalski* [10.41], *Kuipers* [10.25], *Kakas* et al. [10.26] and *Flach* and *Kakas* [10.6]; and for ‘M’ there is *Magnani* [10.19] and *Meheus* et al. [10.30]. Needless to say, there are legions of AKM-proponents whose surnames are ungraced by any of these letters in first positions.”

### 10.3.2 Computational

Of the many computational approaches to abduction, *abductive logic programming* (ALP) is the selected one to be described in detail (see [10.26, 42–44] for an overview to the field). Logic programming works mostly within first-order logic, and it consists of *logic programs, queries*, and an underlying inferential mechanism known as *resolution*. Abduction emerges naturally in logic programming as a repair mechanism, that is, its result is an extension of a logic program with the facts needed for a query to succeed (a query does not

succeed when it is not derivable from the program). In actual ALP, for these facts to be counted as abductions, they have to belong to a predefined set of abducibles, and to be verified by additional conditions (so-called *integrity constraints*), in order to prevent a combinatorial explosion of possible explanations.

Therefore, logic programming does not use blind deduction; different control mechanisms for proof search determine how queries are processed and this is crucial to the efficiency of the enterprise. Hence, different control policies will vary in the abductions produced, their form and the order in which they appear.

In what follows, it will be illustrated how an ALP system works. A notation closer to logic than to logic programs is the one used in what follows. An *abductive logic program* has three components:

- *A set of rules.* Each rule has a head (a predicate) and a body (a set of predicates). A way to prove the head is to prove all predicates of its body. Rules without body are facts and are assumed to be true. Here is a very simple logic program

$$\text{eats}(X, Y) \leftarrow \text{vegetarian}(X), \text{vegetable}(Y) \quad (10.1)$$

$$\text{eats}(X, Y) \leftarrow \text{carnivore}(X), \text{meat}(Y) \quad (10.2)$$

$$\text{vegetarian}(\text{rabbit}). \quad (10.3)$$

There are two rules (10.1) and (10.2) for the predicate  $\text{eats}(X, Y)$ . The first rule, for example, states that a way to prove that  $X$  eats  $Y$  is to prove that  $X$  is vegetarian and  $Y$  is a vegetable. The fact (10.3) states that the rabbit is vegetarian:

- *A set of abducible predicates.* A common restriction is that abducible predicates cannot be in the head of any rule. In the example above, the following may be considered as abducible predicates:  $\text{vegetarian}(X)$ ,  $\text{carnivore}(X)$ ,  $\text{vegetable}(X)$  and  $\text{meat}(X)$ .
- *A set of integrity constraints.* These are rules with  $\perp$  (falsehood) as head. To be satisfied, they require that not all literals in the body are simultaneously true. In the example above, the following constraint may be added in order to avoid that an animal is considered both a vegetarian and a carnivore

$$\perp \leftarrow \text{vegetarian}(X), \text{carnivore}(X). \quad (10.4)$$

An *abductive problem* in ALP manifests itself when there is a query (an instance of some predicate) that cannot succeed with the rules of the logic program



(the query is not derivable from the program). In the example above, the query `eats(rabbit,banana)` is not successful with the program (10.1)–(10.3). An *abductive solution* in ALP is then a set of facts that, together with the program, entail the original query. In above example, there are two possible sets of these:

- {vegetable(banana)}
- {carnivore(rabbit), meat(banana)}.

But only the first one satisfies integrity constraint (10.4), because assuming `carnivore(rabbit)`, together with `vegetarian(rabbit)`, contradicts (10.4).

More technically, in ALP, *standard selective linear definite resolution* [10.45] is extended to build the abductive solutions and check the integrity constraints. Some modified resolution procedures and ALP systems have been developed since the pioneer work of *Bob Kowalski* [10.41], but the usual procedure remains the same: look for abductive solutions at the *dead ends* of the proofs. That is, when a resolution proof cannot be completed because there is no clause to prove some query, if the predicate of that query is abducible, then it is incorporated to the abductive solution. Other procedures use *well-founded semantics* and *stable models*. The overview of ALP is left here.

As already mentioned, there are indeed many computational approaches to abduction. To end this section, here are some final words in regard to the logical framework of *semantic tableaux*, which – when properly extended – allows for a combination of both approaches to abduction previously reviewed, that of abduction as a logical inference and abduction as computation.

As well known in the logical literature, semantic tableaux is a refutation method to test formulae validity. Roughly speaking, it works as follows (see [10.46] for an overview to the field):

“To test if a formula  $\varphi$  follows from a set of premises  $\Theta$ , a tableau tree for the sentences in  $\Theta \cup \{\neg\varphi\}$  is constructed, denoted by  $\mathcal{T}(\Theta \cup \{\neg\varphi\})$ . The tableau itself is a binary tree built from its initial set of sentences by using rules for each of the logical connectives that specify the ways in which the tree branches.

If the tableau closes (every branch contains an atomic formula  $\psi$  and its negation), the initial set is unsatisfiable and the entailment  $\Theta \models \varphi$  holds. Otherwise, if the resulting tableau has open branches, the formula  $\varphi$  is not a valid consequence of  $\Theta$ .”

Within this framework, abduction comes into play as an *extension* of the constructed tableau. When there are open branches, which indicates the condition for a *novel abductive problem* in this framework, the generation of abductive solutions consists in producing

those formulae, which close the open branches (in a consistent way). Abduction in semantic tableaux offers a way of implementing computationally the AKM model (for more details on abduction in semantic tableaux, see [10.27] for its original proposal and [10.8] for a further development). Moreover, abduction in semantic tableaux for a paraconsistent logical theory is to be found in Chap. 15, in which a detailed description of this framework will be offered.

### 10.3.3 Epistemic Change

Abduction has also been characterized as a process for epistemic change, and in this respect an obvious related territory is theories of belief revision in AI (see [10.8, 29, 47, 48] for an introduction and a more detailed account on this topic). These theories describe how to incorporate a new piece of information into a database, a scientific theory, or a set of common sense beliefs. More precisely, given a consistent theory  $\Theta$  closed under logical consequence, called the belief state, and a sentence  $\varphi$ , the incoming belief, there are three *epistemic attitudes* for  $\Theta$  with respect to  $\varphi$ :

1.  $\varphi$  is accepted ( $\varphi \in \Theta$ )
2.  $\varphi$  is rejected ( $\neg\varphi \in \Theta$ )
3.  $\varphi$  is undetermined ( $\varphi \notin \Theta, \neg\varphi \notin \Theta$ ).

Given these attitudes, the following operations characterize the kind of belief change  $\varphi$  brings into  $\Theta$ , thereby effecting an epistemic change in the agent’s currently held beliefs:

- *Expansion*. A new sentence is added to  $\Theta$  regardless of the consequences of the larger set to be formed. The belief system that results from expanding  $\Theta$  by a sentence  $\varphi$  together with the logical consequences is denoted by  $\Theta + \varphi$ .
- *Contraction*. Some sentence in  $\Theta$  is deleted without any addition of new facts. In order to guarantee the deductive closure of the resulting system, some other sentences of  $\Theta$  may be given up. The result of contracting  $\Theta$  with respect to sentence  $\varphi$  is denoted by  $\Theta - \varphi$ .
- *Revision*. A new sentence that is (typically) inconsistent with a belief system  $\Theta$  is added, but in order that the resulting belief system be consistent, some of the old sentences in  $\Theta$  are deleted. The result of revising  $\Theta$  by a sentence  $\varphi$  is denoted by  $\Theta * \varphi$ .

Of these operations, revision is the most complex one. Indeed the three belief change operations can be reduced into two of them, since revision and contraction may be defined in terms of each other. In particular, revision here is defined as a composition of contraction and expansion: First contract those beliefs of  $\Theta$  that are

in conflict with  $\varphi$ , and then expand the modified theory with sentence  $\varphi$  (known as *Levi's identity*). While expansion can be uniquely and easily defined ( $\Theta + \varphi = \{\alpha \mid \Theta \vee \{\varphi\} \vdash \alpha\}$ ), this is not so with contraction or revision, as several formulas can be retracted to achieve the desired effect. Therefore, additional criteria must be incorporated in order to fix which formula to retract. Here, the general intuition is that changes on the theory should be kept *minimal*, in some sense of informational economy (One way of dealing with this issue is based on the notion of *entrenchment*, a preferential ordering which lines up the formulas in a belief state according to their importance [10.49]. Thus, those formulas that are the *least entrenched*, should be retracted first).

Moreover, *epistemic theories* in this tradition observe certain *integrity constraints* (such as those previously shown for ALP), which concern the theory's preservation of consistency, its deductive closure and two criteria for the retraction of beliefs: The loss of information should be kept minimal and the less entrenched beliefs should be removed first. These are the very basics of the AGM approach (This acronym stands for the three initial letters of its proponents: *Alchourrón, Gärdenfors, and Makinson*, authors of the seminal paper which started this tradition [10.50]).

Abduction may be seen as an epistemic process for belief revision. In this context, an incoming sentence  $\varphi$  is not necessarily an observation, but rather a belief for which an explanation is sought. The previously defined *abductive novelty* and *abductive anomaly* correspond respectively, to the epistemic attitudes of undetermination and rejection (provided that  $\Rightarrow$  is  $\vdash$  and  $\Theta$  closed under logical consequence). Both a novel phenomenon and an anomalous one induce a change in the original theory. The latter calls for a revision and the former for expansion. So, the basic operations for abduction are expansion and revision. Therefore, two epistemic attitudes and changes in them are reflected in an abductive model.

## 10.4 Conclusions

This chapter covered, on the one hand, a brief historical overview of the logic of abduction since its origins in Charles Peirce's view to its privileged place in the cognitive sciences. In the context of philosophy of science, logical abduction is relevant in regard to issues of scientific explanation. In a broader context, one including computer science, abduction is immersed in the heuristics of inferential mechanisms. On the other hand, this chapter offered an in-depth overview of logical abduction. A distinction between approaches in which abduction takes the

Here, then, are two abductive operations for epistemic change (as proposed in [10.8]):

- *Abductive expansion.* Given an abductive novelty  $\varphi$ , a consistent explanation  $\alpha$  for  $\varphi$  is computed in such a way that  $\Theta, \alpha \Rightarrow \varphi$ , and then added to  $\Theta$ .
- *Abductive revision.* Given an abductive anomaly  $\varphi$ , a consistent explanation  $\alpha$  is computed as follows: The theory  $\Theta$  is revised into  $\Theta'$  so that it does not explain  $\neg\varphi$ . That is,  $\Theta' \not\Rightarrow \neg\varphi$ , where  $\Theta' = \Theta - (\beta_1, \dots, \beta_n)$  (In many cases, several formulas and not just one must be removed from the theory. The reason is that sets of formulas which entail (explain)  $\varphi$  should be removed. Example: Given  $\Theta = \{\alpha \rightarrow \beta, \alpha, \beta\}$  and  $\varphi = \neg\beta$ , in order to make  $\Theta, \neg\beta$  consistent, one needs to remove either  $\{\beta, \alpha\}$  or  $\{\beta, \alpha \rightarrow \beta\}$ ). Once  $\Theta'$  is obtained, a consistent explanation  $\alpha$  is calculated in such a way that  $\Theta', \alpha \Rightarrow \varphi$  and then added to  $\Theta$ . Thus, the process of revision involves both contraction and expansion.

Some of the previous examples of abduction given in Sect. 10.1.4 may be described as expansions (1–3), where the background theory gets expanded to account for a new fact. Another one of them (4) is clearly a case calling for theory revision, that in which the theory needs to be *revised* in order to account for an anomaly, such as those found in practical settings like diagnostic reasoning [10.15, 39]. Belief revision theories provide an explicit calculus of modification for both cases and applied to abduction, operations for abductive expansion and abductive revision are defined as well.

Note, however, that in this approach changes occur only in the theory, as the situation or world to be modeled is supposed to be static, only new information is coming in. Another important type of epistemic change studied in AI is that of *update*, the process of keeping beliefs up-to-date as the world changes. A recent proposal in this direction in connection to abduction is to be found later in this part of the handbook, in Chap. 13.

form of an argument and those in which it manifests itself as an inference to the best explanation, is a useful one in philosophy, and was described in Sect. 10.2.

Section 10.3 described three *classical* characterizations of abduction, namely as a logical inference, as a computational process and as a process for epistemic change. These characterizations have been the dominant ones in the logical literature and are still a point of reference to new proposals which go beyond the boundaries of the classical way.

The history of the logic of abduction and of its formal modeling is still on the making. The chapters to follow offer an overview of formal tools to model logical and computational abduction.

**Acknowledgments.** Research for this article was supported by the research project *Logics of Discovery, Heuristics, and Creativity in the Sciences* (PAPIIT, IN400514) granted by UNAM.

## References

- 10.1 B. Russell: *The Art of Philosophizing and Other Essays* (Adams, Totowa, Littlefield 1974)
- 10.2 P. Thagard: *Computational Philosophy of Science* (MIT, Cambridge 1988)
- 10.3 J.S. Mill: A system of logic. In: *The Collected Works of John Stuart Mill*, ed. by J.M. Robson (Routledge and Kegan Paul, London 1958), New York, Harper and brothers
- 10.4 G. Harman: The inference to the best explanation, *Philos. Rev.* **74**(1), 88–95 (1965)
- 10.5 E. Shapiro: Inductive inference of theories from facts. In: *Computational Logic: Essays in Honor of Alan Robinson*, ed. by J.L. Lassez, G. Plotkin (MIT, Cambridge 1991)
- 10.6 P. Flach, A. Kakas (Eds.): *Abduction and Induction. Essays on Their Relation and Integration* (Kluwer Academic, Dordrecht 2000)
- 10.7 W. Salmon: Scientific explanation. In: *Introduction to the Philosophy of Science*, Vol. 1–6, ed. by W. Salmon, J. Earman, C. Glymour, J. Lennox, K. Schaffner, W.C. Salmon, J.D. Norton, J.E. McGuire, P. Machamer, J.G. Lennox (Prentice Hall, New York 1992)
- 10.8 A. Aliseda: *Abductive Reasoning. Logical Investigation into Discovery and Explanation*, Vol. 330 (Springer, Dordrecht 2006)
- 10.9 G. Schurz: Patterns of abduction, *Synthese* **164**, 201–234 (2008)
- 10.10 D. Anderson: *The Evolution of Peirce's Concept of Abduction*, *Trans. Charles S. Peirce Society*, Vol. 22 (Indiana Univ. Press, Bloomington 1986) pp. 145–164
- 10.11 J. Hintikka: What is abduction? The fundamental problem of contemporary epistemology, *Trans. Charles S. Peirce Soc.* **34**(3), 503–533 (1998)
- 10.12 A. Aliseda: Abduction as epistemic change: A Peircean model in artificial intelligence. In: *Abduction and Induction*, ed. by P. Flach, A. Kakas (Kluwer Academic, Dordrecht 2000) pp. 45–58
- 10.13 C.S. Peirce: *1867–1913. Collected Papers of Charles Sanders Peirce*. Vols. 1–6, ed. by C. Hartshorne, P. Weiss. (Harvard Univ. Press, Cambridge 1934)
- 10.14 H.E. Pople: *On the Mechanization of Abductive Logic* (Morgan Kaufmann, San Francisco 1973) pp. 147–152
- 10.15 Y. Peng, J.A. Reggia: *Abductive Inference Models for Diagnostic Problem-Solving, Symbolic Computation: Artificial Intelligence* (Springer, New York 1990)
- 10.16 G. Paul: Approaches to abductive reasoning: An overview, *Artif. Intell. Rev.* **7**(2), 109–152 (1993)
- 10.17 K. Konolige: Abductive theories in artificial intelligence. In: *Principles of Knowledge Representation*, ed. by G. Brewka (CSLI Publications, Stanford 1996)
- 10.18 P. Paul: AI approaches to abduction. In: *Abductive Reasoning and Uncertainty Management Systems, Handbook of Defeasible Reasoning and Uncertainty Management Systems*, Vol. 4, ed. by D. Gabbay, R. Kruse (Kluwer Academic, Dordrecht 2000) pp. 35–98
- 10.19 L. Magnani: *Abduction, Reason and Science: Processes of Discovery and Explanation* (Kluwer/Plenum, New York 2001)
- 10.20 H. Simon: *Models of Discovery* (Reidel, Holland 1977)
- 10.21 N. Chomsky: *Language and Mind. Enlarged edition* (Harcourt Brace Jovanovich, New York 1972)
- 10.22 C. Hempel: Aspects of scientific explanation. In: *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, ed. by C. Hempel (The Free Press, New York 1965)
- 10.23 I. Niiniluoto: Statistical explanation reconsidered, *Synthese* **48**(3), 437–472 (1981)
- 10.24 I. Niiniluoto: Hempel's theory of statistical explanation. In: *Science, Explanation, and Rationality: The Philosophy of Carl G. Hempel*, ed. by J.H. Fetzer (Oxford Univ. Press, Oxford 2000) pp. 138–163
- 10.25 T. Kuipers: Abduction aiming at empirical progress or even truth approximation leading to a challenge for computational modelling, *Found. Sci.* **4**(3), 307–323 (1999)
- 10.26 A.C. Kakas, R.A. Kowalski, F. Toni: Abductive logic programming, *J. Logic Comput.* **2**(6), 719–770 (1992)
- 10.27 M.C. Mayer, F. Pirri: First order abduction via tableau and sequent calculi, *Logic J. IGPL* **1**(1), 99–117 (1993)
- 10.28 D. Makinson: General patterns in nonmonotonic reasoning. In: *Handbook of Logic in Artificial Intelligence and Logic Programming, Nonmonotonic Reasoning and Uncertain Reasoning*, Vol. 3, ed. by C.J. Hogger, D.M. Gabbay, J.A. Robinson (Oxford Science Publications, Clarendon, Oxford 1994) pp. 35–110
- 10.29 C. Boutilier, V. Becher: Abduction as belief revision, *Artif. Intell.* **77**(1), 43–94 (1995)
- 10.30 J. Meheus, L. Verhoeven, M. Van Dyck, D. Provijn: Ampliative adaptive logics and the foundation of logic-based approaches to abduction. In: *Logical and Computational Aspects of Model-Based Reasoning*, ed. by L. Magnani, N.J. Nersessian, C. Pizzi (Kluwer Academic, Dordrecht 2002) pp. 39–71
- 10.31 M. Beirlaen, A. Aliseda: A conditional logic for abduction, *Synthese* **191**(15), 3733–3758 (2014)
- 10.32 J.R. Josephson: Smart inductive generalizations are abductions. In: *Abduction and Induction*, ed. by P. Flach, A. Kakas (Kluwer Academic, Dordrecht 2000) pp. 31–44

- 10.33 I. Douven: Abduction. In: *The Stanford Encyclopedia of Philosophy*, Spring 2011 edn., ed. by E. Zalta (2011), <http://plato.stanford.edu/archives/spr2011/entries/abduction/>, Date of last access: June 8th 2016
- 10.34 Y. Shoham: *Reasoning About Change. Time and Causation from the Standpoint of Artificial Intelligence* (MIT, Cambridge 1988)
- 10.35 D. Dubois, H. Prade: Possibilistic logic, preferential models, non-monotonicity and related issues, Proc. 12th Int. Joint Conf. on Artificial Intelligence (Morgan Kaufman, Burlington 1991) pp. 419–424
- 10.36 L. Magnani (Ed.): Special issue: Abduction, practical reasoning, and creative inferences in science, *Logic J. IGPL* **14**(2) (2006)
- 10.37 L. Magnani, W. Carnielli, C. Pizz (Eds.): Special issue: Formal representations in model-based reasoning and abduction, *Logic J. IGPL* **20**(2) (2012)
- 10.38 L. Magnani (Ed.): Special issue: Formal representations in model-based reasoning and abduction, *Logic J. IGPL* **21**(6) (2013)
- 10.39 A. Aliseda, L. Leonides: Hypotheses testing in adaptive logics: An application to medical diagnosis, *Logic J. IGPL* **21**(6), 915–930 (2013)
- 10.40 J. Woods: Ignorance and semantic tableaux: Aliseda on abduction, *Theoria*. **22**(3), 305–318 (2007)
- 10.41 R.A. Kowalski: *Logic for Problem Solving* (Elsevier, New York 1979)
- 10.42 J.W. Lloyd: *Foundations of Logic Programming*, 2nd edn. (Springer, Berlin, Heidelberg 1987)
- 10.43 A.C. Kakas, R.A. Kowalski, F. Toni: The role of abduction in logic programming. In: *Handbook of Logic in Artificial Intelligence and Logic Programming*, Vol. 5, ed. by C.J. Hogger, D.M. Gabbay, J.A. Robinson (Clarendon, Oxford 1998) pp. 235–324
- 10.44 M. Denecker, A.C. Kakas: Abduction in logic programming. In: *Computational Logic: Logic Programming and Beyond*, ed. by A.C. Kakas, F. Sadri (Springer, Berlin, Heidelberg 2002) pp. 402–436
- 10.45 R. Kowalski, D. Kuehner: Linear resolution with selection function, *Artif. Intell.* **2**(3–4), 227–260 (1971)
- 10.46 R. Hähnle, M. D'Agostino, D.M. Gabbay, J. Posegga (Eds.): *Handbook of Tableau Methods* (Kluwer Academic, Dordrecht 1999)
- 10.47 P. Gärdenfors: Belief revision: An introduction. In: *Belief Revision, Cambridge Tracts in Theoretical Computer Science*, ed. by P. Gärdenfors (Cambridge Univ. Press, Cambridge 1992) pp. 1–28
- 10.48 P. Gärdenfors, H. Rott: Belief revision. In: *Handbook of Logic in Artificial Intelligence and Logic Programming*, , Oxford Science Publications, ed. by C.J. Hogger, D.M. Gabbay, J.A. Robinson, Vol. 4, (Clarendon, Oxford 1995) pp. 35–132
- 10.49 P. Gärdenfors: *Knowledge in Flux: Modeling the Dynamics of Epistemic States* (MIT, Cambridge 1988)
- 10.50 C.E. Alchourrón, P. Gärdenfors, D. Makinson: On the logic of theory change: Partial meet contraction and revision functions, *J. Symb. Logic* **50**(2), 510–530 (1985)

# 11. Qualitative Inductive Generalization and Confirmation

Mathieu Beirlaen

Part C | 11.1

Inductive generalization is a defeasible type of inference which we use to reason from the particular to the universal. First, a number of systems are presented that provide different ways of implementing this inference pattern within first-order logic. These systems are defined within the adaptive logics framework for modeling defeasible reasoning. Next, the logics are re-interpreted as criteria of confirmation. It is argued that they withstand the comparison with two qualitative theories of confirmation, Hempel's satisfaction criterion and hypothetico-deductive confirmation.

11.1	<b>Adaptive Logics for Inductive Generalization</b> .....	231
11.2	<b>A First Logic for Inductive Generalization</b> .....	232
11.2.1	<b>General Characterization of the Standard Format</b> .....	232
11.2.2	<b>Proof Theory</b> .....	233
11.2.3	<b>Minimal Abnormality</b> .....	236
11.3	<b>More Adaptive Logics for Inductive Generalization</b> .....	237
11.4	<b>Qualitative Inductive Generalization and Confirmation</b> .....	240
11.4.1	<b>I-Confirmation and Hempel's Adequacy Conditions</b> ....	240
11.4.2	<b>I-Confirmation and the Hypothetico-Deductive Model</b>	242
11.4.3	<b>Interdependent Abnormalities and Heuristic Guidance</b> .....	243
11.5	<b>Conclusions</b> .....	245
11.A	<b>Appendix: Blocking the Raven Paradox?</b> .....	246
	<b>References</b> .....	247

Logics of induction are tools for evaluating the strength of arguments which are not deductively valid. There are many kinds of argument the conclusion of which is not guaranteed to follow from its premises, and there are many ways to evaluate the strength of such arguments. This chapter focusses on one particular kind of non-deductive argument, and on one particular method of implementation. The type of argument under consideration here is that of inductive generalization, as when we reason from the particular to the universal. A num-

ber of logics are discussed which permit us, given a set of objects sharing or not sharing a number of properties, to infer generalizations of the form *All x are P*, or *All x with property P share property Q*. Inductive generalization is a common practice which has proven its use in scientific endeavor. For instance, given the fact that the relatively few electrons measured so far carry a charge of  $-1.6 \times 10^{-19}$  Coulombs, we believe that all electrons have this charge [11.1].

## 11.1 Adaptive Logics for Inductive Generalization

The methods used here for formalizing practices of inductive generalization stem from the adaptive logics framework. Adaptive logics are tools developed for modeling defeasible reasoning, equipped with a proof theory that nicely captures the dynamics of non-monotonic – in this case, inductive – inference. In proofs for adaptive logics for inductive generalization, the conditional introduction of generalizations is allowed. The proof theory is also equipped with a mechanism taking care that conditionally introduced generalizations get retracted in case their condition is violated, for in-

stance when the generalization in question is falsified by the premises.

In Sect. 11.2 and 11.3 the general framework of adaptive logics is introduced, and a number of existing adaptive logics for inductive generalization are defined. The differences between these logics arise from different choices made along one of two dimensions. A first dimension concerns the specific condition required for introducing generalizations in an adaptive proof. A very permissive approach allows for their free introduction, without taking into account the specifics

of the premises. This is the idea behind the logic **LI**. A more economical approach is to permit the introduction of a generalization on the condition that at least one instance of it is present. This is the rationale behind a second logic, **IL**. In an **IL**-proof a generalization *All P are Q* can be introduced only if the premise set contains at least one object which is either *not-P* or *Q*. More economical still is the rationale behind a third logic, **G**, which aims to capture the requirement of knowing at least one *positive* instance of a generalization before introducing it in a proof. That is, in a **G**-proof a generalization *All P are Q* can be introduced if the premise set contains at least one object which is *both P* and *Q*.

The second dimension along which different consequence relations are generated concerns the specific mechanism used for retracting generalizations introduced in adaptive proofs. It is often not sufficient to demand retraction just in case a generalization is falsified by the premises. For instance, if the consequence sets of our logics are to be closed under classical logic, jointly incompatible generalizations should not be derivable, even though none of them is falsified by our premise set. Within the adaptive logics framework, various strategies are available for retracting conditional moves in an adaptive proof. Two such strategies are presented in this chapter: the reliability strategy and the minimal abnormality strategy.

Combining both dimensions, a family of six adaptive logics for inductive generalization is obtained (it contains the systems **LI**, **IL**, and **G**, each of which can be defined using either the reliability or the minimal abnormality strategy). These logics have all been presented elsewhere (for **LI**, see [11.2–4]. For **IL** and **G**, see [11.5]). The original contribution of this chapter consists in a study comparing these systems to some

existing qualitative criteria of confirmation. There is an overlap between the fields of inductive logic and confirmation theory. In 1943 already, Hempel noted that the development of a logical theory of confirmation might be regarded as a contribution to the field of inductive logic [11.6, p. 123]. In Sect. 11.4 the logics from Sect. 11.2 and 11.3 are re-interpreted as qualitative criteria of confirmation, and are related to other qualitative models of confirmation: Hempel's satisfaction criterion (Sect. 11.4.1) and the hypothetico-deductive model (Sect. 11.4.2). Section 11.4 ends with some remarks on the heuristic guidance that adaptive logics for inductive generalization can provide in the derivation and subsequent confirmation of additional generalizations (Sect. 11.4.3).

The following notational conventions are used throughout the chapter. The formal language used is that of first-order logic without identity. A primitive functional formula of rank 1 is an open formula that does not contain any logical symbols ( $\exists, \forall, \neg, \vee, \wedge, \supset, \equiv$ ), sentential letters, or individual constants, and that contains only predicate letters of rank 1. The set of functional atoms of rank 1, denoted  $\mathcal{A}^f$ , comprises the primitive functional formulas of rank 1 and their negations. A *generalization* is the universal closure of a disjunction of members of  $\mathcal{A}^f$ . That is, the set of generalizations in this technical sense is the set  $\{\forall(A_1 \vee \dots \vee A_n) \mid A_1, \dots, A_n \in \mathcal{A}^f; n \geq 1\}$ , where  $\forall$  denotes the universal closure of the subsequent formula. Occasionally the term generalization is also used for formulas equivalent to a member of this set, e.g.,  $\forall x(Px \supset Qx)$ . It is easily checked that generalizations  $\forall(A_1 \vee \dots \vee A_n)$  can be rewritten as formulas of the general form  $\forall((B_1 \wedge \dots \wedge B_j) \supset (C_1 \vee \dots \vee C_k))$ , and vice versa, where all  $B_i$  and  $C_j$  belong to  $\mathcal{A}^f$ .

## 11.2 A First Logic for Inductive Generalization

In this section the standard format (SF) for adaptive logics is introduced and explained. Its features are illustrated by means of the logic **LI** from [11.3, 4], chronologically the first adaptive logic for inductive generalization. A general characterization of the SF is provided, and its proof theory is explained. For a more comprehensive introduction, including the semantics and generic meta-theory of the SF, see, e.g., [11.7, 8].

### 11.2.1 General Characterization of the Standard Format

An adaptive logic (AL) within the SF is defined as a triple, consisting of:

- (i) A *lower limit logic* (LLL), a logic that has static proofs and contains classical disjunction
- (ii) A *set of abnormalities*, a set of formulas that share a (possibly) restricted logical form, or a union of such sets
- (iii) An *adaptive strategy*.

The LLL is the stable part of the AL: anything derivable by means of the LLL is derivable by means of the AL. Explaining the notion of static proofs is beyond the scope of this chapter. For a full account, see [11.9]. (Alternatively, the static proofs requirement can be replaced by the requirement that the lower limit logic has a reflexive, monotonic, transitive, and compact consequence relation [11.8].) In any case, it suffices to know

that the first-order fragment of Classical Logic (**CL**) meets this requirement, as we work almost exclusively with **CL** as a LLL. The lower limit logic of **LI** is **CL**.

Typically, an AL enables one to derive, for most premise sets, some extra consequences on top of those that are LLL-derivable. These supplementary consequences are obtained by interpreting a premise set as *normally as possible*, or, equivalently, by supposing abnormalities to be false *unless and until proven otherwise*. What it means to interpret a premise set as *normally as possible* is disambiguated by the strategy, element (iii).

The normality assumption made by the logics to be defined in this chapter amounts to supposing that the world is in some sense uniform. *Normal* situations are those in which it is safe to derive generalizations. *Abnormal* situations are those in which generalizations are falsified. In fact, the set of **LI**-abnormalities, denoted  $\Omega_{\mathbf{LI}}$ , is just the set of falsified generalizations (the definitions are those from [11.5]; in [11.10, Sect. 4.2.2] it is shown that the same logic is obtained if  $\Omega_{\mathbf{LI}}$  is defined as the set of formulas of the form  $\neg\forall xA(x)$ , where  $A$  contains no quantifiers, free variables, or constants)

$$\Omega_{\mathbf{LI}} =_{\text{df}} \{ \neg\forall (A_1 \vee \dots \vee A_n) \mid A_1, \dots, A_n \in \mathcal{A}^{f1}; n \geq 1 \} . \quad (11.1)$$

In adaptive proofs, it is possible to make conditional inferences assuming that one or more abnormalities are false. Whether or not such assumptions can be upheld in the continuation of the proof is determined by the adaptive strategy. The SF incorporates two adaptive strategies, the *reliability* strategy and the *minimal abnormality* strategy. In the generic proof theory of the SF, adaptive strategies come with a *marking definition*, which takes care of the withdrawal of certain conditional inferences in dynamic proofs. It will be easier to explain the intuitions behind these strategies after defining the generic proof theory for ALs. For now, just note that in the remainder **LI** is ambiguous between **LI<sup>r</sup>** and **LI<sup>m</sup>**, where the subscripts  $r$  and  $m$  denote the reliability strategy, respectively the minimal abnormality strategy. Analogously for the other logics defined below.

### 11.2.2 Proof Theory

Adaptive proofs are *dynamic* in the sense that lines derived at a certain stage of a proof may be withdrawn at a later stage. Moreover, lines withdrawn at a certain stage can become derivable again at an even later stage, and so on. (A stage of a proof is a sequence of lines and a proof is a sequence of stages. Every proof starts off with stage 1. Adding a line to a proof by applying

one of the rules of inference brings the proof to its next stage, which is the sequence of all lines written so far.)

A line in an adaptive proof consists of four elements: a line number, a formula, a justification and a *condition*. For instance, a line

$$j \quad A \quad i_1, \dots, i_n; R \quad \Delta,$$

reads: at line  $j$ , the formula  $A$  is derived from lines  $i_1 - i_n$  by rule  $R$  on the condition  $\Delta$ . The fourth element, the condition, is what permits the dynamics. Intuitively, the condition of a line in a proof corresponds to an assumption made at that line. In the example above,  $A$  was derived on the assumption that the formulas in  $\Delta$  are false. If, later on in the proof, it turns out that this assumption was too bold, the line in question is withdrawn from the proof by a marking mechanism corresponding to an adaptive strategy. Importantly, only members of the set of abnormalities are allowed as elements of the condition of a line in an adaptive proof. Thus, assumptions always correspond to the falsity of one or more abnormalities, or, equivalently, to the truth of one or more generalizations.

Before explaining how the marking mechanism works, the generic inference rules of the SF must be introduced. There are three of them: a premise introduction rule (Prem), an unconditional rule (RU), and a conditional rule (RC). For adaptive logics with **CL** as their LLL, they are defined as follows

$$\begin{array}{ll} \text{Prem} & \text{If } A \in \Gamma : \\ & \frac{\dots \quad \dots}{A \quad \emptyset} \\ \text{RU} & \text{If } A_1, \dots, A_n \vdash_{\mathbf{CL}} B : \\ & \frac{A_1 \quad \Delta_1}{\vdots \quad \vdots} \\ & \frac{A_n \quad \Delta_n}{\hline B \quad \Delta_1 \cup \dots \cup \Delta_n} \\ \text{RC} & \text{If } A_1, \dots, A_n \vdash_{\mathbf{CL}} B \vee \text{Dab}(\Theta) : \\ & \frac{A_1 \quad \Delta_1}{\vdots \quad \vdots} \\ & \frac{A_n \quad \Delta_n}{\hline B \quad \Delta_1 \cup \dots \cup \Delta_n \cup \Theta} \end{array}$$

Where  $\Gamma$  is the premise set, Prem permits the introduction of premises on the empty condition at any time in the proof. Remember that conditions, at the intuitive level, correspond to assumptions, so Prem stipulates that premises can be introduced at any time without making any further assumptions.

Since ALs strengthen their LLL, one or more rules are needed to incorporate LLL-inferences in AL-proofs. In the proof theory of the SF, this is taken care of by the generic rule RU. This rule stipulates that whenever  $B$  is a **CL**-consequence of  $A_1, \dots, A_n$ , and all of  $A_1, \dots, A_n$  have been derived in a proof, then  $B$  is derivable, provided that the conditions attached to the lines at which  $A_1, \dots, A_n$  were derived are carried over. Intuitively, if  $A_1, \dots, A_n$  are derivable assuming that the members of  $\Delta_1, \dots, \Delta_n$  are false, and if  $B$  is a **CL**-consequence of  $A_1, \dots, A_n$ , then  $B$  is derivable, still assuming that all members of  $\Delta_1, \dots, \Delta_n$  are false.

Before turning to RC, here is an example illustrating the use of the rules Prem and RU. Let  $\Gamma_1 = \{Pa \wedge Qa, Pb, \neg Qc\}$ . Suppose we start an **LI**-proof for  $\Gamma_1$  as follows

1	$Pa \wedge Qa$	Prem	$\emptyset$
2	$Pb$	Prem	$\emptyset$
3	$\neg Qc$	Prem	$\emptyset$
4	$Pa$	1; RU	$\emptyset$
5	$Qa$	1; RU	$\emptyset$

Let  $\Theta$  be a finite set of **LI**-abnormalities, that is,  $\Theta \subset \Omega_{\text{LI}}$ . Then  $\text{Dab}(\Theta)$  refers to the classical disjunction of the members of  $\Theta$  ( $\text{Dab}$  abbreviates *disjunction of abnormalities*; in the remainder, such disjunctions are sometimes referred to as  $\text{Dab}$ -formulas). RC stipulates that, whenever  $B$  is **CL**-derivable from  $A_1, \dots, A_n$  in disjunction with one or more abnormalities, then  $B$  can be inferred assuming that these abnormalities are false, i. e., we can derive  $B$  and add the abnormalities in question to the condition set, together with assumptions made at the lines at which  $A_1, \dots, A_n$  were derived.

For instance, (11.2) is **CL**-valid

$$\forall x(Px \vee Qx) \vee \neg \forall x(Px \vee Qx) \quad (11.2)$$

Note that the second disjunct of (11.2) is a member of  $\Omega_{\text{LI}}$ . In the context of inductive generalization the assumption that the world is as *normal* as possible corresponds to an assumption about the uniformity of the world. In adaptive proofs, such assumptions are made explicit by applications of the conditional rule. Concretely, if a formula like (11.2) is derived in an **LI**-proof, RC can be used to derive the first disjunct on the condition that the second disjunct is false. In fact, since (11.2) is a **CL**-theorem, the generalization  $\forall x(Px \vee Qx)$  can be introduced right away, taking its negation to be false (lines 1–5 are not repeated)

$$6 \quad \forall x(Px \vee Qx) \quad \text{RC} \quad \{\neg \forall x(Px \vee Qx)\}$$

In a similar fashion, RC can be used to derive other generalizations

7	$\forall x Px$	RC	$\{\neg \forall x Px\}$
8	$\forall x Qx$	RC	$\{\neg \forall x Qx\}$
9	$\forall x(\neg Px \vee Qx)$	RC	$\{\neg \forall x(\neg Px \vee Qx)\}$
10	$\forall x(Px \vee \neg Qx)$	RC	$\{\neg \forall x(Px \vee \neg Qx)\}$
11	$\forall x(\neg Px \vee \neg Qx)$	RC	$\{\neg \forall x(\neg Px \vee \neg Qx)\}$

Each generalization is derivable assuming that its corresponding condition is false. However, some of these assumptions clearly cannot be upheld. We know, for instance, that the generalizations derived at lines 8 and 11 are falsified by the premises at lines 3 and 1 respectively. So we need a way of distinguishing between *good* and *bad* inferred generalizations. This is where the adaptive strategy comes in. Since distinguishing *good* from *bad* generalizations can be done in different ways, there are different strategies available to us for making the distinction hard. First, the reliability strategy and its corresponding marking definition are introduced. The latter definition takes care of the retraction of *bad* generalizations.

Marking definitions proceed in terms of the minimal inferred  $\text{Dab}$ -formulas derived at a stage of a proof. A  $\text{Dab}$ -formula that is derived at a proof stage by RU at a line with condition  $\emptyset$  is called an *inferred Dab-formula* of the proof stage.

#### Definition 11.1 Minimal inferred Dab-formula

$\text{Dab}(\Delta)$  is a *minimal inferred Dab-formula* at stage  $s$  of a proof iff  $\text{Dab}(\Delta)$  is an inferred  $\text{Dab}$ -formula at stage  $s$  and there is no  $\Delta' \subset \Delta$  such that  $\text{Dab}(\Delta')$  is an inferred  $\text{Dab}$ -formula at stage  $s$ .

Where  $\text{Dab}(\Delta_1), \dots, \text{Dab}(\Delta_n)$  are the minimal inferred  $\text{Dab}$ -formulas derived at stage  $s$ ,  $U_s(\Gamma) = \Delta_1 \cup \dots \cup \Delta_n$  is the set of formulas that are *unreliable* at stage  $s$ .

#### Definition 11.2 Marking for reliability

Where  $\Delta$  is the condition of line  $i$ , line  $i$  is marked at stage  $s$  iff  $\Delta \cap U_s(\Gamma) \neq \emptyset$ .

To illustrate the marking mechanism, consider the following extension of the **LI**-proof for  $\Gamma_1$  (marked lines are indicated by a  $\checkmark$ -sign; lines 1–5 are not repeated in the proof)

6	$\forall x(Px \vee Qx)$	RC	$\{\neg \forall x(Px \vee Qx)\} \checkmark$
7	$\forall x Px$	RC	$\{\neg \forall x Px\} \checkmark$
8	$\forall x Qx$	RC	$\{\neg \forall x Qx\} \checkmark$



9	$\forall x(\neg Px \vee Qx)$ $\{\neg \forall x(\neg Px \vee Qx)\} \checkmark$	RC
10	$\forall x(Px \vee \neg Qx)$ $\{\neg \forall x(Px \vee \neg Qx)\}$	RC
11	$\forall x(\neg Px \vee \neg Qx)$ $\{\neg \forall x(\neg Px \vee \neg Qx)\} \checkmark$	RC
12	$\neg \forall x Qx$ $\emptyset$	3; RU
13	$\neg \forall x(\neg Px \vee \neg Qx)$ $\emptyset$	1; RU
14	$\neg \forall x Px \vee \neg \forall x(\neg Px \vee Qx)$ $\emptyset$	3; RU
15	$\neg \forall x(Px \vee Qx) \vee \neg \forall x(\neg Px \vee Qx)$ $\emptyset$	3; RU

As remarked above, the generalizations derived at lines 8 and 11 are falsified by the premises, so it makes good sense to mark them and thereby consider them not derived anymore. As soon as we derive the negations of these generalizations (lines 12 and 13) Definition 11.2 takes care that lines 8 and 11 are marked. The generalizations derived at lines 6, 7, and 9 are not falsified by the data, yet they are marked according to Definition 11.2, due to the derivability of the minimal inferred Dab-disjunctions at lines 14 and 15. We know, for instance, that the generalizations derived at lines 7 and 9 cannot be upheld together: at line 14 we inferred that they are jointly incompatible in view of the premises. Definition 11.2 takes care that *both* lines 7 and 9 are marked at stage 15, since

$$U_{15}(\Gamma_1) = \{\neg \forall x Px, \neg \forall x Qx, \neg \forall x(Px \vee Qx), \\ \neg \forall x(\neg Px \vee Qx), \neg \forall x(\neg Px \vee \neg Qx)\}. \quad (11.3)$$

The only inferred generalization left unmarked at stage 15 is  $\forall x(Px \vee \neg Qx)$ , derived at line 10.

Due to the dynamics of adaptive proofs, we cannot just take a formula to be an AL-consequence of some premise set  $\Gamma$  once we derived it at some stage on an unmarked line in a proof for  $\Gamma$ , for it may be that there are extensions of the proof in which the line in question gets marked. Likewise, we need to take into account the fact that lines marked at a stage of a proof may become unmarked at a later stage. This is taken care of by using the concept of *final derivability*:

**Definition 11.3 Final derivability**

$A$  is *finally derived* from  $\Gamma$  at line  $i$  of a finite proof stage  $s$  iff (i)  $A$  is the second element of line  $i$ , (ii) line  $i$

is not marked at stage  $s$ , and (iii) every extension of the proof in which line  $i$  is marked may be further extended in such a way that line  $i$  is unmarked.

**Definition 11.4 Logical consequence for  $\mathbf{LI}^f$**

$\Gamma \vdash_{\mathbf{LI}^f} A$  ( $A$  is finally  $\mathbf{LI}^f$ -derivable from  $\Gamma$ ) iff  $A$  is finally derived at a line of an  $\mathbf{LI}^f$ -proof from  $\Gamma$ .

Given the premise set  $\Gamma_1$ , there are no extensions of the proof above in which any of the marked lines become unmarked, nor are there extensions in which line 10 is marked and cannot be unmarked again in a further extension of the proof. Hence, by Definitions 11.3 and 11.4

$$\Gamma_1 \not\vdash_{\mathbf{LI}^f} \forall x Px, \quad (11.4)$$

$$\Gamma_1 \not\vdash_{\mathbf{LI}^f} \forall x Qx, \quad (11.5)$$

$$\Gamma_1 \not\vdash_{\mathbf{LI}^f} \forall x(Px \vee Qx), \quad (11.6)$$

$$\Gamma_1 \vdash_{\mathbf{LI}^f} \forall x(Px \vee \neg Qx), \quad (11.7)$$

$$\Gamma_1 \not\vdash_{\mathbf{LI}^f} \forall x(\neg Px \vee Qx), \quad (11.8)$$

$$\Gamma_1 \not\vdash_{\mathbf{LI}^f} \forall x(\neg Px \vee \neg Qx). \quad (11.9)$$

The logic  $\mathbf{LI}^f$  is non-monotonic: adding new premises may block the derivation of generalizations that were finally derivable from the original premise set. For instance, suppose that we add the premise  $\neg Pd \wedge Qd$  to  $\Gamma_1$ . Since the extra premise provides a counter-instance to the generalization  $\forall x(Px \vee \neg Qx)$ , the latter should no longer be  $\mathbf{LI}^f$ -derivable from the new premise set. The following proof illustrates that this is indeed the case

1	$Pa \wedge Qa$ $\emptyset$	Prem
2	$Pb$ $\emptyset$	Prem
3	$\neg Qc$ $\emptyset$	Prem
4	$\neg Pd \wedge Qd$ $\emptyset$	Prem
5	$\forall x(Px \vee Qx)$ $\{\neg \forall x(Px \vee Qx)\} \checkmark$	RC
6	$\forall x Px$ $\{\neg \forall x Px\} \checkmark$	RC
7	$\forall x Qx$ $\{\neg \forall x Qx\} \checkmark$	RC
8	$\forall x(\neg Px \vee Qx)$ $\{\neg \forall x(\neg Px \vee Qx)\} \checkmark$	RC

9	$\forall x(Px \vee \neg Qx)$ $\{\neg \forall x(Px \vee \neg Qx)\} \checkmark$	RC
10	$\forall x(\neg Px \vee \neg Qx)$ $\{\neg \forall x(\neg Px \vee \neg Qx)\} \checkmark$	RC
11	$\neg \forall x Px$ $\emptyset$	4; RU
12	$\neg \forall x Qx$ $\emptyset$	3; RU
13	$\neg \forall x(\neg Px \vee \neg Qx)$ $\emptyset$	1; RU
14	$\neg \forall x(Px \vee Qx) \vee \neg \forall x(\neg Px \vee Qx)$ $\emptyset$	3; RU
15	$\neg \forall x(Px \vee \neg Qx)$ $\emptyset$	4; RU

Line 9 is marked in view of the Dab-formula derived at line 15. There is no way to extend this proof in such a way that the line in question gets unmarked. Hence,  $\Gamma_1 \cup \{\neg Pd \wedge Qd\} \not\vdash_{\mathbf{L}^{\mathbf{F}}} \forall x(Px \vee \neg Qx)$ . In fact, no nontautological generalizations whatsoever are  $\mathbf{L}^{\mathbf{F}}$ -derivable from the extended premise set  $\Gamma_1 \cup \{\neg Pd \wedge Qd\}$ .

### 11.2.3 Minimal Abnormality

Different interpretations of the same set of data may lead to different views concerning which generalizations should or should not be derivable. Each such view may be driven by its own rationale, and choosing one such rationale over the other is not a matter of pure logic. For that reason, different strategies are available to adaptive logicians, each interpreting a set of data in their own sensible way, depending on the context. The reliability strategy was defined already. The minimal abnormality strategy is slightly less skeptical. Consequently, for some premise sets, generalizations may be  $\mathbf{L}^{\mathbf{m}}$ -derivable, but not  $\mathbf{L}^{\mathbf{F}}$ -derivable.

Like reliability, the minimal abnormality strategy comes with its marking definition. Let a *choice set* of  $\Sigma = \{\Delta_1, \Delta_2, \dots\}$  be a set that contains one element out of each member of  $\Sigma$ . A *minimal choice set* of  $\Sigma$  is a choice set of  $\Sigma$  of which no proper subset is a choice set of  $\Sigma$ . Where  $\text{Dab}(\Delta_1), \text{Dab}(\Delta_2), \dots$  are the minimal inferred Dab-formulas derived from a premise set  $\Gamma$  at stage  $s$  of a proof,  $\Phi_s(\Gamma)$  is the set of minimal choice sets of  $\{\Delta_1, \Delta_2, \dots\}$ .

#### Definition 11.5 Marking for minimal abnormality

Where  $A$  is the formula and  $\Delta$  the condition of line  $i$ , line  $i$  is marked at stage  $s$  iff (i) there is no  $\varphi \in \Phi_s(\Gamma)$  such that  $\varphi \cap \Delta = \emptyset$ , or (ii) for some  $\varphi \in \Phi_s(\Gamma)$ , there

is no line at which  $A$  is derived on a condition  $\Theta$  for which  $\varphi \cap \Theta = \emptyset$ .

An example will clarify matters. Let  $\Gamma_2 = \{Pa \wedge Qa \wedge Ra, \neg Rb \wedge (\neg Pb \vee \neg Qb), \neg Pc \wedge \neg Qc \wedge Rc\}$ .

1	$Pa \wedge Qa \wedge Ra$ $\emptyset$	Prem
2	$\neg Rb \wedge (\neg Pb \vee \neg Qb)$ $\emptyset$	Prem
3	$\neg Pc \wedge \neg Qc \wedge Rc$ $\emptyset$	Prem
4	$\forall x(Px \vee Qx)$ $\{\neg \forall x(Px \vee Qx)\} \checkmark$	RC
5	$\forall x(Px \vee Rx)$ $\{\neg \forall x(Px \vee Rx)\} \checkmark$	RC
6	$\forall x(\neg Px \vee Rx)$ $\{\neg \forall x(\neg Px \vee Rx)\} \checkmark$	RC
7	$\neg \forall x(Px \vee Qx)$ $\emptyset$	3; RU
8	$\neg \forall x(Px \vee Rx) \vee \neg \forall x(\neg Px \vee Rx)$ $\emptyset$	2; RU
9	$\forall x(Px \vee Rx) \vee \forall x(\neg Px \vee Rx)$ $\{\neg \forall x(Px \vee Rx)\}$	5; RU
10	$\forall x(Px \vee Rx) \vee \forall x(\neg Px \vee Rx)$ $\{\neg \forall x(\neg Px \vee Rx)\}$	6; RU

To see what is happening in this proof, we need to understand the markings. Note that there are two minimal choice sets at stage 10

$$\Phi_{10}(\Gamma_2) = \{\{\neg \forall x(Px \vee Qx), \neg \forall x(Px \vee Rx)\}, \{\neg \forall x(Px \vee Qx), \neg \forall x(\neg Px \vee Rx)\}\} . \quad (11.10)$$

Line 4 is marked in view of clause (i) in Definition 11.5, since its condition intersects with each minimal choice set in  $\Phi_{10}(\Gamma_2)$ . Lines 5 and 6 are marked in view of clause (ii) in Definition 11.5. For the minimal choice set  $\{\neg \forall x(Px \vee Qx), \neg \forall x(Px \vee Rx)\}$ , there is no line at which  $\forall x(Px \vee Rx)$  was derived on a condition that does not intersect with this set. Hence line 5 is marked. Analogously, line 6 is marked because, for the minimal choice set  $\{\neg \forall x(Px \vee Qx), \neg \forall x(\neg Px \vee Rx)\}$ , there is no line at which  $\forall x(\neg Px \vee Rx)$  was derived on a condition that does not intersect with this set.

Things change, however, when we turn to lines 9 and 10. In these cases, none of clauses (i) or (ii) of Def-

inition 11.5 apply: for each of these lines, there is a minimal choice set in  $\Phi_{10}(\Gamma_2)$  which does not intersect with the line's condition; and for each of the sets in  $\Phi_{10}(\Gamma_2)$ , we have derived the formula  $\forall x(Px \vee Rx) \vee \forall x(\neg Px \vee Rx)$  on a condition that does not intersect with it. Hence, these lines remain unmarked at stage 10 of the proof.

Things would have been different if we made use of the reliability strategy, since

$$U_{10}(\Gamma_2) = \{ \neg \forall x(Px \vee Qx), \neg \forall x(Px \vee Rx), \\ \neg \forall x(\neg Px \vee Rx) \}. \quad (11.11)$$

In view of  $U_{10}(\Gamma_2)$  and Definition 11.2, all of lines 4–6 and 9–10 would be marked if the above proof were a  $\mathbf{LI}^{\mathbf{F}}$ -proof.

As with the reliability strategy, logical consequence for the minimal abnormality strategy is defined in terms of final derivability (Definition 11.3). A consequence relation for  $\mathbf{LI}^{\mathbf{m}}$  is defined simply by replacing all occurrences of  $\mathbf{LI}^{\mathbf{F}}$  in Definition 11.4 with  $\mathbf{LI}^{\mathbf{m}}$ . Although the proof above can be extended in many interesting ways, showing the (non-)derivability of many more

generalizations than those currently occurring in the proof, nothing will change in terms of final derivability with respect to the formulas derived at stage 10

$$\Gamma_2 \not\vdash_{\mathbf{LI}^{\mathbf{m}}} \forall x(Px \vee Qx), \quad (11.12)$$

$$\Gamma_2 \not\vdash_{\mathbf{LI}^{\mathbf{m}}} \forall x(Px \vee Rx), \quad (11.13)$$

$$\Gamma_2 \not\vdash_{\mathbf{LI}^{\mathbf{m}}} \forall x(Px \vee \neg Rx), \quad (11.14)$$

$$\Gamma_2 \vdash_{\mathbf{LI}^{\mathbf{m}}} \forall x(Px \vee Rx) \vee \forall x(\neg Px \vee Rx), \quad (11.15)$$

$$\Gamma_2 \not\vdash_{\mathbf{LI}^{\mathbf{F}}} \forall x(Px \vee Qx), \quad (11.16)$$

$$\Gamma_2 \not\vdash_{\mathbf{LI}^{\mathbf{F}}} \forall x(Px \vee Rx), \quad (11.17)$$

$$\Gamma_2 \not\vdash_{\mathbf{LI}^{\mathbf{F}}} \forall x(Px \vee \neg Rx), \quad (11.18)$$

$$\Gamma_2 \not\vdash_{\mathbf{LI}^{\mathbf{F}}} \forall x(Px \vee Rx) \vee \forall x(\neg Px \vee Rx). \quad (11.19)$$

At the beginning of Sect. 11.2.3 it was mentioned that the rationale underlying the reliability strategy is slightly more skeptical than that underlying the minimal abnormality strategy. The point is illustrated by the proof for  $\Gamma_2$ . As we saw, the formula  $\forall x(Px \vee Rx) \vee \forall x(\neg Px \vee Rx)$  is  $\mathbf{LI}^{\mathbf{m}}$ -derivable from  $\Gamma_2$ , but not  $\mathbf{LI}^{\mathbf{F}}$ -derivable from  $\Gamma_2$ .

### 11.3 More Adaptive Logics for Inductive Generalization

$\mathbf{LI}$  interprets the world as *uniform* by taking as normal those situations in which a generalization is true, and as abnormal those situations in which a generalization is false. But of course, if uniformity is identified with the truth of *every* generalization in this way, the world can never be completely uniform (for the simple fact that many generalizations are incompatible and cannot be jointly true). Perhaps a more natural way to interpret the uniformity of the world is to take all objects to have the same properties: as soon as one object has property  $P$ , we try to infer that all objects have property  $P$ . This is the rationale behind the logic  $\mathbf{IL}$  from [11.5].

Roughly, the idea behind  $\mathbf{IL}$  is to generalize from instances. Given an instance, the derivation of a generalization is permitted on the condition that no counterinstances are derivable. So abnormal situations are those in which both an instance and a counter-instance of a generalization are present. This is the formal definition of the set of  $\mathbf{IL}$ -abnormalities

$$\Omega_{\mathbf{IL}} =_{\text{df}} \{ \exists(A_1 \vee \dots \vee A_n) \wedge \exists \neg(A_1 \vee \dots \vee A_n) \mid \\ A_1, \dots, A_n \in \mathcal{A}^{f1}; n \geq 1 \}. \quad (11.20)$$

The logic  $\mathbf{IL}$  is defined by the lower limit logic  $\mathbf{CL}$ , the set of abnormalities  $\Omega_{\mathbf{IL}}$ , and the adaptive strategy reliability ( $\mathbf{IL}^{\mathbf{r}}$ ) or minimal abnormality ( $\mathbf{IL}^{\mathbf{m}}$ ).

In an  $\mathbf{IL}$ -proof generalizations cannot be conditionally introduced from scratch, since an instance is required. In this respect,  $\mathbf{IL}$  is more demanding than  $\mathbf{LI}$ . However, it does not follow that for this reason  $\mathbf{IL}$  is a weaker logic, since it is also more difficult to derive (disjunctions of) abnormalities in  $\mathbf{IL}$ . A simple example will illustrate that, for many premise sets,  $\mathbf{IL}$  is in fact stronger than  $\mathbf{LI}$ . Consider the following  $\mathbf{IL}$ -proof from  $\Gamma_3 = \{Pa, \neg Pb \vee Qb\}$

1	$Pa$	Prem
	$\emptyset$	
2	$\neg Pb \vee Qb$	Prem
	$\emptyset$	
3	$\forall xPx$	1; RC
	$\{ \exists xPx \wedge \exists x\neg Px \}$	
4	$Qb$	2, 3; RU
	$\{ \exists xPx \wedge \exists x\neg Px \}$	
5	$\forall xQx$	4; RC
	$\{ \exists xPx \wedge \exists x\neg Px, \exists xQx \wedge \exists x\neg Qx \}$	

In view of  $Pa \vdash_{\mathbf{CL}} \forall xPx \vee (\exists xPx \wedge \exists x\neg Px)$ , we applied RC to line 1 and conditionally inferred  $\forall xPx$  at line 3. Next, we used RU to infer  $Qb$  from this newly obtained generalization together with the premise at line 2. We

now have an instance of  $\forall xQx$ , so we can conditionally infer the latter generalization, taking over the condition of line 4. Importantly, not a single disjunction of members of  $\Omega_{\mathbf{IL}}$  is **CL**-derivable from  $\Gamma_3$ . This means that there is no way to mark any of lines 3–5 in any extension of this proof, independently of which strategy we use.

Consequence relations for  $\mathbf{IL}^r$  and  $\mathbf{IL}^m$  are again definable in terms of final derivability (Definition 11.3). All we need to do is replace all occurrences of  $\mathbf{LI}^r$  in Definition 11.4 with  $\mathbf{IL}^r$ , respectively  $\mathbf{IL}^m$ . Hence

$$\Gamma_3 \vdash_{\mathbf{IL}} \forall xPx, \quad (11.21)$$

$$\Gamma_3 \vdash_{\mathbf{IL}} \forall xQx. \quad (11.22)$$

Compare the **IL**-proof above with the following **LI**-proof from  $\Gamma_3$

1	$Pa$	Prem
	$\emptyset$	
2	$\neg Pb \vee Qb$	Prem
	$\emptyset$	
3	$\forall xPx$	RC
	$\{\neg \forall xPx\} \checkmark$	
4	$Qb$	2, 3; RU
	$\{\neg \forall xPx\} \checkmark$	
5	$\forall xQx$	RC
	$\{\neg \forall xQx\} \checkmark$	
6	$\neg \forall xPx \vee \neg \forall x \neg Qx$	1, 2; RU
	$\emptyset$	
7	$\neg \forall xQx \vee \neg \forall x(\neg Px \vee \neg Qx)$	1, 2; RU
	$\emptyset$	

Independently of the adaptive strategy used (reliability or minimal abnormality), there are no extensions of this **LI**-proof in which any of lines 3–5 become unmarked. Therefore

$$\Gamma_3 \not\vdash_{\mathbf{LI}} \forall xPx, \quad (11.23)$$

$$\Gamma_3 \not\vdash_{\mathbf{LI}} \forall xQx. \quad (11.24)$$

The premise set  $\Gamma_3$  not only serves to show that **IL** is not strictly weaker than **LI** in terms of derivable generalizations. It also illustrates that, although in an **IL**-proof we generalize on the basis of instances, such an instance need not always be **CL**-derivable from the premise set. In the proof from  $\Gamma_3$ , we derived the generalization  $\forall xQx$  even though no instance of this generalization is **CL**-derivable from  $\Gamma_3$ . Instead, we first derived  $\forall xPx$  (of which  $\Gamma_3$  *does* provide us with an instance), and

then used this generalization to *infer* an instance of  $\forall xQx$ . This is perfectly in line with the intuition behind **IL**: If deriving a generalization on the basis of an instance leads us to more instances of other generalizations, then, assuming the world to be as uniform as possible, we take the world to be uniform with respect to these other generalizations as well.

When discussing inductive generalization, confirmation theorists often use the more fine-grained distinction between mere instances of a generalization, positive instances, and negative instances. For example, given a generalization  $\forall x(Px \supset Qx)$ , any  $a$  such that  $Pa \supset Qa$  is an *instance* of  $\forall x(Px \supset Qx)$ ; any  $a$  such that  $Pa \wedge Qa$  is a *positive instance* of  $\forall x(Px \supset Qx)$ ; and any  $a$  such that  $Pa \wedge \neg Qa$  is a *negative instance* of  $\forall x(Px \supset Qx)$ . Instead of requiring a mere instance before introducing a generalization, some confirmation theorists have suggested the stronger requirement for a positive instance, that is, a negative instance of the contrary generalization (Sect. 11.4.3). According to this idea, interpreting the world as uniform as possible amounts to generalizing whenever a positive instance is available to us. Abnormal situations, then, are those in which both a positive and a negative instance of a generalization are available to us. There is a corresponding variant of **IL** that hard-codes this idea in its set of abnormalities: the logic **G** from [11.5]. The latter is defined by the lower limit logic **CL**, the set of abnormalities  $\Omega_{\mathbf{G}}$  and either the reliability strategy ( $\mathbf{G}^r$ ) or the minimal abnormality strategy ( $\mathbf{G}^m$ ).

$$\begin{aligned} \Omega_{\mathbf{G}} =_{\text{df}} & \\ & \{ \exists(A_1 \wedge \dots \wedge A_n \wedge A_0) \wedge \exists(A_1 \wedge \dots \wedge A_n \wedge \neg A_0) \mid \\ & A_0, A_1, \dots, A_n \in \mathcal{A}^{f1}; n \geq 0 \}. \end{aligned} \quad (11.25)$$

In proofs to follow  $\exists(A_1 \wedge \dots \wedge A_n \wedge A_0) \wedge \exists(A_1 \wedge \dots \wedge A_n \wedge \neg A_0)$  is abbreviated as  $A_1 \wedge \dots \wedge A_n \wedge \pm A_0$  (where again  $A_0, A_1, \dots, A_n \in \mathcal{A}^{f1}$ ). As an illustration of the workings of **G**, consider the following **G**-proof from  $\Gamma_4 = \{Pa \wedge Qa, \neg Qb, \neg Pc\}$

1	$Pa \wedge Qa$	Prem
	$\emptyset$	
2	$\neg Qb$	Prem
	$\emptyset$	
3	$\neg Pc$	Prem
	$\emptyset$	
4	$\forall x(Px \supset Qx)$	1; RC
	$\{Px \wedge \pm Qx\}$	

5	$\forall x(Qx \supset Px)$ $\{Qx \wedge \pm Px\}$	1; RC
6	$\forall x(Px \equiv Qx)$ $\{Px \wedge \pm Qx, Qx \wedge \pm Px\}$	4, 5; RU
7	$\exists xPx \wedge \exists x\neg Px$ $\emptyset$	1, 3; RU
8	$\exists xQx \wedge \exists x\neg Qx$ $\emptyset$	1, 2; RU

The formulas derived at lines 4–6 are finally **G**-derivable in the proof. Since **G**-consequence too is defined in terms of final derivability, it follows, independently of the strategy used, that

$$\Gamma_4 \vdash_{\mathbf{G}} \forall x(Px \supset Qx), \quad (11.26)$$

$$\Gamma_4 \vdash_{\mathbf{G}} \forall x(Qx \supset Px), \quad (11.27)$$

$$\Gamma_4 \vdash_{\mathbf{G}} \forall x(Px \equiv Qx). \quad (11.28)$$

Now consider the following **IL**-proof from  $\Gamma_4$  (where  $A_1, \dots, A_n \in \mathcal{A}^{f1}$ ,  $!(A_1 \vee \dots \vee A_n)$  abbreviates  $\exists(A_1 \vee \dots \vee A_n) \wedge \exists \neg(A_1 \vee \dots \vee A_n)$ )

1	$Pa \wedge Qa$ $\emptyset$	Prem
2	$\neg Qb$ $\emptyset$	Prem
3	$\neg Pc$ $\emptyset$	Prem
4	$\forall x(Px \supset Qx)$ $\{!(\neg Px \vee Qx)\} \checkmark$	1; RC
5	$\forall x(Qx \supset Px)$ $\{!(\neg Qx \vee Px)\} \checkmark$	1; RC
6	$\forall x(Px \equiv Qx)$ $\{!(\neg Px \vee Qx), !( \neg Qx \vee Px)\} \checkmark$	4, 4; RU
7	$!Px$ $\emptyset$	1, 3; RU
8	$!Qx$ $\emptyset$	1, 2; RU
9	$!(Px \vee Qx) \vee !( \neg Px \vee Qx)$ $\emptyset$	1, 2; RU
10	$!(\neg Qx \vee Px) \vee !(Px \vee Qx)$ $\emptyset$	1, 3; RU
11	$!(\neg Px \vee \neg Qx)$ $\emptyset$	1, 2; RU

The minimal inferred Dab-formulas inferred at lines 7–11 will remain minimal in any extension of this proof

(none of the disjuncts of any of the formulas derived at lines 9 or 10 is separately derivable). Accordingly, the marks in this proof will not change. Hence, independently of the strategy used

$$\Gamma_4 \not\vdash_{\mathbf{IL}} \forall x(Px \supset Qx), \quad (11.29)$$

$$\Gamma_4 \not\vdash_{\mathbf{IL}} \forall x(Qx \supset Px), \quad (11.30)$$

$$\Gamma_4 \not\vdash_{\mathbf{IL}} \forall x(Px \equiv Qx). \quad (11.31)$$

Two more remarks are in order. First, the example above suggests that **G** is in general stronger than **IL**. This is correct for the minimal abnormality strategy, but false for the reliability strategy. An illustration is provided by the premise set  $\Gamma_5 = \{Pa, Qb, Rb, Qc, \neg Rc\}$ . The generalization  $\forall x(\neg Px \supset Qx)$  cannot be inferred on the condition  $\neg Px \wedge \pm Qx$ , since we lack a positive instance. It *can* be inferred on the conditions  $\pm Qx$  or  $\pm Px$  in view of  $\forall xQx \vdash_{\mathbf{CL}} \forall x(\neg Px \supset Qx)$  and  $\forall xPx \vdash_{\mathbf{CL}} \forall x(\neg Px \supset Qx)$ , but none of these conditions are reliable in view of the derivability of minimal Dab-formulas like  $\pm Px \vee (Px \wedge \pm Rx)$  and  $\pm Qx \vee (Qx \wedge \pm Px) \vee (Px \wedge \pm Rx)$ .

The situation is different in an **IL**<sup>r</sup>-proof, where deriving  $\forall x(\neg Px \supset Qx)$  on the condition  $!(Px \vee Qx)$  in a proof from  $\Gamma_5$  is both possible and final. That is, for every derivable Dab-formula in which  $!(Px \vee Qx)$  occurs, we can derive a shorter (minimal) disjunction of abnormalities in which it no longer occurs. Summing up

$$\Gamma_5 \not\vdash_{\mathbf{G}^r} \forall x(\neg Px \supset Qx), \quad (11.32)$$

$$\Gamma_5 \vdash_{\mathbf{IL}^r} \forall x(\neg Px \supset Qx). \quad (11.33)$$

The second remark is that the requirement for a *positive* instance before generalizing in a **G**-proof is still insufficient to guarantee that for every **G**-derivable generalization a positive instance is **CL**-derivable from the premises. The following proof from  $Pa$  illustrates the point

1	$Pa$	Prem	$\emptyset$
2	$\forall xPx$	1; RC	$\{\pm Px\}$
3	$\forall x(Qx \supset Px)$	2; RU	$\{\pm Px\}$

Independently of the strategy used, no means are available to mark line 3, hence  $Pa \vdash_{\mathbf{G}} \forall x(Qx \supset Px)$ , even though no positive instance of  $\forall x(Qx \supset Px)$  is available. More on this point below (see the discussion on Hempel's raven paradox in Sect. 11.4.1 and in the Appendix).

A total of six logics have been presented so far: the logics **LI**<sup>r</sup>, **LI**<sup>m</sup>, **IL**<sup>r</sup>, **IL**<sup>m</sup>, **G**<sup>r</sup>, and **G**<sup>m</sup>. Each of these systems interprets the claim that the world is uniform in a slightly different way, leading to slightly different log-

ics. Importantly, there is no Carnapian embarrassment of riches here: each of the systems has a clear intuition behind it.

The systems presented here can be combined so as to implement *Popper's* suggestion that more general hypotheses should be given precedence over less general ones [11.11]. For instance, if two generalizations  $\forall x(Px \supset Qx)$  and  $\forall x((Rx \wedge Sx) \supset Tx)$  are jointly incom-

patible with the premises, a combined system gives precedence to the more general hypothesis and delivers only  $\forall x(Px \supset Qx)$  as a consequence. There are various ways to hard-code this idea, resulting in various new combined adaptive logics for inductive generalization, each slightly different from the others. These combinations are not fully spelled out here. For a brief synopsis, see [11.5, Sect. 5].

## 11.4 Qualitative Inductive Generalization and Confirmation

Inductive logic and confirmation theory overlap to some extent. As early as 1943, *Hempel* noted that the development of a logical theory of confirmation might be regarded as a contribution to the field of inductive logic [11.6, p. 123]. Following Carnap and Popper's influential work on inductive logic and corroboration respectively, many of the existing criteria of confirmation are quantitative in nature, measuring the *degree* of confirmation of a hypothesis by the evidence, possibly taking into account auxiliary hypotheses and background knowledge. Here, the logics defined in the previous two sections are presented as *qualitative* criteria of confirmation, and are related to other qualitative models of confirmation. Quantitative criteria of confirmation are not considered. For *Carnap's* views on inductive logic, see [11.12]. For *Popper's*, see [11.11]. For introductions to inductive logic and probabilistic measures of confirmation, see, e.g., [11.13–16].

Let **I** be any adaptive logic for inductive generalization defined in one of the previous sections. (All remarks on **I**-confirmation readily generalize to the combined systems from [11.5, Sect. 5].) Where  $H$  is the hypothesis and  $\Gamma$  contains the evidence, **I**-confirmation is defined in terms of **I**-consequence:

### Definition 11.6 **I**-confirmation

$\Gamma$  **I**-confirms  $H$  iff  $\Gamma \vdash_{\mathbf{I}} H$ .

$\Gamma$  **I**-disconfirms  $H$  iff  $\Gamma \vdash_{\mathbf{I}} \neg H$ .

$\Gamma$  is **I**-neutral with respect to  $H$  iff  $\Gamma \not\vdash_{\mathbf{I}} H$  and  $\Gamma \not\vdash_{\mathbf{I}} \neg H$ .

This definition of **I**-confirmation has the virtue of simplicity and formal precision. The two main qualitative alternatives to **I**-confirmation are *Hempel's* satisfaction criterion and the hypothetico-deductive model of confirmation. In Sect. 11.4.1, **I**-confirmation is compared to *Hempel's* adequacy conditions, which serve as a basis for his satisfaction criterion. In Sect. 11.4.2, **I**-confirmation is compared to hypothetico-deductive confirmation. Section 11.4.3 concerns the use of the

criteria from Definition 11.6 as heuristic tools for hypothesis generation and confirmation.

### 11.4.1 **I**-Confirmation and *Hempel's* Adequacy Conditions

Let an *observation report* consist of a set of molecular sentences (sentences containing no free variables or quantifiers). According to *Hempel*, the following conditions should be satisfied by any adequate criterion for confirmation [11.17]:

- (1) *Entailment condition*: Any sentence which is entailed by an observation report is confirmed by it.
- (2) *Consequence condition*: If an observation report confirms every one of a class  $K$  of sentences, then it also confirms any sentence which is a logical consequence of  $K$ :
  - (a) *Special consequence condition*: If an observation report confirms a hypothesis  $H$ , then it also confirms every consequence of  $H$ .
  - (b) *Equivalence condition*: If an observation report confirms a hypothesis  $H$ , then it also confirms every hypothesis which is logically equivalent to  $H$ .
- (3) *Consistency condition*: Every logically consistent observation report is logically compatible with the class of all the hypotheses which it confirms.

If *logical consequence* is taken to be **CL**-consequence, as *Hempel* did, then **I**-confirmation satisfies conditions (1)-(3) no matter which adaptive logic for inductive generalization is used, due to **I**'s closure under **CL**. So all of the resulting criteria of confirmation meet *Hempel's* adequacy conditions. (For (3) the further property of *smoothness* or *reassurance* is required, from which it follows that the **I**-consequence set of consistent premise sets is consistent as well [11.7, Sect. 6].)

The definition of *Hempel's* own criterion requires some preparation (the formal presentation of *Hempel's* criterion is taken from [11.18]). An atomic formula  $A$

is *relevant* to a formula  $B$  iff there is some model  $M$  of  $A$  such that: if  $M'$  differs from  $M$  only in the value assigned to  $B$ ,  $M'$  is not a model of  $A$ . The *domain* of a formula  $A$  is the set of individual constants that occur in the atomic formulas that are relevant for  $A$ . The *development* of a universally quantified formula  $A$  for another formula  $B$  is the restriction of  $A$  to the domain of  $B$ , that is, the truth value of  $A$  is evaluated with respect to the domain of  $B$ . For instance, the domain of  $Pa \wedge (Pb \vee Qc)$  is  $\{a, b, c\}$  whereas the domain of  $Pa \wedge Qa$  is  $\{a\}$ ; and the development of  $\forall x(Px \supset Qx)$  for  $Pa \wedge \neg Qb$  is  $(Pa \supset Qa) \wedge (Pb \supset Qb)$ .

### Definition 11.7 Hempel's satisfaction criterion

An observation report  $E$  *directly confirms* a hypothesis  $H$  if  $E$  entails the development of  $H$  for  $E$ .

An observation report  $E$  *confirms* a hypothesis  $H$  if  $H$  is entailed by a class of sentences each of which is directly confirmed by  $E$ .

An observation report  $E$  *disconfirms* a hypothesis  $H$  if it confirms the denial of  $H$ .

An observation report  $E$  is *neutral* with respect to a hypothesis  $H$  if  $E$  neither confirms nor disconfirms  $H$ .

There are two reasons for arguing that Hempel's satisfaction criterion is too restrictive, and two reasons for arguing that it is too liberal. Each of these is discussed in turn. First, in order for the evidence to confirm a hypothesis  $H$  according to Hempel's criterion, *all* objects in the development of  $H$  must be known to be instances of  $H$ . This is a very strong requirement. **I**-confirmation is different in this respect. For instance,

$$Pa, Qa, \neg Pb, \neg Qb, Pc \vdash_{\mathbf{I}} \forall x(Px \supset Qx) . \quad (11.34)$$

In (11.34) it is unknown whether  $c$  instantiates the hypothesis  $\forall x(Px \supset Qx)$ , since the premises do not tell us whether  $Pc \supset Qc$ . The development of  $\forall x(Px \supset Qx)$  entails  $Pc \supset Qc$ , whereas the premise set of (11.34) does not. So the hypothesis  $\forall x(Px \supset Qx)$  is not directly confirmed by these premises according to the satisfaction criterion, nor is it entailed by one or more sentences which are directly confirmed by them. Therefore the satisfaction criterion judges the premises to be neutral with respect to the hypothesis  $\forall x(Px \supset Qx)$ , whereas (11.34) illustrates that  $\forall x(Px \supset Qx)$  is **I**-confirmed by these premises.

Second, given the law  $\forall x(Px \supset Rx)$ , the report  $\{Pa, Qa, Pb, Qb\}$ , does not confirm the hypothesis  $\forall x(Rx \supset Qx)$  according to Hempel's original formulation of the satisfaction criterion. The reason is that *auxiliary hypotheses* like  $\forall x(Px \supset Rx)$  contain quantifiers and therefore cannot be elements of observation reports. (The original formulation of Hempel's criterion

can, however, be adjusted so as to take into account background knowledge [11.19, 20].) For problems related to auxiliary hypotheses, see also Sect. 11.4.2. For now, it suffices to note that the criteria from Definition 11.6 do not face this problem, as quantified formulas are perfectly allowed to occur in premise sets. For instance, the set  $\{Pa, Qa, Pb, Qb, \forall x(Px \supset Rx)\}$  **I**-confirms the hypothesis  $\forall x(Rx \supset Qx)$

$$Pa, Qa, Pb, Qb, \forall x(Px \supset Rx) \vdash_{\mathbf{I}} \forall x(Rx \supset Qx) . \quad (11.35)$$

It seems, then, that **I**-confirmation is not too restrictive a criterion for confirmation. However, there are two senses in which **I**-confirmation, like Hempelian confirmation, can be said to be too liberal. The first has to do with Goodman's well-known *new riddle of induction* [11.21]. The family of adaptive logics for inductive generalization makes no distinction between regularities that are *projectible* and regularities that are not. Using Goodman's famous example, let an emerald be *grue* if it is green before January 1st 2020, and blue thereafter. Then the fact that all hitherto observed emeralds are *grue* confirms the hypothesis that all emeralds are *grue*. The latter regularity is not projectible into the future, as we do not seriously believe that in 2020 we will start observing blue emeralds. Nonetheless, it is perfectly fine to define a predicate denoting the property of being *grue*, just as it is perfectly fine to define a predicate denoting the property of being green. Yet the hypothesis *all emeralds are green* is projectible, whereas *all emeralds are grue* is not.

The problem of formulating precise rules for determining which regularities are projectible and which are not is difficult and important, but it is an epistemological problem that cannot be solved by purely logical means. Consequently, it falls outside the scope of this article. See [11.21] for Goodman's formulation and proposed solution of the problem, and [11.22] for a collection of essays on the projectibility of regularities.

Finally, one may argue that **I**-confirmation is too liberal on the basis of Hempel's own *raven paradox*. Where  $Ra$  abbreviates that  $a$  is a raven, and  $Ba$  abbreviates that  $a$  is black, a non-black non-raven **I**-confirms the hypothesis that all ravens are black

$$\neg Ba, \neg Ra \vdash_{\mathbf{I}} \forall x(Rx \supset Bx) . \quad (11.36)$$

Even the logic **G** does not block this inference. The reason is that we are given a positive instance of the generalization  $\forall x(\neg Bx \supset \neg Rx)$ , so we can derive this generalization on the condition  $\exists x(\neg Bx \wedge \neg Rx) \wedge \exists x(\neg Bx \wedge Rx)$ . As the generalization  $\forall x(\neg Bx \supset \neg Rx)$

is **G**-derivable from the premises, so is the logically equivalent hypothesis that all ravens are black,  $\forall x(Rx \supset Bx)$  (remember that **G**, like all logics defined in the previous section, is closed under **CL**).

Hempel's own reaction to the raven paradox was to bite the bullet and accept its conclusion [11.23]. According to Hempel, a non-black non-raven indeed confirms the raven hypothesis in case we did not know beforehand that the bird in question is not a raven. For example, if we observe a grey bird resembling a raven, then finding out that it was a crow confirms the raven hypothesis [11.18]. But as pointed out in [11.19] this defense is insufficient. Even in cases in which it is *known* that a non-black bird is not a raven, the bird in question, although irrelevant to the raven hypothesis, still confirms it.

If – like Hempel – one accepts its conclusion, the raven paradox poses no further problems for **I**-confirmation. Those who disagree are referred to the Appendix, where a relatively simple adaptive alternative to **G**-confirmation is defined which blocks the paradox by means of a non-material conditional invalidating the inference from *all non-black objects are non-ravens* to *all ravens are black*.

### 11.4.2 I-Confirmation and the Hypothetico-Deductive Model

If a hypothesis predicts an event which is observed at a later time, or if it subsumes a given observation report as a consequence of one of its postulates, then this counts as evidence in favor of the hypothesis. The hypothetico-deductive model of confirmation (HD confirmation) is an attempt to formalize this basic intuition according to which a piece of evidence confirms a hypothesis if the latter entails the evidence.

In its standard formulation, HD confirmation also takes into account auxiliary hypotheses. Where  $\Delta$  is a set of background information distinct from the evidence  $E$ ,

#### Definition 11.8 HD-confirmation

$E$  HD-confirms  $H$  relative to  $\Delta$  iff:

- (i)  $\{H\} \cup \Delta$  is consistent,
- (ii)  $\{H\} \cup \Delta$  entails  $E$  ( $\{H\} \cup \Delta \vdash E$ ),
- (iii)  $\Delta$  alone does not entail  $E$  ( $\Delta \not\vdash E$ ).

The intuitive difference conveyed by HD confirmation and Hempelian confirmation becomes concrete if HD confirmation is compared with Hempel's adequacy criteria from Sect. 11.4.1. Let  $H$  abbreviate *Black swans exist*, let  $E$  consist of a black swan, and let  $\Delta$  be

the empty set. Then, according to Hempel's entailment condition,  $H$  is confirmed by  $E$ , since  $E \vdash H$ . Not so according to HD confirmation, for condition (ii) of Definition 11.8 is violated ( $H \not\vdash E$ ) [11.24]. The same example illustrates how HD confirmation violates the following condition, which holds for the satisfaction criterion in view of Definition 11.7 [11.25]:

- (4) Complementarity condition:  $E$  confirms  $H$  iff  $E$  disconfirms  $\neg H$ .

The consequence condition too is clearly invalid for HD confirmation. For instance,  $Ra \supset Ba$  HD confirms  $\forall x(Rx \supset Bx)$ , but it does not HD confirm the weaker hypothesis  $\forall x(Rx \supset (Bx \vee Cx))$ , since  $\forall x(Rx \supset (Bx \vee Cx)) \not\vdash Ra \supset Ba$ .

An advantage of HD confirmation is that it fares better with the raven paradox. The observation of a black raven ( $Ra, Ba$ ) is not deducible from the raven hypothesis  $\forall x(Rx \supset Bx)$ , so black ravens do not in general confirm the raven hypothesis. But birds that are known to be ravens *do* confirm the raven hypothesis once it is established that they are black. For once it is known that an object is a raven, the observation that it is black is entailed by this knowledge together with the hypothesis ( $\forall x(Rx \supset Bx), Ra \vdash Ba$ ). Likewise, a non-black non-raven does not generally confirm the raven hypothesis. Only objects that are known to be non-black can confirm the hypothesis by establishing that they are not ravens. In formulas:  $\forall x(Rx \supset Bx), \neg Ba \vdash \neg Ra$ .

HD confirmation faces a number of standard objections, of which three are discussed here. The first is the problem of irrelevant conjunctions and disjunctions. In view of Definition 11.8 it is easily checked that whenever a hypothesis  $H$  confirms  $E$  relative to  $\Delta$ , so does  $H' = H \wedge K$  for any arbitrary  $K$  consistent with  $\Delta$ . Thus adding arbitrary conjuncts to confirmed hypotheses preserves confirmation. Dually, adding arbitrary disjuncts to the data likewise preserves confirmation. That is, whenever  $H$  confirms  $E$  relative to  $\Delta$ ,  $H$  also confirms  $E'$  relative to  $\Delta$ , where  $E' = E \vee F$  for any arbitrary  $F$ .

Various solutions have been proposed for dealing with such problems of irrelevancy, but as so often the devil is in the details (see [11.20] for a nice overview and further references). For present purposes, it suffices to say that **I**-confirmation is not threatened by problems of irrelevance. Clearly, if the evidence  $E$  **I**-confirms a hypothesis  $H$ , it does not follow that it **I**-confirms  $H \wedge K$  for some arbitrary  $K$  consistent with  $\Delta$ , since from  $\{E\} \cup \Delta \vdash_{\mathbf{I}} H$  it need not follow that  $\{E\} \cup \Delta \vdash_{\mathbf{I}} H \wedge K$ . Nor does it follow that  $E \vee F$  confirms  $H$  relative to  $\Delta$ , since from  $\{E\} \cup \Delta \vdash_{\mathbf{I}} H$  it need not follow that  $\{E \vee F\} \cup \Delta \vdash_{\mathbf{I}} H$ .

A second objection against HD confirmation concerns the inclusion of background information in Def-



inition 11.8. In general, this inclusion is an advantage, since evidence often does not (dis)confirm a hypothesis simpliciter. Rather, evidence (dis)confirms hypotheses with respect to a set of auxiliary (background) assumptions or theories. The vocabulary of a theory often extends beyond what is directly observable. Notwithstanding Hempel's conviction to the contrary, nowadays philosophers largely agree that the use of purely theoretical terms is both intelligible and necessary in science [11.26]. Making the confirmation relation relative to a set of auxiliaries allows for the inclusion of bridging principles connecting observation terms with theoretical terms, permitting purely theoretical hypotheses to be confirmed by pure observation statements [11.27]. However, making confirmation relative to background assumptions makes HD vulnerable to a type of objection often traced back to *Duhem* [11.28] and *Quine* [11.29]. Suppose that a hypothesis  $H$  entails an observation  $E$  relative to  $\Delta$ , and that  $E$  is found to be false. Then either (a)  $H$  is false or (b) a member of  $\Delta$  is false. But the evidence does not tell us which of (a) or (b) is the case, so we always have the option to retain  $H$  and blame some auxiliary hypothesis in the background information. More generally, one may object that what gets (dis)confirmed by observations is not a hypothesis taken by itself, but the conjunction of a hypothesis and a set of background assumptions or theories.

With *Elliott Sober*, we can counter such holistic objections by pointing to the different epistemic status of hypotheses *under* test and auxiliary hypotheses (or hypotheses *used* in a test). Auxiliaries are independently testable, and when used in an experiment we already have good reasons to think of these hypotheses as true. Moreover, they are epistemically independent of the test outcome. So if a hypothesis is disconfirmed by the HD criterion, we can, in the vast majority of cases, maintain that it is the hypothesis we need to retract, and not one of the background assumptions [11.30].

A parallel point can be made concerning **I**-confirmation. Here too, we can add to the premises a set  $\Delta$  of auxiliary or background assumptions. And here too, we can use Sober's defence against objections from evidential holism. A nice feature of **I**-confirmation is that in adaptive proofs the weaker epistemic status of hypotheses inferred from an observation report in conjunction with a set of auxiliaries is reflected by their non-empty condition. Whereas auxiliaries are introduced as premises on the empty condition, inductively generated hypotheses are derived conditionally and may be retracted at a later stage of the proof. For a more fine-grained treatment of background information in adaptive logics for inductive generalization, see [11.5, Sect. 6].

The third objection against HD confirmation dates back to *Hempel's* [11.17], in which he argued that a variant of HD confirmation (which he calls the *prediction criterion* of confirmation) is circular. The problem is that in HD confirmation the hypothesis to be confirmed functions as a premise from which we derive the evidence, and that it is unclear where this premise comes from. The hypothesis is not generated, but given in advance, so HD confirmation presupposes the prior attainment – by inductive reasoning – of a hypothesis. This inductive move, *Hempel* argues, already presupposes the idea of confirmation, making the HD account circular.

The weak step in *Hempel's* argument consists in his assumption that the inductive jump to the original attainment of a hypothesis already presupposes the confirmation of this hypothesis. In testing or generating a hypothesis we need not yet *believe* or *accept* it. Typically, belief and acceptance come only after confirming the hypothesis. Indeed, in probabilistic notions of confirmation the idea is often exactly this: confirming a hypothesis amounts to increasing our degree of belief in it. *Hempel's* circularity objection, it seems, confuses hypothesis generation and hypothesis confirmation.

*Hempel's* circularity objection does not undermine HD confirmation, but it points to the wider scope of the adaptive account as compared to HD confirmation. In an **I**-proof, the conditional rule allows us to *generate* hypotheses. Hypotheses are not given in advance but are computable by the logic itself. Moreover, a clear distinction can be made between hypothesis generation and hypothesis confirmation. Hypotheses generated in an **I**-proof may be derivable at some stage of the proof, but the central question is whether they can be retained – whether they are *finally* derivable. **I**-confirmation, then, amounts to final derivability in an **I**-proof whereas the inductive step of hypothesis generation is represented by retractable applications of RC.

### 11.4.3 Interdependent Abnormalities and Heuristic Guidance

For any of the adaptive logics for inductive generalization defined in this chapter, at most one positive instance is needed to try and derive and, subsequently, confirm a generalization for a given set of premises. This is a feature that **I**-confirmation shares with the other qualitative criteria of confirmation. As a simple illustration, note that an observation report consisting of a single observation  $Pa$  confirms the hypothesis  $\forall xPx$  according to all qualitative criteria discussed in this chapter. Proponents of quantitative approaches to confirmation may object that this is insufficient; that a stronger criterion is needed which requires *more* than one instance for a hypothesis to be confirmed. Against

this view, one can uphold that confirmation is mainly falsification-driven. Rather than confirming hypotheses by heaping up positive instances, we try and test them by searching for negative instances. In the remainder of this section, it is argued by means of a number of examples that **I**-confirmation is sufficiently selective as a criterion for confirming generated hypotheses. The examples moreover allow for the illustration of an additional feature of **I**-confirmation: its use as a heuristic guide for provoking further tests in generating and confirming additional hypotheses.

Simple examples like the one given in the previous paragraph may suggest that, in the absence of falsifying instances, a single instance usually suffices to **I**-confirm a hypothesis. This is far from the truth. Consider the simple premise set  $\Gamma_6 = \{\neg Pa \vee Qa, \neg Qb, Pc\}$ . This premise set contains instances of all of the generalizations  $\forall xPx$ ,  $\forall x\neg Qx$ , and  $\forall x(Px \supset Qx)$ . Not a single one of these is **IL**-confirmed, however, due to the derivability of the following disjunctions of abnormalities

$$!Px \vee !Qx, \quad (11.37)$$

$$!Px \vee !(\neg Px \vee Qx), \quad (11.38)$$

$$!(Px \vee Qx) \vee !(\neg Px \vee Qx), \quad (11.39)$$

$$!Qx \vee !(\neg Px \vee Qx), \quad (11.40)$$

$$!(\neg Px \vee Qx) \vee !(\neg Px \vee \neg Qx). \quad (11.41)$$

Note that  $\Gamma_6$  contains positive instances of both  $\forall xPx$  and  $\forall x\neg Qx$ , so not even a positive instance suffices for a generalization to be finally **IL**-derivable in the absence of falsifying instances. The same is true if we switch from **IL** to **G**. None of  $\forall xPx$ ,  $\forall x\neg Qx$ , or  $\forall x(Px \supset Qx)$  is **G**-confirmed, due to the derivability of the following disjunctions of abnormalities

$$\pm Px \vee \pm Qx, \quad (11.42)$$

$$\pm Px \vee (Px \wedge \pm Qx), \quad (11.43)$$

$$\pm Qx \vee (Qx \wedge \pm Px). \quad (11.44)$$

The reason for the non-confirmation of generalizations like  $\forall xPx$ ,  $\forall x\neg Qx$ , or  $\forall x(Px \supset Qx)$  in this example has to do with the dependencies that exist between abnormalities. Even if a generalization is not falsified by the data, it is often the case that this generalization is not compatible with a different generalization left unfalsified by the data. As a further illustration, consider the premise set  $\Gamma_7 = \{\neg Ra, \neg Ba, Rb\}$ . Again, although no falsifying instance is present, the generalization  $\forall x(Rx \supset Bx)$  is not **IL**-derivable. The reason is the derivability of the following minimal disjunction of abnormalities

$$!(\neg Rx \vee Bx) \vee !(\neg Rx \vee \neg Bx). \quad (11.45)$$

Examples like these illustrate that **I**-confirmation is not too liberal a criterion of confirmation. They also serve to illustrate a different point. Minimal Dab-formulas like (11.45) evoke questions. Which of the two abnormalities is the case? For this particular premise set, establishing which of  $Bb$  or  $\neg Bb$  is the case would settle the matter. For if  $Bb$  were the case, then the second disjunct of (11.45) would be derivable, and (11.45) would no longer be minimal. Consequently, the abnormality  $\exists x(\neg Rx \vee Bx) \wedge \exists x\neg(\neg Rx \vee Bx)$  would no longer be part of a minimal disjunction of abnormalities, and the generalization  $\forall x(Rx \supset Bx)$  would become finally derivable. Analogously, if  $\neg Bb$  were the case, then the first disjunct of (11.45) would become derivable, and, by the same reasoning, the generalization  $\forall x(Rx \supset \neg Bx)$  would become finally derivable. Thus

$$\Gamma_7 \cup \{Bb\} \vdash_{\mathbf{IL}} \forall x(Rx \supset Bx), \quad (11.46)$$

$$\Gamma_7 \cup \{\neg Bb\} \vdash_{\mathbf{IL}} \forall x(Rx \supset \neg Bx). \quad (11.47)$$

Two more comments are in order here. First, this example illustrates that confirming a hypothesis often involves the disconfirmation of the contrary hypothesis. We saw that if we use Hempel's criterion a non-black non-raven confirms the raven hypothesis. But as Goodman pointed out "the prospects for indoor ornithology vanish when we notice that under these same conditions, the contrary hypothesis that no ravens are black is equally well confirmed" [11.21, p. 71]. Thus, according to Goodman, confirming the raven hypothesis  $\forall x(Rx \supset Bx)$  requires disconfirming its contrary  $\forall x(Rx \supset \neg Bx)$ . This is exactly what happens in the example: in order to **IL**-derive  $\forall x(Rx \supset Bx)$ , a falsifying instance for its contrary is needed, as (11.46) illustrates. Goodman's suggestion that the confirmation of a hypothesis requires the falsification/disconfirmation of its contrary was picked up by Israel Scheffler, who developed it further in his [11.31]. Note that falsifying the contrary of the raven hypothesis amounts to finding a positive instance of the raven hypothesis. Thus, in demanding a positive instance before permitting generalization in a **G**-proof, the latter system goes further than **IL** in implementing Goodman's idea. As we saw, however, not even **G** goes all the way: a generalization may be **G**-derivable even in the absence of a positive instance.

Second, if empirical (observational or experimental) means are available to answer questions like  $\{Bb, \neg Bb\}$  in the foregoing example, these questions may be called *tests* [11.2]. Adaptive logics for inductive generalization provide heuristic guidance in the sense that interdependencies between abnormalities evoke such tests. Importantly, further tests may lead to

the derivability of new generalizations. In the example, deciding the question  $\{Bb, \neg Bb\}$  in favor of  $Bb$  leads to the confirmation of  $\forall x(Rx \supset Bx)$  and to the disconfirmation of  $\forall x(Rx \supset \neg Bx)$ , while deciding it in favor of  $\neg Bb$  leads to the confirmation of  $\forall x(Rx \supset \neg Bx)$  and to the disconfirmation of  $\forall x(Rx \supset Bx)$ . This is an important practical advantage of **I**-confirmation over other qualitative criteria: adaptive logics for inductive generalization evoke tests for increasing the number of confirmed generalizations.

The illustrations so far may suggest that this heuristic guidance provided by **I**-confirmation only applies to hypotheses that are logically related or closely connected, like the raven hypothesis and its contrary. But the point is more general, as the following example illustrates.

Consider the premise set

$$\Gamma_8 = \{Pa, Qa, \neg Ra, \neg Pb, \\ \neg Qb, Rb, Pc, Rc, Qd, \neg Pe\}.$$

Despite the fact that  $\Gamma_8$  contains positive instances of the generalizations  $\forall x(Px \supset Qx)$  and  $\forall x(Rx \supset \neg Qx)$ , and despite the fact that these generalizations are not falsified by  $\Gamma_8$ , none of them is **IL**-derivable due to the derivability of the disjunction

$$!(\neg Px \vee Qx) \vee !(\neg Rx \vee \neg Qx). \quad (11.48)$$

## 11.5 Conclusions

A number of adaptive logics for inductive generalization were presented each of which, it was argued, can be re-interpreted as a criterion of confirmation. The logics in question can be classified along two dimensions. The first dimension concerns when it is permitted to introduce a generalization in an adaptive proof. The logic **LI** permits the free introduction of generalizations. **IL** and **G** require instances of a generalization before introducing it in a proof. Interestingly, these stronger requirements do not result in stronger logics.

The second dimension along which the logics defined in this chapter can be classified concerns their

By the same reasoning as in the previous illustration,  $\Gamma_8$  evokes the question  $\{Qc, \neg Qc\}$ . If this question is a test (if it can be answered by empirical means), the answer will confirm one of the generalizations  $\forall x(Px \supset Qx)$  and  $\forall x(Rx \supset \neg Qx)$ , and will disconfirm the other generalization [11.2].

The example generalizes. In **LI** and **G** too, the derivability of  $\forall x(Px \supset Qx)$  and  $\forall x(Rx \supset \neg Qx)$  is blocked due to the **CL**-derivability of the **LI**-minimal Dab-formula (11.49), respectively the **G**-minimal Dab-formula (11.50)

$$\neg \forall x(Px \supset Qx) \vee \neg \forall x(Rx \supset \neg Qx), \quad (11.49)$$

$$(Px \wedge \pm Qx) \vee (Rx \wedge \pm \neg Qx). \quad (11.50)$$

Here too, deciding the question  $\{Qc, \neg Qc\}$  resolves the matter. Thus, where  $\mathbf{I} \in \{\mathbf{LI}, \mathbf{IL}, \mathbf{G}\}$

$$\Gamma_8 \not\vdash_{\mathbf{I}} \forall x(Px \supset Qx), \quad (11.51)$$

$$\Gamma_8 \not\vdash_{\mathbf{I}} \forall x(Rx \supset \neg Qx), \quad (11.52)$$

$$\Gamma_8 \cup \{Qc\} \vdash_{\mathbf{I}} \forall x(Px \supset Qx), \quad (11.53)$$

$$\Gamma_8 \cup \{Qc\} \not\vdash_{\mathbf{I}} \forall x(Rx \supset \neg Qx), \quad (11.54)$$

$$\Gamma_8 \cup \{\neg Qc\} \not\vdash_{\mathbf{I}} \forall x(Px \supset Qx), \quad (11.55)$$

$$\Gamma_8 \cup \{\neg Qc\} \vdash_{\mathbf{I}} \forall x(Rx \supset \neg Qx). \quad (11.56)$$

For some concrete heuristic rules applicable to the logic **LI**, see [11.3].

adaptive strategy. Here, no surprises arise. A logic defined using the reliability strategy is in general weaker than its counterpart logic defined using the minimal abnormality strategy (this was shown to be the case for all adaptive logics defined within the standard format [11.7, Theorem 11]).

When re-interpreted as criteria of confirmation, the logics defined here withstand the comparison with their main rivals, i. e., Hempel's satisfaction criterion and the hypothetico-deductive model of confirmation. In conclusion, the adaptive confirmation criteria defined in Sect. 11.4 offer an interesting alternative perspective on (qualitative) confirmation theory.

## 11.A Appendix: Blocking the Raven Paradox?

If a formalism defined in terms of **CL** behaves overly permissive, a good strategy to remedy this problem is to add further criteria of validity or relevance. For instance, in order to avoid problems of irrelevant conjunctions and disjunctions, hypothetico-deductivists may impose further demands on HD confirmation [11.32–35].

A similar strategy could be adopted with respect to **I**-confirmation and the raven paradox. In this appendix, an alternative adaptive logic of induction, **IC**, is defined, as is a corresponding criterion of confirmation which is slightly less permissive than the criteria from Sect. 11.4. **IC** makes use of a non-classical conditional resembling a number of conditionals originally defined in order to avoid the so-called paradoxes of material implication. First, an extension of **CL** is introduced, including this new conditional connective. Next, the adaptive logic **IC** is defined.

The new conditional,  $\rightarrow$ , is fully characterized by the following rules and axiom schema's

$$\frac{A, (A \rightarrow B)}{B}, \quad (\text{MP})$$

$$\frac{A \equiv B}{(A \rightarrow C) \equiv (B \rightarrow C)}, \quad (\text{RCEA})$$

$$\frac{A \equiv B}{(C \rightarrow A) \equiv (C \rightarrow B)}, \quad (\text{RCEC})$$

$$(A \rightarrow (B \wedge C)) \equiv ((A \rightarrow B) \wedge (A \rightarrow C)), \quad (\text{D}\wedge)$$

$$((A \vee B) \rightarrow C) \equiv ((A \rightarrow C) \wedge (B \rightarrow C)), \quad (\text{D}\vee)$$

((RCEA), (RCEC), and (D $\wedge$ ) fully characterize the conditional of Chellas's logic **CR** from [11.36]. The latter was also used for capturing explanatory conditionals in [11.37]. See also [11.38, Chap. 5] for some closely related conditional logics, including an extension of Chellas's systems that validates (MP).)

Let  $\text{CL}^{\rightarrow}$  be the logic resulting from adding  $\rightarrow$  to the language of **CL**, and from adding (MP)-(D $\vee$ ) to the list of rules and axioms of **CL**. Note that the conditional  $\rightarrow$  is strictly stronger than  $\supset$

$$(A \rightarrow B) \supset (A \supset B). \quad (11.57)$$

(By (MP),  $A, (A \rightarrow B) \vdash_{\text{CL}^{\rightarrow}} B$ . By the deduction theorem for  $\supset$ ,  $A \rightarrow B \vdash_{\text{CL}^{\rightarrow}} A \supset B$ . By the deduction theorem again,  $\vdash_{\text{CL}^{\rightarrow}} (A \rightarrow B) \supset (A \supset B)$ .)

In view of this bridging principle between both conditionals it is easily seen that counter-instances to a formula of the form  $\forall x(A(x) \supset B(x))$  form counter-instances to  $\forall x(A(x) \rightarrow B(x))$ , and falsify the latter formula as well. For instance, if  $Pa \wedge \neg Qa$ , then, by **CL**,  $\neg \forall x(Px \supset Qx)$ , and, by (11.57),  $\neg \forall x(Px \rightarrow Qx)$ .

The adaptive logic **IC** is fully characterized by the lower limit logic  $\text{CL}^{\rightarrow}$ , the set of abnormalities

$$\begin{aligned} \Omega_{\text{IC}} =_{\text{df}} & \{ \exists (A_1 \wedge \dots \wedge A_n \wedge A_0) \\ & \wedge \neg \forall ((A_1 \wedge \dots \wedge A_n) \rightarrow A_0) \mid \\ & A_0, A_1, \dots, A_n \in \mathcal{A}^{\uparrow}; n \geq 0 \}, \end{aligned} \quad (11.58)$$

and the adaptive strategy reliability (**IC**<sup>r</sup>) or minimal abnormality (**IC**<sup>m</sup>). **IC** is defined within the SF. All rules and definitions for its proof theory are as for the other logics defined in this chapter, except that in the definition of RU and RC, **CL** is replaced with  $\text{CL}^{\rightarrow}$ .

The following proof illustrates how formulas are derived conditionally in **IC**

1	$\neg Ra$	Prem
	$\emptyset$	
2	$\neg Ba$	Prem
	$\emptyset$	
3	$\forall x(\neg Bx \rightarrow \neg Rx)$	1, 2; RC
	$\{ \exists x(\neg Bx \wedge \neg Rx) \wedge \neg \forall x(\neg Bx \rightarrow \neg Rx) \}$	

Given only the premises  $\neg Ra$  and  $\neg Ba$ , there is no possible extension of this proof in which line 3 gets marked. Hence

$$\neg Ra, \neg Ba \vdash_{\text{IC}} \forall x(\neg Bx \rightarrow \neg Rx). \quad (11.59)$$

However, contraposition is invalid for the new conditional  $\rightarrow$ , hence we cannot derive the raven hypothesis from the formula derived at line 3. Note also that, in view of (11.60), we cannot use the conditional rule RC to derive  $\forall x(Rx \rightarrow Bx)$  on the condition  $\{ \exists x(Rx \wedge Bx) \wedge \neg \forall x(Rx \rightarrow Bx) \}$  in an **IC**-proof, since

$$\begin{aligned} \neg Ra, \neg Ba \not\vdash_{\text{CL}^{\rightarrow}} \forall x(Rx \rightarrow Bx) \\ \vee \{ \exists x(Rx \wedge Bx) \wedge \neg \forall x(Rx \rightarrow Bx) \}. \end{aligned} \quad (11.60)$$

Therefore

$$\neg Ra, \neg Ba \not\vdash_{\text{IC}} \forall x(Rx \rightarrow Bx). \quad (11.61)$$

Thus, if conditional statements of the form *for all x, if A(x) then B(x)* are taken to be **IC**-confirmed only if the conditional in question is an arrow ( $\rightarrow$ ) instead of a material implication, then the raven paradox, in its original formulation, is blocked.

An additional property of **IC** is that *strengthening the antecedent* fails for  $\rightarrow$ . In Sect. 11.3, for instance, we saw that

$$Pa \vdash_{\text{G}} \forall x(Qx \supset Px). \quad (11.62)$$

In **IC**, (11.62) still holds for the material implication, but not for the new conditional. In an **IC**-proof from  $Pa$  we can still derive  $\forall xPx$  on the condition  $\{\exists xPx \wedge \exists x\neg Px\}$ , and since **IC** extends **CL** it still follows that  $\forall x(Px \supset Qx)$

$$Pa \vdash_{\mathbf{IC}} \forall xPx, \tag{11.63}$$

$$Pa \vdash_{\mathbf{IC}} \forall x(Qx \supset Px). \tag{11.64}$$

However, since  $\forall xPx \not\vdash_{\mathbf{CL}} \forall x(Qx \rightarrow Px)$ , and since we do not have any further means to conditionally derive the formula  $\forall x(Qx \rightarrow Px)$  in an **IC**-proof

$$Pa \not\vdash_{\mathbf{IC}} \forall x(Qx \rightarrow Px). \tag{11.65}$$

Originally, the logics in the **G**-family were constructed as logics requiring a *positive instance* before we are allowed to apply RC. This is reflected in the definition of the set of **G**-abnormalities. In order to derive a formula like  $\forall x(Px \supset Qx)$  on its corresponding condition, a positive instance, e.g.,  $Pa \wedge Qa$ , is needed. Examples like (11.36) and (11.62) show, however, that such a positive instance is not always required in order to **G**-derive a generalization. The logic **IC**, it seems, does much better in this respect. However, it still does not fully live up to the requirement for a positive instance before generalizing, as the following **IC**-proof from  $\Gamma_9 = \{\neg Ra \wedge \neg Ba, Rb, Bc\}$  illustrates (where  $A_0, A_1, \dots, A_n \in \mathcal{A}^I$ ,  $\dagger((A_1 \wedge \dots \wedge A_n) \rightarrow A_0)$  abbreviates  $\exists(A_1 \wedge \dots \wedge A_n \wedge A_0) \wedge \neg \forall((A_1 \wedge \dots \wedge A_n) \rightarrow A_0)$ ).

1	$\neg Ra \wedge \neg Ba$	Prem
	$\emptyset$	
2	$Rb$	Prem
	$\emptyset$	
3	$Bc$	Prem
	$\emptyset$	

**References**

11.1 J. Norton: A little survey of induction. In: *Scientific Evidence*, ed. by P. Achinstein (John Hopkins Univ. Press, Baltimore 2005) pp. 9–34

11.2 D. Batens: The basic inductive schema, inductive truisms, and the research-guiding capacities of the logic of inductive generalization, *Logique et Analyse* **185–188**, appeared **2005**, 53–84 (2004)

11.3 D. Batens: On a logic of induction, *Log. Philos. Sci.* **4(1)**, 3–32 (2006)

11.4 D. Batens, L. Haesaert: On classical adaptive logics of induction, *Logique et Analyse* **173–175**, appeared

4	$\forall x(\neg Bx \rightarrow \neg Rx)$ $\{\dagger(\neg Bx \rightarrow \neg Rx)\}$	1; RC
5	$Bb$ $\{\dagger(\neg Bx \rightarrow \neg Rx)\}$	2, 4; RU
6	$\forall x(Rx \rightarrow Bx)$ $\{\dagger(Rx \rightarrow Bx), \dagger(\neg Bx \rightarrow \neg Rx)\}$	2, 5; RC

The key step in this proof is the derivation of  $Bb$  at line 5, which together with  $Rb$  provides us with a positive instance of the raven hypothesis.  $Bb$  is derivable from lines 2 and 4 in view of **CL** and (11.57). Except for the formulas  $\exists xRx \wedge \exists x\neg Rx$  and  $\exists xBx \wedge \exists x\neg Bx$ , no minimal Dab-formulas are **CL** $\rightarrow$ -derivable from  $\Gamma_9$ . Therefore

$$\Gamma_9 \vdash_{\mathbf{IC}} \forall x(Rx \rightarrow Bx). \tag{11.66}$$

As (11.61) illustrates the logic **IC** avoids the raven paradox in its original formulation. A possible drawback of **IC** is that it does not fully meet the demand for a positive instance when confirming a hypothesis (Sect. 11.4.3). It is left open whether it is possible and desirable to further extend **IC** so as to fully meet this demand.

**Acknowledgments.** The author is greatly indebted to Atocha Aliseda, Cristina Barés-Gómez, Diderik Batens, Matthieu Fontaine, Jan Sprenger, and Frederik Van De Putte for insightful and valuable comments on previous drafts of this chapter. Research for this article was supported by the *Programa de Becas Posdoctorales de la Coordinación de Humanidades* of the National Autonomous University of Mexico (UNAM), by the project *Logics of discovery, heuristics and creativity in the sciences* (PAPIIT, IN400514-3) granted by the UNAM, and by a Sofja Kovalevskaja award of the Alexander von Humboldt-Foundation, founded by the German Ministry for Education and Research.

**2003**, 255–290 (2001)

11.5 D. Batens: Logics for qualitative inductive generalization, *Studia Logica* **97**, 61–80 (2011)

11.6 C.G. Hempel: A purely syntactical definition of confirmation, *J. Symb. Log.* **8(4)**, 122–143 (1943)

11.7 D. Batens: A universal logic approach to adaptive logics, *Logica Universalis* **1**, 221–242 (2007)

11.8 D. Batens: Tutorial on inconsistency-adaptive logics. In: *New Directions in Paraconsistent Logic: 5th WCP, Kolkata, India, February 2014*, Springer Proceedings in Mathematics and Statistics, Vol. 152, ed.

- by J.-Y. Beziau, M. Chakraborty, S. Dutta (Springer India, New Delhi 2015) pp. 3–38
- 11.9 D. Batens: Towards a dialogic interpretation of dynamic proofs. In: *Dialogues, Logics and Other Strange Things. Essays in Honour of Shahid Rahman*, ed. by C. Dégreumont, L. Keiff, H. Rückert (College Publications, London 2009) pp. 27–51
- 11.10 F. Van De Putte, C. Straßer: Adaptive logics: A parametric approach, *Log. J. IGPL* **22**(6), 905–932 (2014)
- 11.11 K. Popper: *The Logic of Scientific Discovery* (Hutchinson, London 1959)
- 11.12 R. Carnap: *Logical Foundations of Probability* (Univ. of Chicago Press, Chicago 1950)
- 11.13 B. Fitelson: Inductive logic. In: *Philosophy of Science: An Encyclopedia*, ed. by J. Pfeifer, S. Sarkar (Routledge, London 2005) pp. 384–394
- 11.14 A. Hájek, N. Hall: Induction and probability. In: *The Blackwell Guide to the Philosophy of Science*, ed. by P. Machamer, M. Silberstein (Blackwell, Oxford 2002) pp. 149–172
- 11.15 R. Jeffrey: *The Logic of Decision*, 2nd edn. (Univ. of Chicago Press, Chicago 1990)
- 11.16 B. Skyrms: *Choice and Chance. An Introduction to Inductive Logic*, 3rd edn. (Wadsworth Publishing Company, Belmont 1986)
- 11.17 C.G. Hempel: Studies in the logic of confirmation II, *Mind* **54**(214), 97–121 (1945)
- 11.18 J. Sprenger: A synthesis of Hempelian and hypothetico-deductive confirmation, *Erkenntnis* **78**(4), 727–738 (2013)
- 11.19 B. Fitelson, J. Hawthorne: How Bayesian confirmation theory handles the paradox of the ravens. In: *The Place of Probability in Science*, ed. by E. Eells, J. Fetzer (Springer, Heidelberg 2010) pp. 247–275
- 11.20 J. Sprenger: Hypothetico-deductive confirmation, *Philos. Compass* **6**17, 497–508 (2011)
- 11.21 N. Goodman: *Fact, Fiction, and Forecast* (Harvard Univ. Press, Cambridge 1955)
- 11.22 D. Stalker (Ed.): *Grue! The New Riddle of Induction* (Open Court, Chicago 1994)
- 11.23 C.G. Hempel: Studies in the logic of confirmation I, *Mind* **54**(213), 1–26 (1945)
- 11.24 J. Sprenger: Hempel and the paradoxes of confirmation. In: *Handbook of the History of Logic*, Vol. 10, ed. by D. Gabbay, S. Hartmann, J. Woods (Elsevier, Amsterdam 2011) pp. 231–260
- 11.25 V. Crupi: Confirmation. In: *The Stanford Encyclopedia of Philosophy*, Spring 2014 edn., ed. by Edward N. Zalta, <http://plato.stanford.edu/archives/spr2014/entries/confirmation/> (2014)
- 11.26 H. Putnam: Craig's theorem, *J. Philos.* **62**(10), 251–260 (1965)
- 11.27 C. Glymour: *Theory and Evidence* (Princeton Univ. Press, Princeton 1980)
- 11.28 P. Duhem: *The Aim and Structure of Physical Theory* (Princeton Univ. Press, Princeton 1991), first published 1906
- 11.29 N.W.V. Quine: Two dogmas of empiricism, *Philos. Rev.* **60**, 20–43 (1951)
- 11.30 E. Sober: Testability, *Proc. Addresses Am. Philos. Assoc.* **73**(2), 47–76 (1999)
- 11.31 I. Scheffler: *The Anatomy of Inquiry* (Knopf, New York 1963)
- 11.32 K. Gemes: Hypothetico-deductivism, content, and the natural axiomatization of theories, *Philos. Sci.* **60**(3), 477–487 (1993)
- 11.33 K. Gemes: Hypothetico-deductivism: The current state of play, *Erkenntnis* **49**, 1–20 (1998)
- 11.34 G. Schurz: Relevant deduction, *Erkenntnis* **35**, 391–437 (1991)
- 11.35 G. Schurz: Relevant deduction and hypothetico-deductivism: A reply to Gemes, *Erkenntnis* **41**, 183–188 (1994)
- 11.36 B. Chellas: Basic conditional logic, *J. Philos. Log.* **4**, 133–153 (1975)
- 11.37 M. Beirlaen, A. Aliseda: A conditional logic for abduction, *Synthese* **191**(15), 3733–3758 (2014)
- 11.38 G. Priest: *An Introduction to Non-Classical Logic*, 2nd edn. (Cambridge Univ. Press, Cambridge 2008)

# Modeling Hypothetical Reasoning by Formal Logics

Tjerk Gauderis

In this chapter, it is discussed to which extent hypothetical reasoning can be modeled by formal logics. It starts by exploring this idea in general (Sects. 12.1 and 12.2), which leads to the conclusion that in order to model this kind of reasoning formally, a more fine-grained classification of reasoning patterns should be in order. After such a classification is provided in Sect. 12.3, a formal framework that has proven successful to capture some of these patterns is described (Sects. 12.4 and 12.6) and some of the specific problems for this procedure are discussed (Sect. 12.5). The chapter concludes by presenting two logics for hypothetical reasoning in an informal way (Sects. 12.7 and 12.8) such that the nontechnically skilled reader can get a flavor of how formal methods can be used to describe hypothetical reasoning.

12.1	<b>The Feasibility of the Project</b> .....	249
12.2	<b>Advantages and Drawbacks</b> .....	251
12.3	<b>Four Patterns of Hypothetical Reasoning</b> .....	252
12.4	<b>Abductive Reasoning and Adaptive Logics</b> .....	255
12.5	<b>The Problem of Multiple Explanatory Hypotheses</b> ..	256
12.6	<b>The Standard Format of Adaptive Logics</b> .....	256
12.6.1	Dynamic Proof Theory .....	257
12.7	<b>LA<sub>s</sub><sup>†</sup>: A Logic for Practical Singular Fact Abduction</b>	258
12.7.1	Lower Limit Logic .....	258
12.7.2	Set of Abnormalities $\Omega$ .....	258
12.7.3	Reliability Strategy .....	259
12.7.4	Practical Abduction .....	260
12.7.5	Avoiding Random Hypotheses .....	260
12.8	<b>MLA<sub>s</sub><sup>‡</sup>: A Logic for Theoretical Singular Fact Abduction</b> .....	261
12.8.1	Formal Language Schema .....	261
12.8.2	Lower Limit Logic .....	261
12.8.3	Intended Interpretation of the Modal Operators .....	261
12.8.4	Set of Abnormalities .....	261
12.8.5	First Proposal $\Omega_1$ .....	262
12.8.6	Simple Strategy .....	262
12.8.7	Contradictory Hypotheses .....	262
12.8.8	Predictions and Evidence .....	262
12.8.9	Contradictions .....	263
12.8.10	Tautologies .....	263
12.8.11	Second Proposal $\Omega_2$ .....	263
12.8.12	Most Parsimonious Explanantia .....	263
12.8.13	Notation .....	264
12.8.14	Final Proposal $\Omega$ .....	264
12.9	<b>Conclusions</b> .....	265
12.A	<b>Appendix: Formal Presentations of the Logics LA<sub>s</sub><sup>†</sup> and MLA<sub>s</sub><sup>‡</sup></b> .....	265
12.A.1	Proof Theory .....	265
12.A.2	Semantics .....	266
	<b>References</b> .....	267

## 12.1 The Feasibility of the Project

To an outsider, the claim that hypothetical or abductive reasoning, that is, the act of *forming* and *suggesting* hypotheses for certain observations or puzzling facts, can be modeled by means of formal logics might sound as outlandish as claiming that computers have the same cognitive and creative abilities as humans. After all, abductive or hypothetical reasoning – by which we always mean the reasoning *toward* (explanatory) hypotheses, not *starting from* certain (possibly counterfactual) hy-

potheses [12.1] – is often considered to be the hallmark of creative ingenuity, leading to our rich and wide diversity of ideas, innovations, and scientific theories. It just seems impossible that this richness can be reconstructed or created by just using formal tools, which are by nature abstracted from the specific semantic content. This argument, in short the *creativity excludes logic* argument, is the main reason why even the field itself is sharply divided between believers and nonbelievers.

This argument, however, is a straw man. Nobody would argue for the claim that hypothetical reasoning can be modeled by means of formal logics along these lines. What is argued in this chapter and the various sources it cites is the more modest claim that certain aspects and forms of hypothetical reasoning can be modeled with the aid of formal systems that are specifically suited for this task.

There are three important ways in which this modest claim differs from the straw man that is attacked by the *creativity excludes logic* argument:

1. Abductive reasoning is not a monolithic concept: it does not consist of a single method or procedure, but consists of many different patterns; formal logics are only used to capture one specific and precisely defined pattern at a time.
2. The logics that are used for this goal are not necessarily classical or deductive.
3. The relation between formal logics and abductive reasoning is not that of an agent and an activity (i. e., formal logics do not display themselves abductive reasoning like humans do) but that of a model and a target: Formal logics are used by (human) agents to model and – to a certain extent – to simulate certain aspects of human abductive reasoning.

The semantic content that is lacking in abductive reasoning is provided by these agents.

In the remainder of this introduction, the types of logics that are suitable for modeling abductive or hypothetical reasoning are discussed in further detail and the framework that is used for the logics in this chapter is introduced in general terms.

For those who are still a bit suspicious how abductive or hypothetical reasoning patterns can be modeled using formal logics, it needs to be stressed that it is not meant that any of these patterns is a valid inference in classical logic (CL) or any other (nontrivial) deductive logic. To model defeasible reasoning steps such as hypothesis formation, one has to use nonmonotonic logics: logics for which an extension of a premise set does not always yield a consequence set that is a superset of the original consequence set. Or, put more simply, logics according to which new information may lead to revoke old conclusions.

It is important to note that the purpose in using logics for this task is not the classical purpose of the discipline of logic. Classically, the discipline of logic studies the correct way to infer further knowledge from already known facts. The correct way should guarantee the truth of the new facts, under the supposition that the old facts are true. Accordingly, this has motivated the search for the right (deductive) logic (whether it be classical logic or another one such as intuitionistic logic).

The purpose here, however, is to model or explicate human reasoning patterns. As these patterns are fallible, leading to conclusions that are not necessarily true even if the premises are assumed to be true, it should be possible to revoke previously derived results; hence, the use of nonmonotonic logics. Also, because there are many patterns of human reasoning, it is natural to conceive of a plenitude of logics in order to describe them.

Let this be explained a bit more formally. A logic can be considered as a function from the power set of the sentences of a language to itself. So, given a language  $L$  and the set  $\mathcal{W}$  of its well-formed formulas

$$\mathbf{L} : \wp(\mathcal{W}) \rightarrow \wp(\mathcal{W}) . \quad (12.1)$$

Hence, a logic determines for every set of sentences (or premise set)  $\Gamma$  the sentences of which can be inferred from it ( $Cn_{\mathbf{L}}(\Gamma) =_{\text{df}} \mathbf{L}(\Gamma)$ ). Therefore, as a reasoning pattern is nothing more than the inference of some statements given some initial statements, in principle, a logic can be devised to model any reasoning pattern in science. If this pattern can be formally described, description by a formal logic is, in principle, possible. It has to be added, though, that in reality, scientific and human reasonings include not only sentences or propositions, but also direct observations, sketches, and various other symbolic representations. Yet, for the purpose of modeling particular reasoning patterns, those sources can be generally represented by suitable propositions.

Deductive logics, such as CL, have the property of monotonicity, that is, for all premise sets  $\Gamma$  and  $\Gamma'$

$$Cn_{\mathbf{L}}(\Gamma) \subseteq Cn_{\mathbf{L}}(\Gamma \cup \Gamma') . \quad (12.2)$$

Most patterns of human reasoning, however, do not meet this criterion. For instance, if an agent infers a hypothesis, s/he is well aware of the fact that it might need to be revoked on closer consideration of the available background knowledge or in light of new information.

Although nonmonotonic reasoning has typically received less attention in the field of logic than monotonic reasoning, various frameworks for defeasible reasoning and nonmonotonic logics are available such as default logic, adaptive logics, and belief revision (see [12.2] for a general overview of the variation in approaches). In this chapter, the progress that has been made on modeling abduction within the adaptive logics framework is overviewed. This is a framework created by *Batens* over the past three decades (see [12.3] or [12.4] for an extensive overview and thorough formal introduction). This framework for devising nonmonotonic logics has some advantages that suit very well the project of modeling abductive reasoning patterns.



First, the focus in the adaptive logic program is, in contrast with other approaches to nonmonotonic reasoning, on proof theory. For these logics, a dynamic proof style has been defined in order to mimic to a certain extent actual human reasoning patterns. More in particular, these dynamic proofs display the two forms of revoking previously derived results that can also be found in human reasoning: revoking old conclusions on closer consideration of the available evidence (internal dynamics) and revoking them in light of new information (external dynamics). One should not be misled, however, by this idea of dynamic proofs in thinking that the consequence set of adaptive logics for a certain premise set depends on the proof. Adaptive logics are proper proof-invariant logics that assign for each premise set  $\Gamma$  exactly one consequence set  $Cn_L(\Gamma)$ .

Second, over the years, a solid meta-theory has been built for this framework, which guarantees that if

an adaptive logic is created according to certain standards (the so-called *standard format*), many important metatheoretical properties are generically proven. This creates an opportunity for projects such as this to focus almost exclusively on the application of these formal methods without having to worry too much about proving their metatheoretical characteristics.

Finally, as the framework is presented as a unified framework for nonmonotonic logics, it has been applied in many different contexts. Over the years, adaptive logics have been devised for, apart from abduction, paraconsistent reasoning, induction, argumentation, deontic reasoning, etc. Most of these applications have been studied at the Centre for Logic and Philosophy of Science (Ghent University). At this center's website, many references can be found to papers in various contexts. The reference works mentioned earlier, [12.3] and [12.5], also give a good overview of the various applications.

## 12.2 Advantages and Drawbacks

Explicating patterns of hypothesis formation by means of formal logics has a clear advantage: By reducing patterns to their formal and structural essence, an insight into the pattern's precise conditions and applications is gained that is hard to achieve purely by studying different cases.

Another great advantage of the formal explication of human reasoning patterns is that it allows for the possibility of providing artificially intelligent agents (which, in general, lack the human capacity for context awareness unless it is explicitly provided) with formal patterns to simulate human reasoning. In the case of hypothesis formation, this possibility has presently already found applications in the artificial intelligence subfields of abduction (diagnosis), planning, and machine learning (see [12.6] for an overview).

The method of explicating patterns of hypothesis formation by means of formal adaptive logics also has certain drawbacks, however.

First, formal logics are expressed in terms of a formal language, in which not all elements of human reasoning processes can be represented. This leads inevitably to certain losses. A very obvious example is that in general only propositions can be represented in logics. It means that all observations, figures, or other symbolic representations must be reduced to descriptions of them. A more important example in the case of abduction is the implication relation. The adaptive logics framework that is used in this chapter is, certainly for ampliative logics such as those for abduction

or induction, largely built around the use of a classical material implication (mostly to keep things sufficiently simple). As a result of this, all relations between a hypothesis and the observations that led to their formation (their triggers) are modeled by material implications. It is clear that this is a strong reduction of the actual richness of such relations. Hypotheses do not have to imply their triggers: they can also just be correlated with them or be probabilistically likely; or the relation can be much more specific, as in the case of an explanatory or causal relation. This issue is relevant beyond the field of adaptive logics. *Paul* [12.6, p. 36] has claimed that most approaches to abduction use a material implication that is implicitly interpreted as some kind of explanatory or causal relation. See also [12.7] for an attempt to better capture the explanatory conditional.

Second, if one sets out to model actual historical human reasoning processes by means of dynamic logical proofs (as the adaptive framework allows us to do), one quickly finds that it is not an easy task to boil down those actual processes to the microstructure of their individual reasoning steps. As human agents often combine individual steps and seldom take note of each individual step, these types of models always contain an aspect of simulation.

Human reasoning also does not proceed linearly step by step as proofs do: It contains circular motions, off-topic deviations, and irrational connections that cannot be captured by formal logics. Therefore, models of

such reasoning processes are always, to a great extent, idealized.

Natural languages are also immensely more complex than any formal language can aspire to be. Therefore, models of human reasoning are unavoidably simplifications. Furthermore, as formal logics state everything explicitly, any modeler of human reasoning has to simplify deliberately the actual cases, only to achieve a certain degree of comprehensibility.

Altogether, it is clear that formal models of human reasoning processes are, in fact, only models: They contain abstractions, simulations, simplifications, and idealizations. And although these techniques are the key characteristics of models, such as those used in science, it is not always easy to evade the criticism that formal logics can only handle toy examples.

Third, certain patterns of creative hypothesis formation, that is, those that introduce the hypothetical existence of new concepts, cannot be modeled by first-order logics. They seem to require at least the use of second-order logics, and this is a possibility of which, at present, the adaptive logics framework is not capable.

Fourth, as one is here purely concerned with hypothesis formation and not with hypothesis selection, formal methods will generate sets of possible hypotheses that may grow exponentially in relation to the

growth of the agent's background knowledge. It is clear that this also poses a limit to the application of these methods to real-world problems.

Finally, one might question the normativity of this project (and more generally of the adaptive logics program). By aiming to describe actual human reasoning processes, this branch of logics appears to put a descriptive ideal first, which contrasts sharply with the strongly normative ideals in the field of logic in general. The standard answer to this question is that adaptive logics attempt to provide both: On the one hand, they aim to describe actual reasoning patterns; on the other, once these patterns are identified, they aim to prescribe how these patterns should be rationally applied. Yet, this does not answer how the trade-off between these two goals of description and normativity should be conceived. Is it better to have a large set of logics that is able to describe virtually any pattern actually found in human reasoning, or should one keep this set trimmed and qualify most actual human reasoning as failing to accord with the highest normative standards? Therefore, it remains a legitimate criticism that the goals of description and prescription cannot be so easily joined: how their trade-off should be dealt with needs further theoretical underpinning.

## 12.3 Four Patterns of Hypothetical Reasoning

The quest to characterize abduction under a single schema was abandoned around 1980. The main reasons were that such attempts (*Hanson's* [12.8, 9] proposal to call abduction *the logic of discovery*) often did not provide much detailed guidance for actual discovery processes, and that even these general attempts always captured only a part of the discovery process (e.g., *inference to the best explanation*, which was first emphasized by *Harman* [12.10], describes only the selection of hypotheses, not their formation).

Around the same time, research from different fields such as the philosophy of science based on historical cases, artificial intelligence, and cognitive science resulted in a new consensus that there is a plenitude of patterns, heuristics, and methods of discovery, which are open to normative guidance; yet, this guidance might be content-, subject-, or context dependent [12.11, 12].

Various authors in the literature on abduction have tried to provide classifications of various patterns of abductions [12.13–15]. For an overview of several logic-based characterizations of abductions, see also the introductory chapter of this section. Although these

attempts differ slightly, some general patterns clearly stand out. In the following, the author's personal interpretation of these major general patterns is given. The main reason this deviates from the previous classifications is that an attempt is made to simplify the rather prolific classifications, yet to provide a sufficient basis for formal modeling. This is possible because it is not attempted to give a fully exhaustive list or a list the elements of which are mutually exclusive. The only purpose was to give a simple list as a basis that covers most instances of abductive reasoning and can serve as the basis of formal modeling.

Before the classification of these major patterns found in abductive reasoning is given, it is important to note that abductive inferences form explanatory hypotheses for observed facts using the agent's background beliefs (or knowledge). Therefore, these patterns have the structure of the inference of a hypothesis (HYP) from some observed facts (OBS) and a part of the agent's background beliefs (or knowledge) (BBK). These latter are, apart from toy examples, typically more than a few factual statements and often encompass a whole explanatory framework of (shared) assump-

tions and knowledge that provides the explanatory link between hypothesis and observations (see [12.16] for an elaborate discussion of the role of an explanatory framework in logical approaches to explanation).

In line with the Fregean tradition, *factual statements* are considered as statements of a *concept* with regard to one or more *objects* (or a logical combination of such statements). For instance, the statement *there was a civil war in France in 1789* can be analyzed as the concept *a country in civil war* applied to or *with regard to* the object *France in 1789*. A *fact* is a true factual statement. As such, concepts can also be considered as the *class* of all objects (or tuples of objects) for which the concept with regard to that object (or tuple of objects) is a fact. An *observed fact* is a factual statement describing an agent's observation that she considers to be true. This can be broadly conceived to include also, for instance, a graph or a table of measurements in a paper. Together, the observed facts form the *trigger* for the agent.

In this semiformal description of these patterns, that  $p$  should be considered as a hypothesis is expressed by using a formulation of the form *it might be that p*; beliefs and observed facts can be expressed simply by stating their content. Concepts such as *a country in civil war* or *a bipedal hominid* are denoted by uppercase letters (typically  $F$  for observed, factual concepts and  $E$  for explanatory concepts) and objects such as *France in 1789* or *Lucy* by lowercase letters such as  $x$  or  $y$ . A finite set or list of (related) objects or concepts can then be expressed, for example, by  $x_1, \dots, x_n$  or  $F_1, \dots, F_n$  where generally  $n \geq 1$  (hence, including the possibility of a single object or concept; the other case is indicated by  $n \geq 2$ ). Finally, that a concept applies to certain objects will be indicated by the phrase *with regard to*:

1. *Abduction of a singular fact*

(OBS)	$F$ with regard to $x_1, \dots, x_n$ ( $n \geq 1$ )
(BBK)	$E$ with regard to $x_1, \dots, x_n$ explains $F$ with regard to those objects in a certain explanatory framework EF.
(HYP) It might be that $E$ with regard to $x_1, \dots, x_n$	

Some examples of this pattern, which has also been called *simple abduction* by *Thagard* [12.13], *factual abduction* by *Schurz* [12.14], and *selective fact abduction* by *Hoffmann* [12.15], are as follows:

- The inference that (HYP) the hominid who has been dubbed Lucy ( $x_1$ ) might have been bipedal ( $E$ ) from (OBS) observing the particular structure of her pelvis and knee bones ( $F$ ) and (BBK)

knowledge about how the structure of pelvis and knee bones relates to the locomotion of animals (EF).

- The inference that (HYP) two particles ( $x_1$  and  $x_2$ ) might have opposite electric charges ( $E$ ), from (OBS) observing their attraction ( $F$ ) and (BBK) knowledge of the Coulomb force (EF).

2. *Abduction of a Generalization*

(OBS)	$F$ with regard to <i>all observed</i> objects of class $D$
(BBK)	$E$ with regard to <i>some</i> objects explains $F$ with regard to those objects in a certain explanatory framework EF.
(HYP) It might be that $E$ with regard to <i>all existing</i> objects of class $D$	

Some examples of this pattern, which has also been called *rule abduction* [12.13], *law abduction* [12.14], and *selective law abduction* [12.15], are as follows:

- The inference that (HYP) *all* hominids of the last three million years ( $D$ ) might have been bipedal ( $E$ ), from (OBS) observing the similar structure of the pelvis and knee bones ( $F$ ) of *all observed* hominid skeletons dated to be younger than three million years ( $D$ ) and (BBK) knowledge about how the structure of pelvis and knee bones relates to the locomotion of animals (EF).
- The inference that (HYP) *all* emitted radiation from a particular chemical element ( $D$ ) might be electrically neutral ( $E$ ), from (OBS) observing in *all* experiments *conducted so far* that radiation emitted by this element ( $D$ ) continues in a straight path in an external magnetic field perpendicular to the stream of radiation ( $F$ ) and (BBK) knowledge of the Lorentz force and Newton's second law (EF).

3. *Existential abduction*, or the abduction of *the existence* of unknown objects from a particular class

(OBS)	$F$ with regard to $x_1, \dots, x_n$ ( $n \geq 1$ )
(BBK)	The existence of objects $y_1, \dots, y_m$ ( $m \geq 1$ ) of class $E$ would explain $F$ with regard to $x_1, \dots, x_n$ in a certain explanatory framework EF.
(HYP) It might be that there exist objects $y_1, \dots, y_m$ of class $E$	

Some examples of this pattern, which was already called *existential abduction* by *Thagard* [12.13],

and has also been called *first-order existential abduction* [12.14] and *selective type abduction* [12.15], are as follows:

- The inference that (HYP) a hominid ( $y_1$ ) of the genus *Australopithecus* ( $E$ ) might have lived in this area, from (OBS) observing a set of vulcanized foot imprints ( $x_1, \dots, x_n$  of class  $F$ ) and (BBK) the belief that these foot imprints are of an *Australopithecus* (EF).
  - The inference that (HYP) there might be other charged particles ( $y_1, \dots, y_m$  of class  $E$ ) in the chamber, from (OBS) observing deflections in the path ( $F$ ) of a charged particle ( $x_1$ ) in a chamber without external electric or magnetic fields and (BBK) the knowledge of the Coulomb and Lorentz forces and Newton's second law (EF).
4. *Conceptual abduction* or the abduction of a *new concept*

(OBS)  $F_1, \dots, F_m$  ( $m \geq 2$ ) with regard to each of  $x_1, \dots, x_n$  ( $n \geq 2$ )

(BBK) No known concept explains why  $F_1, \dots, F_m$  with regard to each of  $x_1, \dots, x_n$

---

(HYP) It might be that there is a *similarity* between the  $x_1, \dots, x_n$ , which can be labeled with a *new concept*  $E$  that explains why  $F_1, \dots, F_m$  with regard to each of  $x_1, \dots, x_n$  in a certain explanatory framework EF.

It was *Schurz* [12.14] who pointed out that this pattern is rational and useful for science only if the observation concerns several objects each individually having the same or similar properties, so that some form of conceptual unification is obtained. Otherwise, for each fact, it could be suggested that there exists an ad hoc power that explains (only) this single fact.

Some examples of this pattern, which largely coincides with the various types of *second-order abduction*, as *Schurz* [12.14] suggests and several types of *creative abduction* conceived by *Hoffmann* [12.15], are as follows:

- The inference that (HYP) there might be a new species of hominids ( $E$ ), from (OBS) observing various hominid fossils ( $x_1, \dots, x_n$ ) that are similar in many ways ( $F_1, \dots, F_m$ ) and (BBK) believing that these fossils cannot be classified in the current taxonomy of hominids (EF).
- The inference that (HYP) there might exist a new type of interaction ( $E$ ), from (OBS) observing similar interactive behavior

( $F_1, \dots, F_m$ ) between certain types of particles ( $x_1, \dots, x_n$ ) in similar experiments and (BBK) believing that this behavior cannot be explained by the already known interactions, properties of the involved particles and properties of the experimental setup (EF).

Using the terminology of *Magnani* [12.18] and following the distinction of *Schurz* [12.14], the first two patterns, the abduction of a singular fact and that of a generalization, can be considered as instances of *selective abduction*, as the agent selects an appropriate hypothesis in her background knowledge, while the latter two, existential abduction and conceptual abduction, can be called *creative abduction*, as the agent creates a new hypothetical concept or object. It has to be added that *Hoffmann* [12.15] would dispute this distinction, as he sees the third pattern (existential abduction) in the first place as the selection of an already known type (e.g., the genus *Australopithecus*), and not so much as the creation of a new token (someone of this genus of which his/her existence is now hypothesized).

As stated before, this list is not exhaustive. Further patterns have been identified, such as the abduction of a new perspective [12.15], for example, suggesting that a problem might have a geometrical solution instead of an algebraic one; *analogical abduction* [12.13], for example, explaining similar properties of water and light, by hypothesizing that light could also be wave-like; or *theoretical model abduction* [12.14], for example, explaining some observation by suggesting suitable initial conditions given some governing principles or laws. Some have even considered *visual abduction*, the inference from the observation itself to a statement describing this observation, as a separate pattern [12.19]. For some of these patterns (or instances of them), it is possible to argue that they are a special case of one of the patterns mentioned earlier. For instance, the suggestion of the wave nature of light can also be seen as an instance of conceptual abduction, in which the (mathematical) concept *wave behavior* is constructed to explain the similar properties of water and light; yet, it is true that the analogical nature of this inference makes it a special subpattern with interesting properties in itself. This is also how *Schurz* [12.14] presents it: In his classification, analogical abduction is one of the types of second-order existential abduction he conceives of.

Perhaps more important to note is that these patterns are not mutually exclusive given a particular instance of abductive reasoning. For instance, the inference that leads to the explanation of why a particular piece of iron is rusted can be described both as singular fact abduction (this piece of iron underwent a reaction

with oxygen) and as existential abduction (there were oxygen atoms present with which this piece of iron reacted). But in essence it describes the same explanation for the same explanandum.

Also, combinations occur. For instance, if a new particle is hypothesized as an explanation for an experimental anomaly (such as, for instance, *Wolfgang Pauli's* suggestion of the neutrino in the case of the anomalous  $\beta$  spectrum [12.20]), then this is both an instance of existential abduction – there is a not yet observed particle that causes the observed phenomenon – and an instance of conceptual abduction – these hypothesized particles are of a new kind of combination, which coincides with *Hoffmann's* [12.15] pattern of *creative fact abduction*. Yet in the mind of the scientist, this process of hypothesis formation might have occurred in a single reasoning step.

One should not, however, be too worried about these issues, if it is remembered that these patterns are categories for linguistic descriptions of actual reasoning processes. Any actual instance of hypothesis formation can be described in several ways by means of natural language, and some of these expressions can be formally analyzed in more than one way. Therefore, one should not focus too much on the exact classification of particular instances of hypothesis formation. Yet, this does not render meaningless the project of explicating various patterns of hypothesis formation. The goal of this project is to provide normative guidance for future hypothesis formation. If particular problems or observations can be looked at from different perspectives and, therefore, expressed in various ways, it is only beneficial for an agent to have multiple patterns of hypothesis formation at her disposal.

## 12.4 Abductive Reasoning and Adaptive Logics

Now, the various attempts to model abductive reasoning by means of adaptive logics are presented, though it first needs to be explained why the framework of adaptive logics is fit for this job.

First, adaptive logics allow for a direct implementation of defeasible reasoning steps (in casual applications of *affirming the consequent*). This makes it possible to construct logical proofs that nicely integrate defeasible (in this case ampliative) and deductive inferences. This corresponds to natural reasoning processes.

Second, the formal apparatus of an adaptive logic instructs exactly which formulas would falsify a (defeasible) reasoning step. As these formulas are assumed to be false (so long as one cannot derive them), they are called *abnormalities* in the adaptive logic literature. So, if one or a combination of these abnormalities is derived in a proof, it instructs in a formal way which defeasible steps cannot be maintained. This possibility of defeating previous reasoning steps mirrors nicely the dynamics found in actual human reasoning.

Third, for all adaptive logics in standard format, such as the presented logics  $\mathbf{LA}_s^r$  and  $\mathbf{MLA}_s^s$ , there are generic proofs for most of the important metatheoretical properties such as soundness and completeness [12.4].

So far, most research effort has been focused on modeling singular fact abduction, which already proves to be, even it appears to be the easiest case, a rich and

fruitful point of departure. This is not exclusive to the adaptive logics framework: In general, very little logics have been proposed for other forms besides singular abduction. Some of the few exceptions are [12.13] and [12.16]. (This last logic, which is an adaptive logic, suffers, however, from some complications [12.7, appendix B] and [12.21, p. 140].) For these reasons, this overview is limited to the various attempts to model singular fact abduction within the framework of adaptive logics.

The history of research into singular fact abduction within the adaptive logics community dates back to the early 2000s and can be traced through the articles [12.22–28]. Besides presenting early logics for singular fact abduction, this research has also shown that there actually exist two types of singular fact abduction (see also Sect. 12.5). In recent years, for each of these two types of abductions, an adaptive logic in a standard format (see Sect. 12.6) has been developed:  $\mathbf{LA}_s^r$  for practical abduction [12.29] and  $\mathbf{MLA}_s^s$  for theoretical abduction [12.30]. These will be the two logics that will be presented and explained in this chapter (see Sects. 12.7 and 12.8). It further needs to be noted that recent research has even pushed further by considering abduction from inconsistent theories [12.31], adaptations for use in AI ([12.32], improved version in [12.21, Ch. 5]), and a first logic for propositional singular fact abduction [12.7].

## 12.5 The Problem of Multiple Explanatory Hypotheses

The early research into logics for abduction has shown that two types of abduction logics can actually be constructed, depending on how the logic deals with multiple explanatory hypotheses for a single observation.

To explain this problem, consider the following example. Suppose one has to form hypotheses for the puzzling fact  $Pa$ , while one's background knowledge contains both  $(\forall x)(Qx \supset Px)$  and  $(\forall x)(Rx \supset Px)$ . There are two ways in which one can proceed. First, one could construct a logic in which one could derive only the disjunction  $(Qa \vee Ra)$  and not the individual hypotheses  $Qa$  and  $Ra$ . This first way, called *practical abduction* (according to the definition suggested in [12.24, pp. 224–225] and first used in [12.27]) and modeled by the logic  $\text{LA}_s^r$  ([12.29], Sect. 12.7), is suitable for modeling situations in which one has to *act* on the basis of the conclusions before having the chance to find out which hypothesis actually is the case. A good example is how people react to unexpected behavior. If someone suddenly starts to shout, people will typically react in a hesitant way, taking into account that either they themselves are somehow at fault or that the shouting person is just frustrated or crazy and acting inappropriately.

Second, someone with a theoretical perspective (for instance, a scientist or a detective) is interested in finding out which of the various hypotheses is the actual explanation. Therefore, it is important that s/he can *abduce* the individual hypotheses  $Qa$  and  $Ra$  in order to examine them further one by one. Early work on these kinds of logics has been done in [12.27, 28]. Yet, these logics have a quite complex proof theory. This is because, on the one hand, one has to be able to derive  $Qa$  and  $Ra$  separately, but on the other, one has to prevent

the derivation of their conjunction  $(Qa \wedge Ra)$ , because it seems counterintuitive to take the conjunction of two possible hypotheses as an explanation: For instance, if the street is wet, it would be weird to suggest that it has rained and that the fire department also just held an exercise. Moreover, if the two possible hypotheses are actually incompatible, it would lead to logical explosion in a classical logical context.

Logical explosion is the situation that just any statement can be derived from a certain premise set. In CL this occurs when a premise set contains a contradiction, that is, both a particular statement and its negation can be derived from the premise set, which makes an *ex falso quodlibet* argument possible. Briefly, such an argument goes as follows: suppose one's premise set contains both the statements  $p$  and  $\neg p$ . Then, by means of addition, one can first derive  $p \vee q$  for any random  $q$ . (Informally, as one already knows that  $p$  is true, any statement of the form *p or . . .* will also be true.) But, as  $\neg p$  also holds, one can derive  $q$  from this disjunction by means of a disjunctive syllogism (the logical rule that if you know that one side of disjunction is false, the other side has to be true to make the disjunction true).

The logic  $\text{MLA}_s^s$  [12.30] presented in this overview (Sect. 12.8) solves this problem by adding modalities to the language and deriving the hypotheses  $\diamond Qa$  and  $\diamond Ra$  instead of  $Qa$  and  $Ra$ . By conceiving of hypotheses as logical possibilities, the conjunction problem is automatically solved because  $\diamond Qa \wedge \diamond Ra$  does not imply  $\diamond(Qa \wedge Ra)$  in any standard modal logic. This approach also nicely coincides with the common idea that hypotheses are possibilities. These features make the logic  $\text{MLA}_s^s$  very suitable for the modeling of actual theoretical abductive reasoning processes.

## 12.6 The Standard Format of Adaptive Logics

Before the logics for abduction  $\text{LA}_s^r$  and  $\text{MLA}_s^s$  are presented, the reader should first be provided with the necessary background about the adaptive logics framework, and, more in particular, with the nuts and bolts of its standard format. This will, of course, be a limited introduction, and the reader is referred to, for example, [12.3] or [12.4] for a thorough introduction.

### Definition 12.1

An *adaptive logic in the standard format* is defined by a triple:

- (i) A *lower limit logic* (henceforth **LLL**): a reflexive, transitive, monotonic, and compact logic that has a characteristic semantics
- (ii) A *set of abnormalities*  $\Omega$ : a set of **LLL**-contingent formulas (that are not theorems of **LLL**) characterized by a logical form, or a union of such sets
- (iii) An adaptive *strategy*.

The lower limit logic **LLL** specifies the stable part of the adaptive logic. Its rules are unconditionally valid

in the adaptive logic, and anything that follows from the premises by **LLL** will never be revoked. Apart from that, it is also possible in an adaptive logic to derive defeasible consequences. These are obtained by assuming that the elements of the set of abnormalities are *as much as possible* false. The adaptive strategy is needed to specify *as much as possible*. This will become clearer further on.

Strictly speaking, the standard format for adaptive logics requires that a lower limit logic contains, in addition to the **LLL** operators, also the operators of **CL**. However, these operators have merely a technical role (in the generic meta-theory for adaptive logics) and are not used in the applications presented here. Therefore, given the introductory nature of this section, this will not be explained into further detail. In the logics presented in this chapter, the condition is implicitly assumed to be satisfied.

### 12.6.1 Dynamic Proof Theory

As stated before, a key advantage of adaptive logics is their *dynamic proof theory* which models human reasoning. This dynamics is possible because a *line* in an adaptive proof has – along with a line number, a formula and a justification – a fourth element, that is, the *condition*. A condition is a finite subset of the set of abnormalities and specifies which abnormalities need to be assumed to be false for the formula on that line to be derivable.

The inference rules in an adaptive logic reduce to three generic rules. Where  $\Gamma$  is the set of premises,  $\Theta$  is a finite subset of the set of abnormalities  $\Omega$  and  $Dab(\Theta)$  the (classical) disjunction of the abnormalities in  $\Theta$ , and where

$$A \quad \Delta. \quad (12.3)$$

indicates that  $A$  occurs in the proof on the condition  $\Delta$ , the inference rules are given by the generic rules

$$\begin{array}{l} \text{PREM} \quad \text{If } A \in \Gamma : \\ \quad \vdots \\ \quad \vdots \\ \hline A \quad \emptyset \end{array}$$

$$\begin{array}{l} \text{RU} \quad \text{If } A_1, \dots, A_n \vdash_{\text{LLL}} B : \\ A_1 \quad \Delta_1 \\ \quad \vdots \\ \quad \vdots \\ A_n \quad \Delta_n \\ \hline B \quad \Delta_1 \cup \dots \cup \Delta_n \end{array}$$

$$\begin{array}{l} \text{RC} \quad \text{If } A_1, \dots, A_n \vdash_{\text{LLL}} B \vee Dab(\Theta) \\ A_1 \quad \Delta_1 \\ \quad \vdots \\ \quad \vdots \\ A_n \quad \Delta_n \\ \hline B \quad \Delta_1 \cup \dots \cup \Delta_n \cup \Theta \end{array}$$

The premise rule **PREM** states that a premise may be introduced at any line of a proof on the empty condition. The unconditional inference rule **RU** states that, if  $A_1, \dots, A_n \vdash_{\text{LLL}} B$  and  $A_1, \dots, A_n$  occur in the proof on the conditions  $\Delta_1, \dots, \Delta_n$ , one may add  $B$  on the condition  $\Delta_1 \cup \dots \cup \Delta_n$ . The strength of an adaptive logic comes from the third rule, the conditional inference rule **RC**, which works analogously to **RU**, but introduces new conditions. So, it allows one to take defeasible steps based on the assumption that the abnormalities are false (this rule also makes clear that any adaptive proof can be transformed into a Fitch-style proof in the **LLL** by writing down for each line the disjunction of the formula and all of the abnormalities in the condition). Several examples of how these rules are employed will follow.

The only thing that is still needed is a criterion that defines when a line of the proof is considered to be defeated. At first sight, it seems straightforward to mark lines of which one of the elements of the condition is *unconditionally* derived from the premises, this means that it is derived on the empty condition (defeated lines in a proof are marked instead of deleted, because, in general, it is possible that they may later become unmarked in an extension of the proof). But this strategy, called the *simple strategy*, usually has a serious flaw. If it is possible to derive unconditionally a disjunction of abnormalities  $Dab(\Delta)$  that is *minimal*, that is, if there is no  $\Delta' \subset \Delta$  such that  $Dab(\Delta')$  can be unconditionally derived, the simple strategy would ignore this information. This is problematic, however, because at least one of the disjuncts of the ignored disjunction has to be true. Therefore, one can use the simple strategy only in cases where

$$\begin{array}{l} \Gamma \vdash_{\text{LLL}} Dab(\Delta) \\ \text{only if} \\ \text{there is an } A \in \Delta \text{ such that } \Gamma \vdash_{\text{LLL}} A \end{array} \quad (12.4)$$

with  $Dab(\Delta)$  being any disjunction of abnormalities out of  $\Omega$ . This condition will be met for the logic  $\mathbf{MLA}_s^s$  (Sect. 12.8); this logic will, hence, employ the simple strategy.

The majority of logics, however, do not meet this criterion and for those logics, more advanced strategies

have been developed. The best known of these are *reliability* and *minimal abnormality*. The logic  $\mathbf{LA}_s^r$  uses the reliability strategy. This strategy, which will be explained and illustrated in the following, orders to mark any line of which one of the elements is unconditionally derived as a disjunct from a minimal disjunction of abnormalities.

At this point, all elements are introduced to explain the naming of the two logics that will be presented in this chapter: As might be expected  $\mathbf{LA}$  and  $\mathbf{MLA}$  stand for *logic for abduction* and *modal logic for abduc-*

*tion* and the superscripts  $r$  and  $s$  stand for the adaptive strategies *reliability* and *simple strategy*, respectively. The subscript  $s$  originally denoted that the logic was formulated in the standard format for adaptive logics, but in [12.21], it is argued that it is more useful to interpret this  $s$  as that they are logics for *singular fact abduction*. After all, most adaptive logics are nowadays formulated in the standard format anyhow, and this allows us to contrast these logics with the logic  $\mathbf{LA}_\forall^r$  which is a logic for *abduction of generalizations* [12.16, 21].

## 12.7 $\mathbf{LA}_s^r$ : A Logic for Practical Singular Fact Abduction

In this section, the reader is introduced to the logic  $\mathbf{LA}_s^r$  [12.29] in an informal manner. This will allow the reader to gain a better understanding of the framework of adaptive logics and the functioning of its dynamic proof theory. In the next section, the same approach will be used for the logic  $\mathbf{MLA}_s^r$ . The formal definitions of both logics will be presented in the appendix for those who are interested.

In order to model abductive reasoning processes of singular facts, the logic  $\mathbf{LA}_s^r$  (as will the logic  $\mathbf{MLA}_s^r$ ) contains, in addition to deductive inference steps, defeasible reasoning steps based on an argumentation schema known as *affirming the consequent* (combined with Universal Instantiation)

$$(\forall\alpha)(A(\alpha) \supset B(\alpha)), B(\beta)/A(\beta). \quad (12.5)$$

The choice for a predicate logic is motivated by the fact that a material implication is used to model the relation between *explanans* and *explanandum*. As is well known that  $B \vdash_{\mathbf{CL}} A \supset B$ , a propositional logic would allow one to derive anything as a hypothesis. In the predicative case, the use of the universal quantifier can avoid this. This can be seen if we compare  $\vdash_{\mathbf{CL}} B(\beta) \supset (A(\beta) \supset B(\beta))$  with  $\not\vdash_{\mathbf{CL}} B(\beta) \supset (\forall\alpha)(A(\alpha) \supset B(\alpha))$  (see [12.7] for a propositional logic for abduction that solves this problem in another way).

Let the list of desiderata for this logic first be overviewed. This is important because in specifying the set of abnormalities and the strategy, one has to check whether they allow one to model practical abductive reasoning according to one's expectations. Apart from the fact that by means of this logic one should be able to derive hypotheses according to the schema of *affirming the consequent*, one has to make sure that one cannot derive – as a side effect – random hypotheses which

are not related to the explanandum. Finally, as has been pointed out in the introduction, it is a nice feature of adaptive logics that they enable one to integrate defeasible and deductive steps.

### 12.7.1 Lower Limit Logic

The lower limit logic of  $\mathbf{LA}_s^r$  is a classical first-order logic  $\mathbf{CL}$ . This means that the deductive inferences of this logic are the reasoning steps modeled by classical logic. Also, as this logic is an extension of classical logic, any classical consequence of a premise set will also be a consequence of the premise set according to this logic.

### 12.7.2 Set of Abnormalities $\Omega$

If one takes (here and in further definitions) the metavariables  $A$  and  $B$  to represent (well-formed) formulas,  $\alpha$  a variable and  $\beta$  a constant of the language in which the logic is defined  $\mathcal{L}$ , we can define the set of abnormalities of the logics  $\mathbf{LA}_s^r$  as

$$\begin{aligned} \Omega = \{ & (\forall\alpha)(A(\alpha) \supset B(\alpha)) \\ & \wedge (B(\beta) \wedge \neg A(\beta)) \} \mid \\ & \text{No predicate occurring in } B \\ & \text{occurs in } A \} \end{aligned} \quad (12.6)$$

The first line is the logical form of the abnormality; the second line in the definition is to prevent self-explanatory hypotheses. To understand the functioning of this logical form, consider the following example proof starting from the premise set,  $\{Qa, \forall x(Px \supset Qx)\}$ ,  $\forall x(Px \supset Rx)$ . The official layout of this chapter in two columns forces to split each line of the proof over



two lines and write the condition of the line on a second line, starting with an  $\rightarrow$  arrow

1	$\forall x(Px \supset Qx)$	-;PREM	
	$\rightarrow \emptyset$		
2	$Qa$	-;PREM	
	$\rightarrow \emptyset$		
3	$Pa \vee \neg Pa$	-; RU	
	$\rightarrow \emptyset$		
4	$Pa \vee (\forall x(Px \supset Qx) \wedge (Qa \wedge \neg Pa))$	1,2,3;RU	
	$\rightarrow \emptyset$		
5	$Pa$	4;RC	
	$\rightarrow \{\forall x(Px \supset Qx) \wedge (Qa \wedge \neg Pa)\}$		

From this premise set, one would like to be able to form the hypothesis  $Pa$ . One obtains this hypothesis as follows. One starts by writing two premises on the first two lines and a tautology on the third line (all these lines are not dependent on earlier lines, indicated by the dash). These three lines allow one then to derive the disjunction on line 4 by means of the unconditional inference rule RU. This disjunction has the exact form that allows one now to derive conditionally the hypothesis  $Pa$  from it by applying the rule RC.

From this hypothesis, one can reason further in a deductive way by applying, for example, modus ponens (note that the result of this inference has also a nonempty condition)

5	$Pa$	4;RC	
	$\rightarrow \{\forall x(Px \supset Qx) \wedge (Qa \wedge \neg Pa)\}$		
6	$\forall x(Px \supset Rx)$	-; PREM	
	$\rightarrow \emptyset$		
7	$Ra$	5,6;RU	
	$\rightarrow \{\forall x(Px \supset Qx) \wedge (Qa \wedge \neg Pa)\}$		

Suppose now that one comes to know that  $\neg Pa$  is the case and add this premise to the premise set and continue the proof. Strictly speaking, this is not what is actually done. What is actually done is to start a new proof with another premise set (the extended set). But it is easily seen that one can start this new proof with exactly the same lines as the old proof. This way, it looks as if one has extended the old proof. Therefore, it made sense to use the phrase *adding premises and continuing a proof* as it also nicely mirrors how human beings deal with incoming information: They do not start over their reasoning but incorporate the new information at

the point where they have arrived.

5	$Pa$	4;RC	
	$\rightarrow \{\forall x(Px \supset Qx) \wedge (Qa \wedge \neg Pa)\}$		
6	$\forall x(Px \supset Rx)$	-; PREM	
	$\rightarrow \emptyset$		
7	$Ra$	5, 6; RU	
	$\rightarrow \{\forall x(Px \supset Qx) \wedge (Qa \wedge \neg Pa)\}$		
8	$\neg Pa$	-;PREM	
	$\rightarrow \emptyset$		
9	$\forall x(Px \supset Qx) \wedge (Qa \wedge \neg Pa)$	1,2,6;RU	
	$\rightarrow \emptyset$		

This new premise makes it possible to derive unconditionally on line 9 the condition of the hypothesis  $Pa$ . At this point, it is clear that one should not trust anymore the hypothesis formed on line 5, which one indicates by marking this line with a checkmark, indicating that one lost one's confidence in this formula once one wrote down line 9. As the formula  $Ra$  is arrived at by reasoning further upon the hypothesis  $Pa$ , it has (at least) the same condition, and is, hence, at this point also marked.

In summary, each time one defeasibly derives a hypothesis, one has to state explicitly the condition the (suspected) truth of which would defeat the hypothesis. Therefore, one can assume the hypothesis to be true as long as one can assume the condition to be false; but as soon as one has evidence that the condition might be true, one should withdraw the hypothesis.

### 12.7.3 Reliability Strategy

In the previous example, one withdrew the hypothesis because its condition was explicitly derived. However, have a look at the following example proof from the premise set  $\{Qa, Ra, \forall x(Px \supset Qx), \forall x(\neg Px \supset Rx)\}$

1	$\forall x(Px \supset Qx)$	-;PREM	
	$\rightarrow \emptyset$		
2	$\forall x(\neg Px \supset Rx)$	-;PREM	
	$\rightarrow \emptyset$		
3	$Qa$	-;PREM	
	$\rightarrow \emptyset$		
4	$Ra$	-;PREM	
	$\rightarrow \emptyset$		
5	$Pa$	1,3;RC	
	$\rightarrow \{\forall x(Px \supset Qx) \wedge (Qa \wedge \neg Pa)\}$		
6	$\neg Pa$	2,4;RC	
	$\rightarrow \{\forall x(\neg Px \supset Rx) \wedge (Ra \wedge Pa)\}$		

There is clearly something fishy about this situation. As the conditions on lines 5 and 6 are not derivable from this premise set, logical explosion would allow one to derive anything from this premise set, if one were to use the simple strategy. Still, it is quite obvious that at least one of those two conditions has to be false, as the disjunction of these two conditions is a theorem of the lower limit logic. Yet, as one does not know from these premises which disjunct is true and which one is false, the most reliable thing to do is to mark both lines

⋮	⋮	⋮					
5	$Pa$	1,3;RC	✓ <sup>7</sup>				
	$\rightarrow \{\forall x(Px \supset Qx) \wedge (Qa \wedge \neg Pa)\}$						
6	$\neg Pa$	2,4;RC	✓ <sup>7</sup>				
	$\rightarrow \{\forall x(\neg Px \supset Rx) \wedge (Ra \wedge Pa)\}$						
7	$(\forall x(Px \supset Qx) \wedge (Qa \wedge \neg Pa))$ $\vee (\forall x(\neg Px \supset Rx) \wedge (Ra \wedge Pa))$	1-4;RU					
	$\rightarrow \emptyset$						

This marking strategy is called the *reliability strategy* and it orders one to mark lines for which an element of the condition has been unconditionally derived as a disjunct of a minimal disjunction of abnormalities (or in short, a minimal *Dab* formula). It is important to note that (1) the disjunction should only hold disjuncts that have the form of an abnormality (otherwise, a defeating disjunction could be constructed for every hypothesis) and (2) that this disjunction should be minimal (as disjunctions can always be extended by applications of the addition rule). To clarify this last point: Suppose one was able to derive the condition of line 5 by itself, then the disjunction on line 7 would not be minimal anymore and there would be no reason anymore to mark line 6.

### 12.7.4 Practical Abduction

The logic  $\mathbf{LA}_s^r$  is a logic for practical abduction (Sect. 12.5). This means that it solves the problem of multiple explanatory hypotheses by only allowing the disjunction of the various hypotheses to be derived. Consider the following example from the premise set  $\{Ra, \forall x(Px \supset Rx), \forall x(Qx \supset Rx)\}$

1	$\forall x(Px \supset Rx)$	-;PREM					
	$\rightarrow \emptyset$						
2	$\forall x(Qx \supset Rx)$	-;PREM					
	$\rightarrow \emptyset$						
3	$Ra$	-;PREM					
	$\rightarrow \emptyset$						

4	$Pa$	1,3;RC	✓ <sup>6</sup>				
	$\rightarrow \{\forall x(Px \supset Rx) \wedge (Ra \wedge \neg Pa)\}$						
5	$Qa$	2,3;RC	✓ <sup>7</sup>				
	$\rightarrow \{\forall x(Qx \supset Rx) \wedge (Ra \wedge \neg Qa)\}$						
6	$(\forall x(Px \supset Rx) \wedge (Ra \wedge \neg Pa))$ $\vee (\forall x((Qx \wedge \neg Px) \supset Rx)$ $\wedge (Ra \wedge \neg(Qa \wedge \neg Pa)))$	1-3;RC					
	$\rightarrow \emptyset$						
7	$(\forall x(Qx \supset Rx) \wedge (Ra \wedge \neg Qa))$ $\vee (\forall x((Px \wedge \neg Qx) \supset Rx)$ $\wedge (Ra \wedge \neg(Pa \wedge \neg Qa)))$	1-3;RC					
	$\rightarrow \emptyset$						
8	$\forall x((Px \vee Qx) \supset Rx)$	1,2;RU					
	$\rightarrow \emptyset$						
9	$Pa \vee Qa$	3,8;RC					
	$\rightarrow \{\forall x((Px \vee Qx) \supset Rx)$ $\wedge (Ra \wedge \neg(Pa \vee Qa))\}$						

Because of the fact that the minimal *Dab* formulas on lines 6 and 7 could be derived from the premises, the individual hypotheses *Pa* and *Qa* have to be withdrawn; yet, the condition of their disjunction on line 9 is not part of a minimal *Dab* formula from these premises. This shows that this logic only allows one to derive a disjunction in the case of multiple explanatory hypotheses, and none of the individual disjuncts.

### 12.7.5 Avoiding Random Hypotheses

Another important feature of a logic for abduction is that it prevents from allowing one to derive random hypotheses. The three most common ways to introduce random hypotheses is:

1. By deriving an explanation for a tautology, for example, deriving *Xa* from the theorems  $Pa \vee \neg Pa$  and  $\forall x(Xx \supset (Px \vee \neg Px))$
2. By deriving contradictions as explanations, which leads to logical explosion, for example, deriving  $Xa \wedge \neg Xa$  from *Pa* and the theorem  $\forall x((Xx \wedge \neg Xx) \supset Px)$
3. By deriving hypotheses that are not the most parsimonious ones, for example, deriving  $Pa \wedge Xa$  from *Qa* and  $\forall x(Px \supset Qx)$  (and its consequence  $\forall x((Px \wedge Xx) \supset Qx)$ ).

The logic  $\mathbf{LA}_s^r$  prevent these three ways by similar mechanisms as the mechanism to block individual hypotheses illustrated above. Elaborate examples for each of these three ways can be found in [12.29].

## 12.8 $\mathbf{MLA}_s^s$ : A Logic for Theoretical Singular Fact Abduction

In this section, the reader will be introduced to the logic  $\mathbf{MLA}_s^s$  [12.30] in a similar informal manner. Formal definitions can again be found in the appendix. Analogously, this logic also models deductive steps combined with applications of *affirming the consequent* (combined with universal instantiation); yet, it treats the problem of multiple explanatory hypotheses now in a different way: It allows one to derive these hypotheses individually; yet to avoid logical explosion caused by mutually exclusive hypotheses, it treats them as modal possibilities (Sect. 12.5).

The list of desiderata for this logic is very analogous as the one for the logic  $\mathbf{LA}_s^s$ , except for treating the problem of multiple explanatory hypotheses in a different manner. Specific for this logic (as this logic is aimed at modeling the reasoning of, e.g., scientists or detectives [12.33]) is the desideratum that it handles contradictory hypotheses, predictions, and counterevidence in a natural way.

### 12.8.1 Formal Language Schema

As this logic is a modal logic, the language of this logic is an extension of the language of the classical logic  $\mathbf{CL}$ . Let the standard predicative language of the classical logic be denoted with  $\mathcal{L}$ .  $C$ ,  $\mathcal{V}$ ,  $\mathcal{F}$ , and  $\mathcal{W}$  will further be used to refer, respectively, to the sets of individual constants, individual variables, all (well-formed) formulas of  $\mathcal{L}$ , and the closed (well-formed) formulas of  $\mathcal{L}$ .

$\mathcal{L}_M$ , the language of the logic  $\mathbf{MLA}_s^s$ , is  $\mathcal{L}$  extended with the modal operator  $\Box$ .  $\mathcal{W}_M$ , the set of closed formulas of  $\mathcal{L}_M$  is the smallest set that satisfies the following conditions:

1. If  $A \in \mathcal{W}$ , then  $A, \Box A \in \mathcal{W}_M$
2. If  $A \in \mathcal{W}_M$ , then  $\neg A \in \mathcal{W}_M$
3. If  $A, B \in \mathcal{W}_M$ ,  
then  $A \wedge B, A \vee B, A \supset B, A \equiv B \in \mathcal{W}_M$ .

It is important to notice that there are no occurrences of modal operators within the scope of another modal operator or a quantifier. The set  $\mathcal{W}_\Gamma$ , the subset of  $\mathcal{W}_M$ , the elements of which can act as premises in the logic, is further defined as

$$\mathcal{W}_\Gamma = \{\Box A \mid A \in \mathcal{W}\}. \quad (12.7)$$

It is easily seen that

$$\mathcal{W}_\Gamma \subset \mathcal{W}_M. \quad (12.8)$$

### 12.8.2 Lower Limit Logic

The **LLL** of  $\mathbf{MLA}_s^s$  is the predicative version of  $D$ , restricted to the language schema  $\mathcal{W}_M$ .  $D$  is characterized by a full axiomatization of predicate CL together with two axioms, an inference rule, and a definition

$$\mathbf{K} \quad \Box(A \supset B) \supset (\Box A \supset \Box B), \quad (12.9)$$

$$\mathbf{D} \quad \Box A \supset \neg \Box \neg A, \quad (12.10)$$

$$\mathbf{NEC} \quad \text{if } \vdash A, \text{ then } \vdash \Box A, \quad (12.11)$$

$$\diamond A =_{\text{df}} \neg \Box \neg A \quad (12.12)$$

This logic is one of the weakest normal modal logics that exist and is obtained by adding the  $D$ -axiom to the axiomatization of the better known minimal normal modal logic  $\mathbf{K}$ .

The semantics for this logic can be expressed by a standard possible world Kripke semantics where the accessibility relation  $R$  between possible worlds is *serial*, that is, for every world  $w$  in the model, there is at least one world  $w'$  in the model such that  $Rww'$ .

### 12.8.3 Intended Interpretation of the Modal Operators

As indicated above, explanatory hypotheses – the results of abductive inferences – will be represented by formulas of the form  $\diamond A$  ( $A \in \mathcal{W}$ ). Formulas of the form  $\Box B$  are used to represent explananda, other observational data, and relevant background knowledge. Otherwise, this information would not be able to revoke derived hypotheses (for instance,  $\neg A$  and  $\diamond A$  are not contradictory, whereas  $\Box \neg A$  and  $\diamond A$  are). The reason **D** is chosen instead of **K** is that it is assumed that the explananda and background information are together consistent. This assumption is modeled by the **D**-axiom (for instance, the premise set  $\{\Box \neg Pa, \Box (\forall x)Px\}$  is a set modeling an inconsistent set of background knowledge and observations, but in the logic **K**, this set would not be considered inconsistent, because anything cannot be derived from this set by *Ex Falso Quodlibet*. To be able to do this, the **D**-axiom is needed.)

### 12.8.4 Set of Abnormalities

Since the final form of the abnormalities is quite complex – although the idea behind is straightforward – two more basic proposals that are constitutive for the final form will first be considered and it will be shown why they are insufficient. Obviously, only closed well-formed formulas can be an element of any set of

abnormalities. This will not be explicitly stated each time.

### 12.8.5 First Proposal $\Omega_1$

This first proposal is a modal version of the set of abnormalities of the logic  $\mathbf{LA}_s^+$

$$\Omega_1 = \{\Box(\forall\alpha(A(\alpha) \supset B(\alpha)) \wedge (B(\beta) \wedge \neg A(\beta))) \mid \text{No predicate occurring in } B \text{ occurs in } A\} \quad (12.13)$$

Analogous to the logic  $\mathbf{LA}_s^+$ , this means that a derived hypothesis will be defeated if one shows explicitly that the hypothesis cannot be the case.

### 12.8.6 Simple Strategy

For this logic, the *simple strategy* can be used, which means, as stated before, that one has to mark lines for which one of the elements of the condition is unconditionally derived. It can easily be seen that the condition for the use of the simple strategy, that is,

$$\begin{aligned} &\Gamma \vdash_{\text{LLL}} \text{Dab}(\Delta) \\ &\text{only if} \\ &\text{there is an } A \in \Delta \text{ such that } \Gamma \vdash_{\text{LLL}} A, \end{aligned} \quad (12.14)$$

is fulfilled here. Since all premises have the form  $\Box A$ , the only option to derive a disjunction of abnormalities would be to apply addition, that is, to derive  $(\Box A \vee \Box B)$  from  $\Box A$  (or  $\Box B$ ), because it is well known that  $\Box(A \vee B) \not\vdash \Box A \vee \Box B$  in any standard modal logic (it is also possible to derive a disjunction from the premises by means of the *K*-axiom. For instance,  $\Box(A \supset B) \vdash \neg\Box A \vee \Box B$ , but the first disjunct will always be equivalent to a possibility ( $\Diamond\neg A$ ) and can, hence, not be an abnormality).

### 12.8.7 Contradictory Hypotheses

As a first example of the functioning of this logic, consider the following example starting from the premise set  $\{\Box Qa, \Box Ra, \Box\forall x(Px \supset Qx), \Box\forall x(\neg Px \supset Rx)\}$ . As the reader is by now probably accustomed to the functioning of the abnormalities, it is also shown how this logic is able to handle contradictory hypotheses without causing explosion.

$$\begin{aligned} 1 &\quad \Box\forall x(Px \supset Qx) && \text{-;PREM} \\ &\quad \longrightarrow \emptyset \\ 2 &\quad \Box\forall x(\neg Px \supset Rx) && \text{-;PREM} \\ &\quad \longrightarrow \emptyset \end{aligned}$$

$$\begin{aligned} 3 &\quad \Box Qa && \text{-;PREM} \\ &\quad \longrightarrow \emptyset \\ 4 &\quad \Box Ra && \text{-;PREM} \\ &\quad \longrightarrow \emptyset \\ 5 &\quad \Diamond Pa && 1,3;\text{RC} \\ &\quad \longrightarrow \{\Box(\forall x(Px \supset Qx) \wedge (Qa \wedge \neg Pa))\} \\ 6 &\quad \Diamond \neg Pa && 2,4;\text{RC} \\ &\quad \longrightarrow \{\Box(\forall x(\neg Px \supset Rx) \\ &\quad \quad \wedge (Ra \wedge \neg\neg Pa))\} \\ 7 &\quad \Diamond Pa \wedge \Diamond \neg Pa && 5,6;\text{RU} \\ &\quad \longrightarrow \{\Box(\forall x(Px \supset Qx) \wedge (Qa \wedge \neg Pa)), \\ &\quad \quad \Box(\forall x(\neg Px \supset Rx) \wedge (Ra \wedge \neg\neg Pa))\} \end{aligned}$$

$\Diamond Pa$  and  $\Diamond \neg Pa$  are derivable hypotheses because both the conditions on lines 5–7 are not unconditionally derivable from the premise set. It is also interesting to note that, because of the properties of the lower limit *D*, it is not possible to derive from these premises that  $\Diamond(Pa \wedge \neg Pa)$ . The conjunction of two hypotheses is never considered as a hypothesis itself, unless there is further background information that links these two hypotheses in some way.

### 12.8.8 Predictions and Evidence

To show that this logic handles predictions and (counter) evidence for these predictions in a natural way, let the premise set be extended with the additional implication  $\Box\forall x(Px \supset Sx)$

$$\begin{aligned} 8 &\quad \Box\forall x(Px \supset Sx) && \text{-;PREM} \\ &\quad \longrightarrow \emptyset \\ 9 &\quad \Diamond Sa && 5,8;\text{RU} \\ &\quad \longrightarrow \{\Box(\forall x(Px \supset Qx) \wedge (Qa \wedge \neg Pa))\} \end{aligned}$$

With this extra implication, the prediction  $\Diamond Sa$  can be derived. As long as one has no further information about this prediction (for instance, by observation), it remains a hypothesis derived on the same condition as  $\Diamond Pa$ . If one would test this prediction, one would have two possibilities. On the one hand, if the prediction turns out to be false, the premise  $\Box\neg Sa$  could be added to the premise set

$$\begin{aligned} &\vdots && \vdots \\ &\vdots && \vdots \\ 5 &\quad \Diamond Pa && 1,3;\text{RC} \quad \checkmark^{12} \\ &\quad \longrightarrow \{\Box(\forall x(Px \supset Qx) \\ &\quad \quad \wedge (Qa \wedge \neg Pa))\} \\ &\vdots && \vdots \\ &\vdots && \vdots \\ 9 &\quad \Diamond Sa && 5,8;\text{RU} \quad \checkmark^{12} \\ &\quad \longrightarrow \{\Box(\forall x(Px \supset Qx) \\ &\quad \quad \wedge (Qa \wedge \neg Pa))\} \end{aligned}$$

10	$\Box \neg Sa$	PREM
	$\longrightarrow \emptyset$	
11	$\Box \neg Pa$	8,10;RU
	$\longrightarrow \emptyset$	
12	$\Box (\forall x(Px \supset Qx)$	
	$\wedge (Qa \wedge \neg Pa))$	1,3,11;RU
	$\longrightarrow \emptyset$	

In this case, one could subsequently derive  $\Box \neg Pa$ , which would falsify the hypothesis  $\diamond Pa$ . On the other hand, if the prediction  $Sa$  turned out to be true, the premise  $\Box Sa$  could have been added, but this extension of the premise set would not allow us to derive  $\Box Pa$ . Since true predictions only *corroborate* the hypothesis but do not *prove* it, while false predictions directly *falsify* the hypothesis, one can say that this logic handles predictions in a *Popperian* way, although in using this vocabulary, the reader has to be reminded that  $MLA_s^s$  is a logic for modeling abduction and handling explanatory hypotheses, not a formal methodology of science. This logic has nothing to say about the confirmation of theories for which Popper actually employed the concepts of corroboration and falsification [12.34].

### 12.8.9 Contradictions

One of the three ways a logic of abduction could generate random hypotheses as a side effect is by allowing for the abduction of contradictions. How this is possible and how the logic prevents this are illustrated in the following proof from the premise set  $\{\Box Qa\}$

1	$\Box Qa$	-;PREM
	$\longrightarrow \emptyset$	
2	$\Box \forall x((Xx \wedge \neg Xx) \supset Qx)$	-;RU
	$\longrightarrow \emptyset$	
3	$\diamond (Xa \wedge \neg Xa)$	1,2;RC $\checkmark^4$
	$\longrightarrow \{\Box (\forall x((Xx \wedge \neg Xx) \supset Qx)$	
	$\wedge (Qa \wedge \neg (Xa \wedge \neg Xa)))\}$	
4	$\Box (\forall x((Xx \wedge \neg Xx) \supset Qx)$	1;RU
	$\wedge (Qa \wedge (\neg Xa \vee Xa)))$	
	$\longrightarrow \emptyset$	

### 12.8.10 Tautologies

Still, there are other ways to derive random hypotheses that are not prevented by the first proposal for the set of abnormalities  $\Omega_1$ . For instance,  $\Omega_1$  does not prevent that random hypotheses can be derived from a tautology, as illustrated by the following example. As is impossible in the following proof from the premise set  $\emptyset$  to unconditionally derive the abnormality in the condition of line 3 from the premises, the formula of line

3, the random hypothesis  $\diamond Xa$ , remains derived in every possible extension of the proof.

1	$\Box (Qa \vee \neg Qa)$	-;RU
	$\longrightarrow \emptyset$	
2	$\Box \forall x(Xx \supset (Qx \vee \neg Qx))$	-;RU
	$\longrightarrow \emptyset$	
3	$\diamond Xa$	1,2;RC
	$\longrightarrow \{\Box (\forall x(Xx \supset (Qx \vee \neg Qx))$	
	$\wedge ((Qa \vee \neg Qa) \wedge \neg Xa))\}$	

Therefore, let the set of abnormalities be adjusted to obtain the second proposal  $\Omega_2$ .

### 12.8.11 Second Proposal $\Omega_2$

No hypothesis can be abduced from a tautology if the abnormalities have the following form

$$\Omega_2 = \{ \Box (\forall \alpha (A(\alpha) \supset B(\alpha)) \wedge (B(\beta) \wedge \neg A(\beta))) \vee \Box \forall \alpha B(\alpha) \mid \text{No predicate occurring in } B \text{ occurs in } A \}$$
(12.15)

It is clear that one can keep using the simple strategy with this new set of abnormalities. It is also easily seen that all of the advantages and examples described above still hold. Each time one can derive an abnormality of  $\Omega_1$ , one can derive the corresponding abnormality of  $\Omega_2$  by a simple application of the *addition* rule. Finally, the problem raised by tautologies, as illustrated in the previous example, is solved in an elegant way, because the form of abnormalities makes sure that the abnormality will always be a theorem in case the explanandum is a theorem. So, nothing can be abduced from tautologies.

### 12.8.12 Most Parsimonious Explanantia

Still, there is third way to derive random hypotheses that cannot be prevented by  $\Omega_2$ . Consider, for instance, the following proof from the premise set  $\{\Box Ra, \Box \forall x(Px \supset Rx)\}$

1	$\Box Ra$	-;PREM
	$\longrightarrow \emptyset$	
2	$\Box \forall x(Px \supset Rx)$	-;PREM
	$\longrightarrow \emptyset$	
3	$\Box \forall x((Px \wedge Xx) \supset Rx)$	2;RU
	$\longrightarrow \emptyset$	
4	$\diamond (Pa \wedge Xa)$	1,3;RC
	$\longrightarrow \{\Box (\forall x((Px \wedge Xx) \supset Rx)$	
	$\wedge (Ra \wedge \neg (Pa \wedge Xa))) \vee \Box \forall x Rx\}$	

$$5 \quad \diamond Xa \quad 4;RU \\ \longrightarrow \{\Box(\forall x((Px \wedge Xx) \supset Rx) \\ \wedge (Ra \wedge \neg(Pa \wedge Xa))) \vee \Box\forall xRx\}$$

The reason why the random hypothesis  $\diamond Xa$  can be derived is the absence of a mechanism to ensure that the abduced hypothesis is the most parsimonious one and not the result of *strengthening the antecedent* of an implication. Before defining the final and actual set of abnormalities that also prevent this way of generating random hypotheses, a new notation has to be introduced to keep things as perspicuous as possible.

### 12.8.13 Notation

Suppose  $A_{PCN}(\alpha)$  is the *prenex conjunctive normal* form of  $A(\alpha)$ . This is an equivalent form of the formula  $A(\alpha)$  where all quantifiers are first moved to the front of the expression and where, consequently, the remaining (quantifier-free) expression is written in a conjunctive normal form, that is, as a conjunction of disjunctions of literals. Hence, apart from the quantifiers which are all at the front, it is entirely made up of a big conjunction of subformulae.

$$A_{PCN}(\alpha) = (Q_1\gamma_1) \cdots (Q_m\gamma_m) \\ (A_1(\alpha) \wedge \cdots \wedge A_n(\alpha)) \\ \text{and } \vdash A_{PCN}(\alpha) \equiv A(\alpha) \quad (12.16)$$

with  $m \geq 0$ ,  $n \geq 1$ ,  $Q_i \in \{\forall, \exists\}$  for  $i \leq m$ ,  $\gamma_i \in \mathcal{V}$  for  $i \leq m$ ,  $\alpha \in \mathcal{V}$  and  $A_i(\alpha)$  disjunctions of literals in  $\mathcal{F}$  for  $i \leq n$ .

Then, the new notation  $A_i^{-1}(\alpha)$  ( $1 \leq i \leq n$ ) can be introduced so that there is a way to take out one of the conjuncts of a formula in the PCN form. In cases where the conjunction consists of only one conjunct (and, obviously, no more parsimonious explanation is possible), the substitution with a random tautology will make sure that the condition for parsimony, added in the next set of abnormalities, is satisfied trivially

$$\text{if } n > 1: \\ A_i^{-1}(\alpha) =_{\text{df}} (Q_1\gamma_1) \cdots (Q_m\gamma_m) (A_1(\alpha) \wedge \cdots \\ \wedge A_{i-1}(\alpha) \wedge A_{i+1}(\alpha) \wedge \cdots \wedge A_n(\alpha)) \\ \text{with } A_j (1 \leq j \leq n) \text{ the } j\text{-th conjunct} \\ \text{of } A_{PCN}(\alpha) \quad (12.17)$$

$$\text{if } n = 1: \\ A_1^{-1}(\alpha) =_{\text{df}} \top \\ \text{with } \top \text{ any tautology of } \mathbf{CL}. \quad (12.18)$$

### 12.8.14 Final Proposal $\Omega$

With this notation, the logical form of the set of abnormalities  $\Omega$  of the logic  $\mathbf{MLA}_s^s$  can be written as

$$\Omega = \{\Box(\forall\alpha(A(\alpha) \supset B(\alpha)) \wedge (B(\beta) \wedge \neg A(\beta))) \\ \vee \Box\forall\alpha B(\alpha) \vee \bigvee_{i=1}^n \Box\forall\alpha(A_i^{-1}(\alpha) \supset B(\alpha)) \mid \\ \text{No predicate occurring in } B \text{ occurs in } A\} \quad (12.19)$$

This form might look complex, but its functioning is quite straightforward. What is constructed is the disjunction of the three reasons why one should refrain from considering  $A(\beta)$  as a good explanatory hypothesis for the phenomenon  $B(\beta)$ , even if one has  $(\forall\alpha)(A(\alpha) \supset B(\alpha))$ . The disjunction will make sure that the hypothesis  $A(\beta)$  is rejected as soon as one of the following is the case:

1. When  $\neg A(\beta)$  is derived
2. When  $B(\beta)$  is a tautology (and obviously, does not need an explanatory hypothesis)
3. When  $A(\beta)$  has a redundant part and is, therefore, not an adequate explanatory hypothesis.

For the same reasons, as stated in the description of  $\Omega_2$ , one can keep using the simple strategy and all of the advantages and examples described above will still hold.

Let one has a look at how this final set of abnormalities solve the previous problem. As the condition is fully written out, one can easily see that the third disjunct  $\Box\forall x(Px \supset Rx)$  is actually a premise, and that, hence, the abnormality on line 4 unconditionally derivable is

$$\begin{array}{ll} 1 & \Box Ra \quad \text{---;PREM} \\ & \longrightarrow \emptyset \\ 2 & \Box\forall x(Px \supset Rx) \quad \text{---;PREM} \\ & \longrightarrow \emptyset \\ 3 & \Box\forall x((Px \wedge Qx) \supset Rx) \quad 2;RU \\ & \longrightarrow \emptyset \\ 4 & \diamond(Pa \wedge Qa) \quad 1,3;RC \checkmark^5 \\ & \longrightarrow \{\Box(\forall x((Px \wedge Qx) \supset Rx) \\ & \quad \wedge (Ra \wedge \neg(Pa \wedge Qa))) \vee \Box\forall xRx \\ & \quad \vee \Box\forall x(Px \supset Rx) \vee \Box\forall x(Qx \supset Rx)\} \\ 5 & \Box(\forall x((Px \wedge Qx) \supset Rx) \wedge \quad 2;RU \\ & \quad (Ra \wedge \neg(Pa \wedge Qa))) \vee \Box\forall xRx \\ & \quad \vee \Box\forall x(Px \supset Rx) \vee \Box\forall x(Qx \supset Rx) \\ & \longrightarrow \emptyset \end{array}$$

This concludes the informal presentation of this logic, which, in its final form, meets all desiderata put up front.

## 12.9 Conclusions

There is quite some ground covered in this chapter, the main purpose of which was to show in a direct yet nuanced fashion the feasibility and the limits of modeling hypothetical reasoning by means of formal logics. It started with an argument for this claim in a general way, showing which assumptions one has to assume or reject to take this view. As far as there is argued for the feasibility of this project, the attention was also drawn to certain limits, pitfalls, and disadvantages of it. This discussion was then expanded by identifying four main abduction patterns, which showed that no pattern of hypothetical reasoning can be easily modeled.

In the second part of this chapter, gears were shifted and a glimpse was shown of what is already possible today with current logical techniques, by explaining in detail two logics originating in the adaptive logics framework:  $\mathbf{LA}_s^r$  for practical fact abduction and  $\mathbf{MLA}_s^s$  for theoretical singular fact abduction. The purpose of including the full details of these logics is threefold: First, it shows the reader how certain steps, which are

admittedly modest, can be taken in the project of formally modeling hypothetical reasoning. At the same time, the reader is introduced to the unificational framework of adaptive logics that shows promise to take some further steps along the road. Finally, it also shows that the use of formal models draws the attention to various issues about these reasoning patterns which were previously left unattended, for example, the difference between practical and theoretical abduction or the importance of avoiding random hypotheses by restricting the use of tautologies and contradictions.

However, if one looks at the prospect of modeling abductive reasoning by means of formal (adaptive) logics, one has to conclude that so far only the tip of the iceberg has been scratched. At present, apart from a single exception, only logics have been devised for singular fact abduction, which is, in fact, the most easy of the various patterns of abduction. Yet, the complications that already arise on this level warn dreamers that the road ahead will be steep and arduous.

## 12.A Appendix: Formal Presentations of the Logics $\mathbf{LA}_s^r$ and $\mathbf{MLA}_s^s$

In this appendix, the logics  $\mathbf{LA}_s^r$  and  $\mathbf{MLA}_s^s$  will, for the sake of completeness be defined in a formal and precise way. This section is limited to what is needed to present these specific logics. For a more general formal presentation of adaptive logics in the standard format, see [12.4].

Like any adaptive logic in the standard format, the logics  $\mathbf{LA}_s^r$  and  $\mathbf{MLA}_s^s$  are characterized by the triple of a lower limit logic, a set of abnormalities, and an adaptive strategy.

For  $\mathbf{LA}_s^r$ , the lower limit logic is  $\mathbf{CL}$ , the strategy is the *reliability strategy* and the set of abnormalities  $\Omega_{\mathbf{LA}_s^r}$  is defined by

$$\begin{aligned} \Omega_{\mathbf{LA}_s^r} = \{ & (\forall \alpha (A(\alpha) \supset B(\alpha)) \\ & \wedge (B(\beta) \wedge \neg A(\beta))) \mid \\ & \text{No predicate occurring in } B \\ & \text{occurs in } A \} \end{aligned} \quad (12.20)$$

For  $\mathbf{MLA}_s^s$ , the lower limit logic is  $\mathbf{D}$ , the strategy is the *simple strategy* and the set of abnormalities  $\Omega_{\mathbf{MLA}_s^s}$  is, relying on the previously introduced abbreviation, defined by

$$\begin{aligned} \Omega_{\mathbf{MLA}_s^s} = \{ & \Box (\forall \alpha (A(\alpha) \supset B(\alpha)) \\ & \wedge (B(\beta) \wedge \neg A(\beta))) \\ & \vee \Box \forall \alpha B(\alpha) \\ & \vee \bigvee_{i=1}^n \Box \forall \alpha (A_i^{-1}(\alpha) \supset B(\alpha)) \mid \\ & \text{No predicate occurring in } B \\ & \text{occurs in } A \} \end{aligned} \quad (12.21)$$

### 12.A.1 Proof Theory

The proof theory of these logics is characterized by the three generic inference rules introduced in Sect. 12.2 and the following definitions.

Within adaptive logics, proofs are considered to be chains of subsequent stages. A *stage of a proof* is a sequence of lines obtained by application of the three generic rules. As such, every proof starts off with the first stage which is an empty sequence. Each time a line is added to the proof by applying one of the inference rules, the proof comes to its next stage, which is the sequence of lines written so far extended with the new line.

**Definition 12.2 (Minimal Dab formula at stage  $s$ )**

A Dab formula  $Dab(\Delta)$  ( $Dab(\Theta)$  is the (classical) disjunction of the abnormalities in a finite subset  $\Theta$  of the set of abnormalities  $\Omega$ ) is a *minimal Dab formula at stage  $s$*  if and only if  $Dab(\Delta)$  is derived on the empty condition at stage  $s$ , and there is no  $\Delta' \subset \Delta$  for which  $Dab(\Delta')$  is derived on the empty condition at stage  $s$ .

**Definition 12.3 (Set of unreliable formulas  $U_s(\Gamma)$  at stage  $s$ )**

The *set of unreliable formulas  $U_s(\Gamma)$  at stage  $s$*  is the union of all  $\Delta$  for which  $Dab(\Delta)$  is a minimal Dab formula at stage  $s$ .

**Definition 12.4 (Marking for the reliability strategy)**

Line  $i$  with condition  $\Theta$  is *marked for the reliability strategy at stage  $s$*  of a proof if and only if  $\Theta \cap U_s(\Gamma) \neq \emptyset$ .

**Definition 12.5 (Marking for the simple strategy)**

Line  $i$  with condition  $\Theta$  is *marked for the simple strategy at stage  $s$*  of a proof, if stage  $s$  contains a line of which  $A \in \Theta$  is the formula and  $\emptyset$  the condition.

**Definition 12.6 (Derivation of a formula at stage  $s$ )**

A formula  $A$  is *derived from  $\Gamma$  at stage  $s$*  of a proof if and only if  $A$  is the formula of a line that is unmarked at stage  $s$ .

**Definition 12.7 (Final derivation of a formula at stage  $s$ )**

A formula  $A$  is *finally derived from  $\Gamma$  at stage  $s$*  of a proof if and only if  $A$  is derived at line  $i$ , line  $i$  is not marked at stage  $s$  and every extension of the proof in which  $i$  is marked may be further extended in such a way that line  $i$  is unmarked.

Using the simple strategy, it is not possible that a marked line becomes unmarked at a later stage of a proof. Therefore, the final criterion reduces for this strategy to the requirement that the line remains unmarked in every extension of the proof.

**Definition 12.8 (Final derivability for  $\mathbf{LA}_s^r$ )**

For all  $\Gamma \subset \mathcal{W}_\Gamma$ :  $\Gamma \vdash_{\mathbf{LA}_s^r} A$  ( $A \in \mathcal{Cn}_{\mathbf{LA}_s^r}(\Gamma)$ ) if and only if  $A$  is finally derived in an  $\mathbf{LA}_s^r$ -proof from  $\Gamma$ .

**Definition 12.9 (Final derivability for  $\mathbf{MLA}_s^s$ )**

For all  $\Gamma \subset \mathcal{W}_\Gamma$ :  $\Gamma \vdash_{\mathbf{MLA}_s^s} A$  ( $A \in \mathcal{Cn}_{\mathbf{MLA}_s^s}(\Gamma)$ ) if and only if  $A$  is finally derived in a  $\mathbf{MLA}_s^s$ -proof from  $\Gamma$ .

**12.A.2 Semantics**

The semantics of an adaptive logic is obtained by a selection on the models of the lower limit logic. For a more elaborate discussion of the following definitions, the reader is referred to the original articles and the aforementioned theoretical overviews of adaptive logics.

**Definition 12.10**

A  $\mathbf{CL}$ -model  $M$  of the premise set  $\Gamma$  is *reliable* if and only if  $\{A \in \Omega \mid M \models A\} \subseteq \Delta_1 \cup \Delta_2 \cup \dots$  with  $\{Dab(\Delta_1), Dab(\Delta_2), \dots\}$  the set of minimal Dab-consequences of  $\Gamma$ .

**Definition 12.11**

A  $\mathbf{D}$ -model  $M$  of the premise set  $\Gamma$  is *simply all right* if and only if  $\{A \in \Omega \mid M \models A\} = \{A \in \Omega \mid \Gamma \vdash_{\mathbf{D}} A\}$ .

**Definition 12.12 (Semantic consequence of  $\mathbf{LA}_s^r$ )**

For all  $\Gamma \subset \mathcal{W}_\Gamma$ :  $\Gamma \models_{\mathbf{LA}_s^r} A$  if and only if  $A$  is verified by all *reliable* models of  $\Gamma$ .

**Definition 12.13 (Semantic consequence of  $\mathbf{MLA}_s^s$ )**

For all  $\Gamma \subset \mathcal{W}_\Gamma$ :  $\Gamma \models_{\mathbf{MLA}_s^s} A$  if and only if  $A$  is verified by all *simply all right* models of  $\Gamma$ .

The fact that these two logics are in a standard form warrants that the following theorems hold.

**Theorem 12.1 Soundness and completeness of  $\mathbf{LA}_s^r$ )**

$\Gamma \vdash_{\mathbf{LA}_s^r} A$  if and only if  $\Gamma \models_{\mathbf{LA}_s^r} A$ .

**Theorem 12.2 (Soundness and completeness of  $\mathbf{MLA}_s^s$ )**

$\Gamma \vdash_{\mathbf{MLA}_s^s} A$  if and only if  $\Gamma \models_{\mathbf{MLA}_s^s} A$ .



## References

- 12.1 N. Rescher: *Hypothetical Reasoning* (North-Holland, Amsterdam 1964)
- 12.2 R. Koons: Defeasible reasoning. In: *The Stanford Encyclopedia of Philosophy*, ed. by E. Zalta (Stanford Univ., Stanford 2014), Spring 2014 edn.
- 12.3 C. Straßer: *Adaptive Logics for Defeasible Reasoning: Applications in Argumentation, Normative Reasoning and Default Reasoning* (Springer, Dordrecht 2013)
- 12.4 D. Batens: A universal logic approach to adaptive logics, *Logica Universalis* **1**, 221–242 (2007)
- 12.5 D. Batens: *Adaptive Logics and Dynamic Proofs. Mastering the Dynamics of Reasoning with Special Attention to Handling Inconsistency* (Ghent Univ., Ghent 2010), <http://logica.ugent.be/adlog/book.html>
- 12.6 P. Paul: AI approaches to abduction. In: *Abductive Reasoning and Uncertainty Management Systems, Handbook of Defeasible Reasoning and Uncertainty Management Systems, Vol. 4*, ed. by D. Gabbay, R. Kruse (Kluwer Acad., Dordrecht 2000) pp. 35–98
- 12.7 M. Beirlaen, A. Aliseda: A conditional logic for abduction, *Synthese* **191**(15), 3733–3758 (2014)
- 12.8 N.R. Hanson: *Patterns of Discovery: An Inquiry into the Conceptual Foundations of Science* (Cambridge Univ. Press, Cambridge 1958)
- 12.9 N.R. Hanson: Is there a logic of scientific discovery? In: *Current Issues in the Philosophy of Science*, ed. by H. Feigl, G. Maxwell (Holt, Rinehart and Winston, New York 1961) pp. 20–35
- 12.10 G. Harman: The inference to the best explanation, *Philosophical Rev.* **74**(1), 88–95 (1965)
- 12.11 T. Nickles: Introductory essay: Scientific discovery and the future of philosophy of science. In: *Scientific Discovery, Logic and Rationality*, ed. by T. Nickles (Reidel, Dordrecht 1980) pp. 1–59
- 12.12 H. Simon: Does scientific discovery have a logic?, *Philosophy Sci.* **40**, 471–480 (1973)
- 12.13 P. Thagard: *Computational Philosophy of Science* (MIT, Cambridge 1988)
- 12.14 G. Schurz: Patterns of abduction, *Synthese* **164**, 201–234 (2008)
- 12.15 M. Hoffmann: Theoric transformations and a new classification of abductive inferences, *Trans. Charles S. Peirce Soc.* **46**(4), 570–590 (2010)
- 12.16 T. Gauderis, F. Van De Putte: Abduction of generalizations, *Theoria* **27**(3), 345–363 (2012)
- 12.17 A. Aliseda: *Abductive Reasoning. Logical Investigation into Discovery and Explanation*, Vol. 330 (Springer, Dordrecht 2006), Synthese Library
- 12.18 L. Magnani: *Abduction, Reason and Science: Processes of Discovery and Explanation* (Kluwer/Plenum, New York 2001)
- 12.19 P. Thagard, C. Shelley: Abductive reasoning: Logic, visual thinking, and coherence. In: *Logic and Scientific Methods*, Vol. 259, ed. by M.L. Dalla Chiara, K. Doets, D. Mundici, J. van Benthem (Kluwer Acad., Dordrecht 1997) pp. 413–427
- 12.20 T. Gauderis: To envision a new particle or change an existing law? Hypothesis formation and anomaly resolution for the curious spectrum of the  $\beta$  decay spectrum, *Stud. Hist. Philos. Mod. Phys.* **45**(1), 27–45 (2014)
- 12.21 T. Gauderis: Patterns of Hypothesis Formation: At the Crossroads of Philosophy of Science, Logic, Epistemology, Artificial Intelligence and Physics, Ph.D. Thesis (Ghent Univ., Ghent 2013)
- 12.22 J. Meheus, L. Verhoeven, M. Van Dyck, D. Provijn: Ampliative adaptive logics and the foundation of logic-based approaches to abduction. In: *Logical and Computational Aspects of Model-Based Reasoning*, ed. by L. Magnani, N.J. Nersessian, C. Pizzi (Kluwer Academic, Dordrecht 2002) pp. 39–71
- 12.23 D. Batens, J. Meheus, D. Provijn, L. Verhoeven: Some adaptive logics for diagnosis, *Log. Log. Philos.* **11/12**, 39–65 (2003)
- 12.24 J. Meheus, D. Batens: A formal logic for abductive reasoning, *Log. J. IGPL* **14**, 221–236 (2006)
- 12.25 J. Meheus: Adaptive logics for abduction and the explication of explanation-seeking processes. In: *Abduction and the Process of Scientific Discovery*, ed. by O. Pombo, A. Gerne (Centro de Filosofia das Ciências, Lisboa 2007) pp. 97–119
- 12.26 J. Meheus, D. Provijn: Abduction through semantic tableaux versus abduction through goal-directed proofs, *Theoria* **22**(3), 295–304 (2007)
- 12.27 H. Lycke: The Adaptive logics approach to abduction. In: *Logic Philosophy and History of Science in Belgium*, Proc. Young Res. Days, ed. by E. Weber, T. Libert, P. Marage, G. Vanpaelmel (KVAB, Brussels 2009) pp. 35–41
- 12.28 H. Lycke: A formal explication of the search for explanations: The adaptive logics approach to abductive reasoning, *Log. J. IGPL* **20**(2), 497–516 (2012)
- 12.29 J. Meheus: A formal logic for the abduction of singular hypotheses. In: *Explanation, Prediction, and Confirmation*, ed. by D. Dieks, W. Gonzalez, S. Hartmann, T. Uebel, M. Weber (Springer, Dordrecht 2011) pp. 93–108
- 12.30 T. Gauderis: Modelling abduction in science by means of a modal adaptive logic, *Found. Sci.* **18**(4), 611–624 (2013)
- 12.31 D. Provijn: The generation of abductive explanations from inconsistent theories, *Log. J. IGPL* **20**(2), 400–416 (2012)
- 12.32 T. Gauderis: An adaptive logic based approach to abduction in AI, Proc. 9th Int. Workshop Nonmonotonic Reasoning, Action Change (NARC), Barcelona, ed. by S. Sardina, S. Vassos (2011) pp. 1–6, Online Publication, <http://ijcai-11.iiia.csic.es/files/proceedings/W4-%20NRAC11-Proceedings.pdf>
- 12.33 T. Gauderis: The problem of multiple explanatory hypotheses, future directions for logic, Proc. PhDs in Logic III, Brussels, 2011, ed. by L. Demey, J. Devuyt (College Publications, London 2012) pp. 45–54
- 12.34 K. Popper: *The Logic of Scientific Discovery* (Routledge, London 1959)

# 13. Abductive Reasoning in Dynamic Epistemic Logic

Angel Nepomuceno–Fernández, Fernando Soler–Toscano, Fernando R. Velázquez–Quesada

This chapter proposes a study of abductive reasoning addressing it as an epistemic process that involves both an agent's information and the actions that modify this information. More precisely, this proposal presents and discusses definitions of an abductive problem and an abductive solution in terms of an agent's information (her knowledge and beliefs) and the involved epistemic actions (observation and belief revision). The discussion is then formalized with tools from dynamic epistemic logic; under such framework, the properties of the given definitions are studied, an epistemic action representing the application of an abductive step is introduced, and an illustrative example is provided. A number of the most interesting properties of abductive reasoning (those highlighted by Peirce) are shown to be better modeled within this approach.

13.1	<b>Classical Abduction</b> .....	270
13.2	<b>A Dynamic Epistemic Perspective</b> .....	272
13.2.1	What Is an Abductive Problem? .....	272
13.2.2	What Is an Abductive Solution? .....	273
13.2.3	How is the Best Explanation Selected? .....	273
13.2.4	How is the Best Explanation Incorporated Into the Agent's Information? .....	274
13.2.5	Abduction in a Picture .....	274
13.3	<b>Representing Knowledge and Beliefs</b> .....	275
13.3.1	Language and Models .....	275
13.3.2	Operations on Models .....	276
13.4	<b>Abductive Problem and Solution</b> .....	278
13.4.1	Abductive Problem .....	278
13.4.2	Classifying Problems .....	279
13.4.3	Abductive Solutions .....	279
13.4.4	Classifying Solutions .....	280
13.5	<b>Selecting the Best Explanation</b> .....	281
13.5.1	Ordering Explanations .....	282
13.6	<b>Integrating the Best Solution</b> .....	284
13.6.1	Abduction in a Picture, Once Again .....	285
13.6.2	Further Classification .....	285
13.6.3	Properties in a Picture .....	287
13.7	<b>Working with the Explanations</b> .....	287
13.7.1	A Modality .....	288
13.8	<b>A Brief Exploration to Nonideal Agents</b> .....	289
13.8.1	Considering Inference .....	290
13.8.2	Different Reasoning Abilities .....	290
13.9	<b>Conclusions</b> .....	290
	<b>References</b> .....	292

Within logic, abductive reasoning has been studied mainly from a purely syntactic perspective. Definitions of an abductive problem and its solution(s) are given in terms of a theory and a formula, and therefore most of the formal logical work on the subject has focused on:

1. Discussing what a theory and a formula should satisfy in order to constitute an abductive problem, and what a formula should satisfy in order to be an abductive solution [13.1]; see also Chap. 10
2. Proposing algorithms to find abductive solutions [13.2–6]
3. Analyzing the structural properties of abductive consequence relations [13.7–9].

In all these studies, which follow the so-called Aliseda–Kakas/Kowalski–Magnani/Meheus (AKM)-schema of abduction Chap. 10, explanationism and consequentialism are considered, but the epistemic character of abductive reasoning seems to have been

pushed into the background. Such character is considered crucial in this chapter, as it will be discussed.

This chapter's main proposal is an *epistemic and dynamic* approach to abductive reasoning. The proposal is close to the ideas of [13.10–13] in that it stresses the key role that agents play within the abductive reasoning scenario; after all, at the heart, abduction deals with agents and their (individual or collective) information. In this sense, this collaboration is closer to the Gabbay–Woods (GW)-schema [13.14, 15], see also Chap. 14, which is based on the concept of *ignorance problem* that arises when a cognitive agent has a cognitive target that cannot be attained from what she currently knows, and thus highlights the distinctive epistemic feature of abduction that is key to this chapter's considerations. Even so, this presentation goes one step further, as it fully adopts a dynamic perspective by making explicit the actions involved in the abductive process; after all, abduction studies the way agents react epistemically (as individuals or groupwise) to new observations.

More precisely, this proposal argues (Sect. 13.2) that abductive reasoning can be better understood as a *process* that involves *an agent's information*. To this end, it presents definitions of an abductive problem and an abductive solution in terms of an agent's knowledge and her beliefs as well as a *subjective* criteria for selecting *the agent's* best explanation, and outlines a policy through which the chosen abductive solution can be integrated into the agent's information. Then, the discussed ideas and definitions are formalized using tools from *dynamic epistemic logic* (DEL). This choice is not accidental: classical epistemic logic (EL [13.16, 17]) with its possible worlds semantic model is a powerful framework that allows to represent an agent's knowledge and beliefs not only about propositional facts but also about her own information. Its dynamic extension, DEL [13.18, 19], allows the representation of diverse epistemic actions (as diverse forms of announcements and different policies for belief revision) that make such

information change. Section 13.3 introduces the needed tools, and then the ideas and definitions discussed in Sect. 13.2 are formalized in Sects. 13.4, 13.5, 13.6, and 13.7. The chapter closes with a brief exploration (Sect. 13.8) of the epistemic and dynamic aspects of abductive reasoning that are brought to light when non-ideal agents are considered.

*Abductive reasoning* The concept of abductive reasoning has been discussed in various fields, and this has led to different ideas of what abduction should consist of (see [13.20], among others). For example, while certain authors claim that there is an abductive problem only when neither the observed  $\chi$  nor its negation follows from a theory [13.2], others say that there is also an abductive problem when, though  $\chi$  does not follow, its negation does [13.1], a situation that has been typically called a belief revision problem. There are also several opinions of what an abductive solution is. Most of the work on strategies for finding abductive solutions focuses on formulas that are already part of the system (the aforementioned [13.2–6]), while some others take a broader view, allowing not only changes in the underlying logical consequence relation [13.21] but also the creation and modification of concepts [13.22].

The present proposal focuses on a simple account: Abductive reasoning will be understood as a reasoning process that goes from a single unjustified fact to its abductive explanations, where an explanation is a formula of the system that satisfies certain properties. Still, similar epistemic and dynamic approaches can be made to other interpretations of abduction, as those that involve the creation of new concepts or changes in awareness [13.23, 24].

*Abductive reasoning in dynamic epistemic logic* This contribution is a revised version of a proposal whose different parts have been presented in diverse venues. While Sects. 13.2, 13.4, and 13.6 are based on [13.25], Sects. 13.5 and 13.7 are based on [13.26] and Sect. 13.8 is based on [13.27].

## 13.1 Classical Abduction

After Peirce's formulation of abductive reasoning (see [13.28] and Chap. 10), he immediately adds [13.29, p. 231] that:

“[The abductive solution] cannot be abductively inferred, or if you prefer the expression, cannot be abductively conjectured, until its entire content is already present in the premises, *If [the abductive solution] were true, [the abductive problem] would be a matter of course.*”

According to these ideas, abduction is a process that is triggered when a surprising fact is observed by an epistemic agent. Although the process returns an explicative hypothesis, the genuine result of an abductive inference is the plausibility of such hypothesis. The truth of the obtained hypothesis is thereby conjectured as plausible, which makes abduction an inferential process of a nonmonotonic character whose conclusion is rather a provisional proposal that could be revised in the light of new information.

When formalized within logical frameworks, the key concepts in abductive reasoning have traditionally taken the following form (Chap. 10). First, it is said that an abductive problem arises when there is a formula that does not follow from the current theory.

### Definition 13.1 Abductive problem

Let  $\Phi$  and  $\chi$  be a theory and a formula, respectively, in some language  $\mathcal{L}$ . Let  $\vdash$  be a consequence relation on  $\mathcal{L}$ :

- The pair  $(\Phi, \chi)$  constitutes a (*novel*) *abductive problem* when neither  $\chi$  nor  $\neg\chi$  are consequences of  $\Phi$ , that is, when

$$\Phi \not\vdash \chi \quad \text{and} \quad \Phi \not\vdash \neg\chi.$$

- The pair  $(\Phi, \chi)$  constitutes an *anomalous abductive problem* when, though  $\chi$  is not a consequence of  $\Phi$ ,  $\neg\chi$  is, that is, when

$$\Phi \not\vdash \chi \quad \text{and} \quad \Phi \vdash \neg\chi.$$

It is typically assumed that the theory  $\Phi$  is a set of formulas closed under logical consequence, and that  $\vdash$  is a truth-preserving consequence relation.

Consider a novel abductive problem. The observation of a  $\chi$  about which the theory  $\Phi$  does not have any opinion shows that  $\Phi$  is incomplete. Further information that completes  $\Phi$  making  $\chi$  a consequence of it solves the problem, as now the theory is strong enough to *explain*  $\chi$ . Consider now an anomalous abductive problem. The observation of a  $\chi$  whose negation is entailed by the theory shows that the theory contains a mistake. Now two steps are needed. First, perform a *theory revision* that stops  $\neg\chi$  from being a consequence of  $\Phi$ ; this turns the anomalous problem into a novel one, and now the search for further information that completes the theory, making  $\chi$  a consequence of it, can be performed. Here are the formal definitions.

### Definition 13.2 Abductive solution

- Given a *novel* abductive problem  $(\Phi, \chi)$ , the formula  $\eta$  is said to be an *abductive solution* when

$$\Phi \cup \{\eta\} \vdash \chi.$$

- Given an *anomalous* abductive problem  $(\Phi, \chi)$ , the formula  $\eta$  is an *abductive solution* when it is possible to perform a theory revision to get a *novel* problem  $(\Phi', \chi)$  for which  $\eta$  is a solution.

This definition of an abductive solution is often considered as too weak:  $\eta$  can take many trivial forms, as anything that contradicts  $\Phi$  (then everything, including  $\chi$ , follows from  $\Phi \cup \{\eta\}$ ) and even  $\chi$  itself (clearly,  $\Phi \cup \{\chi\} \vdash \chi$ ). Further conditions can be imposed in order to define more satisfactory solutions; here are some of them [13.1] (Chap. 10).

### Definition 13.3 Classification of abductive solutions

Let  $(\Phi, \chi)$  be an abductive problem. An abductive solution  $\eta$  is

consistent	iff	$\Phi, \eta \not\vdash \perp$
explanatory	iff	$\eta \not\vdash \chi$
minimal	iff	for every other solution $\zeta$ , $\eta \vdash \zeta$ implies $\zeta \vdash \eta$

The *consistency* requirement discards solutions that are inconsistent with the theory, something a reasonable explanation should not do. In a similar way, the *explanatory* requirement discards those explanations that would justify the problem by themselves, since it is preferred that the explanation only complements the current theory. Finally, the *minimality* requirement works as Occam's razor, looking for the simplest explanation: A solution is minimal when it is in fact logically equivalent to any other solution it implies. For further details on these definitions, the reader is referred to Chap. 10.

## 13.2 A Dynamic Epistemic Perspective

The present contribution proposes an approach to abductive reasoning from an epistemic and dynamic perspective. Instead of understanding abductive reasoning as a process that *modifies a theory* whenever *there is a formula that is not entailed by the theory under some particular consequence relation*, as the traditional definition of an abductive problem does, the proposed approach understands abductive reasoning as a process that *changes an agent's information* whenever, *due to some epistemic action, the agent has come to know or believe a fact that she could not have predicted otherwise*.

Such an epistemic and dynamic approach is natural. First, abduction, as other forms of nonmonotonic reasoning (e.g., belief revision, default reasoning), is classified as form of a common-sense reasoning rather than a mathematical one, and most of its classic examples involve *real* agents and their information (e.g., Mary observes that the light does not go on; Karen observes that the lawn is wet; Holmes observes that Mr. Wilson's right cuff is very shiny). Thus, even though abductive reasoning has been linked to scientific theories (as interpreted in philosophy of science), in its most basic forms it deals with an agent's (or a set of agents') information. Second, abductive reasoning implies a change in the agent's information (Mary assumes that the electricity supply has failed; Karen assumes it has rained; Holmes assumes Mr. Wilson has done a lot of writing lately), and thus it is essential to distinguish the different stages during the abductive process: the stage before the observation, the stage after the observation has raised the abductive problem (and thus the one when the agent starts looking for an explanation), and the stage in which the explanation that has been chosen is incorporated into the agent's information. This describes, of course, a dynamic process.

There is a final issue that is crucial for an epistemic approach to abductive reasoning. From this contribution's perspective, abductive reasoning involves not one epistemic attitude (as is typically assumed in most approaches) but rather (at least) two: that of those propositions about which the agent has full certainty; and that of those propositions that she considers very likely but she still cannot be certain about. The reason is that an agent typically tries to explain facts she has come to *know* due to some observation, but the chosen solution, being a *hypothesis* that might be dropped in the light of further observations, should not attain the full certainty status. The use of different epistemic notions also gives more flexibility to deal with a wider variety of abductive problems and abductive solutions,

making the analysis closer, from the authors' perspective, to Peirce's original formulation.

All in all, the abductive process can be studied by asking four questions:

1. What is an abductive problem?
2. What is an abductive solution?
3. How is the *best* solution(s) selected?
4. How does the agent assimilate the chosen solution(s)?

In the following, answers to these questions are discussed.

### 13.2.1 What Is an Abductive Problem?

There are, from an epistemic and dynamic perspective, two important concepts in the definition of an abductive problem. The first is what a formula  $\chi$  should satisfy in order to become an abductive problem. The second is the action that triggers the abductive problem, that is, the action that turns a formula  $\chi$  into an abductive problem.

For the former concept, a formula is typically said to be an abductive problem when it is surprising. There are different ways to define *a surprising observation of  $\chi$*  (some of them in a DEL setting [13.30]). Most of the approaches that define this notion in terms of what the agent knows (believes) understand a surprise as something that does not follow from such knowledge (beliefs). In other words, it is said that a given  $\chi$  is surprising whenever the agent does not know (believe) it, or, more radically, whenever the agent knows (believes)  $\neg\chi$ .

Now, note how in the context of abductive reasoning it is not reasonable to define a surprising observation in terms of what the agent knows (believes) *after such epistemic action*. The reason is that, after observing  $\chi$ , an agent would typically come to know (believe) it. Thus, if the mentioned definitions are followed focusing on the agent's information after the observation, no  $\chi$  would be surprising and there would be no abductive problems at all! It is more reasonable to define a surprising observation not in terms of what the agent knows (believes) as a result of the observation, but rather in terms of what she knew (believed) *before* it. More precisely, it will be said that a known (believed)  $\chi$  is surprising with respect to an agent whenever she could not have come to know (believe) it.

Of course, the meaning of the sentence *the agent could have come to know (believe)  $\chi$*  still needs to be clarified. This is a crucial notion, as it will indicate not only when a formula  $\chi$  is an abductive problem (the

agent could not have come to know (believe)  $\chi$ ), but also what a formula  $\eta$  needs in order to be an abductive solution (with the help of  $\eta$ , the agent could have come to know (believe)  $\chi$ ). Here the ability to come to know (believe) a given formula will be understood as the ability to infer it, and the simplest way to state this idea is the following: An agent could have come to know (believe)  $\chi$  if and only if there is an implication  $\eta \rightarrow \chi$  such that the agent knew both the implication and its antecedent. Other formulations that do not use the material implication  $\rightarrow$  are also possible (e.g., the agent may know both  $\neg\eta \vee \chi$  and  $\eta$  to come to know  $\chi$ ), but in the semantic model this contribution uses (Sect. 13.3), they are logically equivalent to the proposed one.

With respect to the action that triggers an abductive problem  $\chi$ , this action is typically assumed to be the observation of  $\chi$  itself. Here a more general idea will be considered: The action that triggers the abductive problem will be simply the observation of *some* formula  $\psi$ . Thus, though  $\psi$  should indeed be related to  $\chi$  (after all,  $\chi$  is an abductive problem because the agent comes to know  $\chi$  by observing  $\psi$ ), the agent will not be restricted to look for explanations of the formula that has been observed: She will also be able to look for explanations of *any* formula  $\chi$  she has come to know (believe) through the observation but could not have come to know (believe) by herself before. Note how other actions are also reasonable, as the agent might want to explain a belief she attained after a belief revision (Sect. 13.4.1).

Here is the intuitive definition of an abductive problem in full detail:

“Let  $s_1$  represent the epistemic state of an agent, and let  $s_2$  be the epistemic state that results from the agent observing some given  $\psi$ . A formula  $\chi$  constitutes an abductive problem for the agent at  $s_2$  whenever  $\chi$  is known and there is no implication  $\eta \rightarrow \chi$  such that the agent knew both the implication and its antecedent at  $s_1$ .”

It is important to emphasize how an abductive problem has been defined *with respect to an agent and stage* (i.e., *some epistemic situation*). Thus, whether a formula is an abductive problem depends on the formula *but also on the information* of that given agent at that given stage. The definition is given purely in terms of the agent’s knowledge, but it can also be given purely in terms of her beliefs, or even in terms of both, as it will be seen later.

The presented definition could seem very restrictive. Even if the reader agrees with the basic idea ( $\chi$  is an abductive problem for a given agent whenever she knows  $\chi$  but she could not have come to know (believe) it), she/he does not need to agree with the way key parts of it are understood. Nevertheless, as stated

in the introduction, this contribution does not intend on providing a full account of the many different understandings of what abductive reasoning does. Rather, its aim is to show how an epistemic and dynamic perspective can shed a new light on the way abductive reasoning is understood, even when assuming its simplest interpretation.

### 13.2.2 What Is an Abductive Solution?

In this proposal’s setting, an abductive solution for a given  $\chi$  will be defined in terms of what the agent could have been able to infer *before* the observation that raised the problem. As mentioned before, it will be said that  $\eta$  is a solution for the abductive problem  $\chi$  when the agent could have come to know (believe)  $\chi$  with the help of  $\eta$ . In this simple case in which the ability to come to know (believe) a given formula is understood as the ability to infer the formula by means of a simple modus ponens step, the following definition is obtained:

“A formula  $\eta$  constitutes an abductive solution for the abductive problem  $\chi$  at some given state  $s_2$  if the agent knew  $\eta \rightarrow \chi$  at the previous state  $s_1$ . Thus, the set of solutions for an abductive problem  $\chi$  is the set of antecedents of implications which have  $\chi$  as consequent and were known before the observation that triggered the abductive problem.”

Note how abductive solutions are looked for not when the agent has come to know (believe)  $\chi$ , but rather at the stage immediately before it. Thus,  $\eta$  is a solution when, had it been known (believed) before, would have allowed the agent to come to know (believe) (to predict/expect)  $\chi$ .

### 13.2.3 How is the Best Explanation Selected?

Although there are several notions of explanation for modeling the behavior of why-questions in scientific contexts (e.g., the law model, the statistical relevance model, or the genetic model), most of these consider a consequence (entailment) relation; *explanation* and *consequence* go typically hand in hand. However, finding suitable and reasonable criteria for selecting *the best* explanation has constituted a fundamental problem in abductive reasoning [13.31–33], and in fact many authors consider it to be the heart of the subject. Many approaches are based on logical criteria, but beyond requisites to avoid triviality and certain restrictions to the syntactic form, the definition of suitable criteria is still an open problem. Some approaches have suggested the use of *contextual aspects*, such as an ordering among formulas or among full theories. In particular,

for the latter, a typical option is the use of *preferential models* based on qualitative properties that are beyond the pure causal or deductive relationship between the abductive problem and its abductive solution. However, these preference criteria are seen as an external device which works on top of the deductive part of the explanatory mechanism, and as such they have been criticized because they seem to fall outside the logical framework.

Approaching abductive reasoning from an epistemic point of view provides a different perspective. It has been discussed already how the explanation an agent will choose for a given abductive problem does not depend on how the problematic formula could have been predicted, but rather on how *the agent* could have predicted it. In general, different agents have different information, and thus they might disagree in what each one calls *the best explanation* (and even in what each one calls *explanation* at all). This suggests that, instead of looking for criteria to select *the best explanation*, the goal should be a criterion to select *the agent's best explanation*. Now, once the agent has a set of formulas that explain the abductive problem from her point of view, how can she choose the best? This proposal's answer makes use of the fact that the considered agents have not only knowledge but also beliefs: Among all these explanations, some are more plausible than others *from her point of view*. These are precisely the ones the agent will choose when trying to explain a surprising observation: The best explanation can be defined in terms of a preference ordering among the agent's epistemic possibilities. It could be argued that this criterion is not *logical in the classic sense* because it is not based exclusively on the *deductive* relationship between the observed fact and different ways in which it could have been derived. Nevertheless, it is *logical in a broader sense* since it does depend on the agent's information: her knowledge and, crucially, her beliefs.

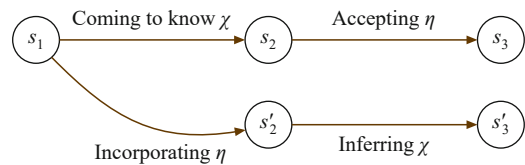
### 13.2.4 How is the Best Explanation Incorporated Into the Agent's Information?

Once the best explanation has been selected, it has to be incorporated into the agent's information. One of

the features that distinguishes abductive reasoning from deductive reasoning is its nonmonotonic nature: The chosen explanation does not need to be true, and in fact can be discarded in the light of further information. This indicates that an abductive solution cannot be assimilated as knowledge. Nevertheless, an epistemic agent has not only this hard form of information which is not subjected to modifications; she also has a soft form that can be revised as many times as it is needed: *beliefs*. Therefore, once the best abductive solution  $\eta$  has been chosen, the agent's information can be changed, leading her to *believe* that  $\eta$  is the case.

### 13.2.5 Abduction in a Picture

It is interesting to notice how the stated definitions of abductive problem and abductive solution rely on some form of *counterfactivity*, as in Peirce's original formulation (and also as discussed in [13.15]): A given  $\eta$  is a solution of a problem  $\chi$  if it would have allowed the agent to predict  $\chi$ . This can be better described with the following diagram.



The upper path is the real one: By means of an observation, the agent goes from the epistemic state  $s_1$  to the epistemic state  $s_2$  in which she knows  $\chi$ , and by accepting the abductive solution  $\eta$  she goes further to  $s_3$ . The existence of this path, the fact that  $\chi$  is an abductive problem and  $\eta$  is one of its abductive solutions, indicates that, at  $s_1$ , the lower path would have been possible: Incorporating  $\eta$  to the agent's information would have taken her to an epistemic state  $s'_2$  where she would have been able to infer  $\chi$ . Of course,  $s'_3$  is not identical to  $s_3$ : In  $s'_3$  both  $\eta$  and  $\chi$  are *equally reliable* because the second is inferred from the first, but in  $s_3$ ,  $\eta$  is less reliable than  $\chi$  since although the second is obtained via an observation, the first is just a hypothesis that is subject to revision in the light of further information.

### 13.3 Representing Knowledge and Beliefs

As mentioned, the most natural framework for formalizing the discussed ideas is that of DEL, the *dynamic extension of epistemic logic*. In particular, with the *plausibility models* of [13.34] it is possible to represent an agent’s knowledge and beliefs as well as acts of observation and belief revision, all of which are crucial to the stated understanding of the abductive process. This section introduces these needed tools; the discussed definitions will be formalized in Sect. 13.4.

#### 13.3.1 Language and Models

##### Definition 13.4 Language

Given a set of atomic propositions  $P$ , formulas  $\varphi$  of the language  $\mathcal{L}$  are given by

$$\varphi ::= p \mid \neg\varphi \mid \varphi \vee \varphi \mid \langle \leq \rangle \varphi \mid \langle \sim \rangle \varphi,$$

where  $p \in P$ . Formulas of the form  $\langle \leq \rangle \varphi$  are read as *there is a world at least as plausible (as the current one) where  $\varphi$  holds*, and those of the form  $\langle \sim \rangle \varphi$  are read as *there is a world epistemically indistinguishable (from the current one) where  $\varphi$  holds*. Other Boolean connectives ( $\wedge, \rightarrow, \leftrightarrow$ ) as well as the universal modalities,  $[\leq]$  and  $[\sim]$ , are defined as usual ( $[\leq]\varphi := \neg\langle \leq \rangle\neg\varphi$  and  $[\sim]\varphi := \neg\langle \sim \rangle\neg\varphi$  for the latter).

The modalities  $\langle \leq \rangle$  and  $\langle \sim \rangle$ , respectively, make it possible to define the notions of belief and knowledge within  $\mathcal{L}$ . The language’s semantic model, a *plausibility model*, is defined as follows.

##### Definition 13.5 Plausibility model

Let  $P$  be a set of atomic propositions. A *plausibility model* is a tuple  $M = \langle W, \leq, V \rangle$ , where:

1.  $W$  is a nonempty set of *possible worlds*
2.  $\leq \subseteq (W \times W)$  is a locally connected and conversely well-founded preorder, the agent’s *plausibility relation*, representing the plausibility order of the worlds from her point of view ( $w \leq u$  is read as *u is at least as plausible as w*)
3.  $V : W \rightarrow \wp(P)$  is an *atomic valuation function*, indicating the atoms in  $P$  that are true at each possible world.

A *pointed plausibility model*  $(M, w)$  is a plausibility model with a distinguished world  $w \in W$ .

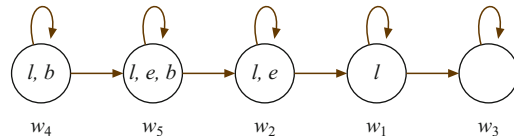
Before proceeding, recall that a relation  $R \subseteq (W \times W)$  is *locally connected* when every two elements that are  $R$ -comparable to a third are also  $R$ -comparable. It is *conversely well-founded* when there is no infinite  $\bar{R}$ -

ascending chain of elements in  $W$ , where  $\bar{R}$ , the *strict* version of  $R$ , is defined as  $\bar{R}wu$  iff  $Rwu$  and not  $Ruw$ . Finally, it is a *preorder* when it is reflexive and transitive.

The key idea behind plausibility models is that an agent’s beliefs can be defined as what is true in the *most plausible worlds from the agent’s perspective*, and modalities for the plausibility relation  $\leq$  will allow this definition to be formed. In order to define the agent’s knowledge, the approach is to assume that two worlds are epistemically indistinguishable for the agent if and only if she considers one of them at least as plausible as the other (if and only if they are comparable via  $\leq$ ). The *epistemic indistinguishability relation*  $\sim$  can therefore be defined as the union of  $\leq$  and its converse, that is, as  $\sim := \leq \cup \geq$ . Thus,  $\sim$  is the symmetric closure of  $\leq$  and hence  $\leq \subseteq \sim$ . Moreover, since  $\leq$  is reflexive and transitive,  $\sim$  is an *equivalence* relation. This epistemic indistinguishability relation  $\sim$  should not be confused with the *equal plausibility* relation, denoted by  $\simeq$ , and defined as the *intersection* of  $\leq$  and  $\geq$ , that is,  $\simeq := \leq \cap \geq$ . For further details and discussion on these models, their requirements and their properties, the reader is referred to [13.34, 35].

##### Example 13.1

The following diagram represents a plausibility model  $M$  based on the atomic propositions  $P := \{l, e, b\}$ . Circles represent possible worlds (named  $w_1$  up to  $w_5$ ), and each one of them includes exactly the atomic propositions that are true at that world (e.g., at  $w_2$ , the atomic propositions  $l$  and  $e$  are true, but  $b$  is false). Arrows represent the plausibility relation, with transitive arcs omitted (so  $w_4 \leq w_5 \leq w_2 \leq w_1 \leq w_3$ , but also  $w_4 \leq w_2, w_4 \leq w_1, w_4 \leq w_3$  and so on). Moreover,  $\sim$  is then the full Cartesian product, that is, for every worlds  $u$  and  $v$  in the model,  $u \sim v$ .



For the semantic interpretation, the two modalities  $\langle \leq \rangle$  and  $\langle \sim \rangle$  are interpreted with the help of their respective relations in the standard modal way.

##### Definition 13.6 Semantic interpretation

Let  $(M, w)$  be a pointed plausibility model with  $M = \langle W, \leq, V \rangle$ . Atomic propositions and Boolean operators



are interpreted as usual. For the remaining cases,

$$(M, w) \Vdash \langle \leq \rangle \varphi \text{ iff } \exists u \in W \text{ s.t. } w \leq u \text{ \& } (M, u) \Vdash \varphi$$

$$(M, w) \Vdash \langle \sim \rangle \varphi \text{ iff } \exists u \in W \text{ s.t. } w \sim u \text{ \& } (M, u) \Vdash \varphi .$$

### Defining Knowledge and Beliefs

The notion of knowledge in plausibility models is defined by means of the epistemic indistinguishability relation in the standard way: The agent knows  $\varphi$  at some world  $w$  if and only if  $\varphi$  is the case in every world she considers to be epistemically possible from  $w$ . (This makes knowledge a very strong notion, corresponding to an “absolutely unrevisable belief” [13.34]). The modality  $[\sim]$  can be used to this end. For the notion of beliefs, the idea is, as stated before, that the agent believes  $\varphi$  at a given  $w$  if and only if  $\varphi$  is the case in the most plausible worlds from  $w$ . Thanks to the properties of the plausibility relation (a locally connected and conversely well-founded preorder),  $\varphi$  is true in the most plausible (i. e., the  $\leq$ -maximal) worlds from  $w$  if and only if, in accordance with the plausibility order, from some moment onward there are only  $\varphi$ -worlds (see [13.34, 36, 37] for the technical details). The modalities  $\langle \leq \rangle$  and  $[\leq]$  can be used to this end. Summarizing,

$$\begin{array}{ll} \text{The agent knows } \varphi & K\varphi := [\sim]\varphi \\ \text{The agent believes } \varphi & B\varphi := \langle \leq \rangle [\leq]\varphi \end{array}$$

Observe how, since  $\leq \subseteq \sim$ , the formula  $K\varphi \rightarrow B\varphi$  is valid (but its converse is not).

The dual of these notions, epistemic possibility and most likely possibility, can be defined as the correspondent modal duals

$$\hat{K}\varphi := \langle \sim \rangle \varphi \quad \hat{B}\varphi := [\leq] \langle \leq \rangle \varphi .$$

### Example 13.2

Consider the plausibility model  $M$  of Example 13.1, and take  $w_2$  as the evaluation point. Since  $w_2 \sim u$  holds for every possible world  $u$  in the model, every world is epistemically possible from  $w_2$ 's perspective. But every world in the model satisfies  $b \rightarrow l$  (the implication is true at  $w_2, w_1$ , and  $w_3$  because the antecedent  $b$  is false, and true at  $w_4$  and  $w_5$  because the consequent  $l$  is true), so  $[\sim](b \rightarrow l)$ , that is,  $K(b \rightarrow l)$  is true at  $w_2$ : *The agent knows  $b \rightarrow l$  at  $w_2$ .* On the other hand,  $\neg l$  is not true in every world, but it is true in  $w_3$ , the most plausible one from  $w_2$ 's perspective, so  $\langle \leq \rangle [\leq] \neg l$ , that is,  $B\neg l$ , is true at  $w_2$ : *The agent believes  $\neg l$  at  $w_2$ .* Moreover, observe how  $b$  is neither known (it is not true in every

epistemically indistinguishable world) nor believed (it is not true in the most plausible worlds) at  $w_2$ . Still, it is true in some epistemic possibilities from  $w_2$  (e.g.,  $w_5$ ); hence,  $\langle \sim \rangle b$  (i. e.,  $\hat{K}b$ ) holds at  $w_2$ : *At that world, the agent considers  $b$  possible.*

A more detailed description of this framework, a number of the epistemic notions that can be defined within it, its technical details and its axiom system can be found in [13.34].

Following the DEL idea, actions that modify an agent's information can be represented as operations that transform the underlying semantic model. In the rest of this section, operations that can be applied over plausibility models will be recalled, and extensions of the language that allow to describe the changes such operations bring about will be provided. These will be used in Sect. 13.4 to represent and describe abductive reasoning.

## 13.3.2 Operations on Models

### Update, Also Known as Observation

The most natural operation over Kripke-like semantic models is that of *update*. This operation reduces the domain of the model, and is typically given in terms of the formula the worlds should satisfy in order to survive the operation.

#### Definition 13.7 Update operation

Let the tuple  $M = \langle W, \leq, V \rangle$  be a plausibility model and let  $\psi$  be a formula in  $\mathcal{L}$ . The *update* operation yields the plausibility model  $M_{\psi!} = \langle W', \leq', V' \rangle$  where  $W' := \{w \in W \mid (M, w) \Vdash \psi\}$ ,  $\leq' := \leq \cap (W' \times W')$  and, for every  $w \in W'$ ,  $V'(w) := V(w)$ .

This operation reduces the domain of the model (preserving only those worlds that satisfy the given  $\psi$ ) and restricts the plausibility relation and the atomic valuation function accordingly. Since a submodel is obtained, the operation preserves the (universal) properties of the plausibility relation and hence it preserves plausibility models: If  $M$  is a plausibility model, then so is  $M_{\psi!}$ .

In order to describe the effects of an update within the language, existential modalities of the form  $\langle \psi! \rangle$  are used, for every formula  $\psi$ . Here is their semantic interpretation

$$(M, w) \Vdash \langle \psi! \rangle \varphi \text{ iff } (M, w) \Vdash \psi$$

$$\text{and } (M_{\psi!}, w) \Vdash \varphi .$$

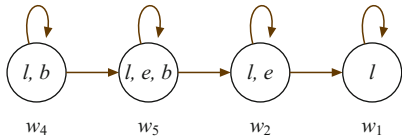
In words, an update formula  $\langle \psi! \rangle \varphi$  holds at  $(M, w)$  if and only if  $\psi$  is the case (i. e., the evaluation point

will survive the operation) and, after the update,  $\varphi$  is the case. The universal modality  $[\psi!]$  is defined as the modal dual of  $\langle\psi!\rangle$ , that is,  $[\psi!]\varphi := \neg\langle\psi!\rangle\neg\varphi$ .

In addition to being the most natural operation over Kripke-like models, an update also has a straightforward epistemic interpretation: it works as an act of a public announcement [13.38, 39] or, as it will be called here, an act of *observation*. When the agent observes a given  $\psi$ , she can discard those epistemically possible worlds that fail to satisfy this formula, thereby obtaining a model with only worlds that satisfied  $\psi$  before the operation. More details on this operation and its modalities (including an axiom system) can be found in the papers [13.38, 39] or in the textbooks [13.18, 19].

### Example 13.3

Consider the model  $M$  in Example 13.1 again. Suppose the agent observes  $l$ ; this can be modeled as an update with  $l$ , which yields the following model  $M_l$



The most plausible world in  $M$  has been discarded in  $M_l$ . As explained in Example 13.2, the agent believes  $\neg l$  in  $M$ , but after the observation this is not the case anymore:  $\neg l$  does not hold in the unique most plausible world of the new model  $M_l$ . In fact,  $\neg l$  does not hold in any epistemically possible world, and thus after the observation the agent knows  $l$ ; in symbols

$$(M_l, w_2) \models Kl, \text{ that is, } (M, w_2) \models [l!]Kl.$$

### Upgrade, Also Known as Belief Revision

Another natural operation over plausibility-like models is the rearrangement of worlds within an epistemic partition. Of course, there are several ways in which a new order can be defined. The following rearrangement, taken from [13.40], is one of the many possibilities.

#### Definition 13.8 Upgrade operation

Let the tuple  $M = \langle W, \leq, V \rangle$  be a plausibility model and let  $\psi$  be a formula in  $\mathcal{L}$ . The *upgrade* operation produces the plausibility model  $M_{\psi\uparrow} = \langle W, \leq', V \rangle$ , which differs from  $M$  just in the plausibility order, given now by

$$\leq' := \{(w, u) \mid w \leq u \text{ and } (M, u) \models \psi\} \cup \{(w, u) \mid w \leq u \text{ and } (M, w) \models \neg\psi\}$$

$$\{(w, u) \mid w \sim u, (M, w) \models \neg\psi \text{ and } (M, u) \models \psi\}.$$

The new plausibility relation states that after an upgrade with  $\psi$ , all  $\psi$ -worlds become more plausible than all  $\neg\psi$ -worlds, and within the two zones the old ordering remains [13.40]. More precisely, a world  $u$  will be at least as plausible as a world  $w$ ,  $w \leq' u$ , if and only if they already are of that order and  $u$  satisfies  $\psi$ , or they already are of that order and  $w$  satisfies  $\neg\psi$ , or they are comparable,  $w$  satisfies  $\neg\psi$  and  $u$  satisfies  $\psi$ . This operation preserves the properties of the plausibility relation and hence preserves plausibility models, as shown in [13.35].

In order to describe effects of this operation within the language, an existential modality  $\langle\psi\uparrow\rangle$  is introduced for every formula  $\psi$ ,

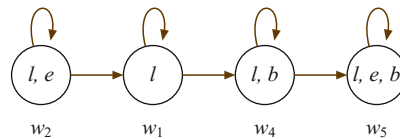
$$(M, w) \models \langle\psi\uparrow\rangle\varphi \text{ iff } (M_{\psi\uparrow, w}) \models \varphi.$$

In words, an upgrade formula  $\langle\psi\uparrow\rangle\varphi$  holds at  $(M, w)$  if and only if  $\varphi$  is the case after an upgrade with  $\psi$ . The universal modality  $[\psi\uparrow]$  is defined as the modal dual of  $\langle\psi\uparrow\rangle$ , as in the update case.

This operation also has a very natural epistemic interpretation. The plausibility relation defines the agent's beliefs, and hence any changes in the relation can be interpreted as changes in the agent's beliefs [13.34, 40, 41]. In particular, an act of *revising* beliefs after a reliable and yet fallible source has suggested  $\psi$  can be represented by an operation that puts  $\psi$ -worlds at the top of the plausibility order. Moreover, each one of the different methods to obtain a relation with former  $\psi$ -worlds at the top can be seen as a different policy for revising beliefs. Details on the operation and its modalities (including an axiom system) can be found in the papers [13.34, 40] or in the textbook [13.19].

### Example 13.4

Consider the model  $M_l$  in Example 13.3, that is, the model that results from the agent observing  $l$  at the initial model  $M$  in Example 13.1. Suppose the agent performs a belief revision toward  $b$ ; this can be modeled as an upgrade with  $b$ , which yields the following model  $(M_l)_{b\uparrow}$ :



The ordering of the worlds has changed, making those worlds that satisfy  $b$  ( $w_4$  and  $w_5$ ) more plausible than

those that do not ( $w_2$  and  $w_1$ ), keeping the old ordering with these two zones ( $w_5$  strictly above  $w_4$  and  $w_1$  strictly above  $w_2$ ). In  $M_{!}$  the agent believed  $\neg b \wedge \neg e$ , as such formula was the case in the model's unique most plausible world  $w_1$ , but this is not the case anymore in  $(M_{!})_{b\uparrow}$ : The unique most plausible world,  $w_5$ , satisfies

$b \wedge e$ , and thus the formula is part of the agent's beliefs. In symbols,

$$\begin{aligned} ((M_{!})_{b\uparrow}, w_2) \Vdash B(b \wedge e), \\ \text{that is, } (M_{!}, w_2) \Vdash [! \uparrow]B(b \wedge e). \end{aligned}$$

## 13.4 Abductive Problem and Solution

Given the intuitive definitions discussed and the formal tools introduced, it is now time to formalize the ideas.

### 13.4.1 Abductive Problem

First, the definition of what an abductive problem is.

#### Definition 13.9 Abductive problem

Let  $(M, w)$  be a pointed plausibility model, and consider  $(M_{\psi!}, w)$ , the pointed plausibility model that results from observing a given  $\psi$  at  $(M, w)$ .

A formula  $\chi$  is an abductive problem at  $(M_{\psi!}, w)$  if and only if it is known at such stage but it was not known before, that is, if and only if

$$(M_{\psi!}, w) \Vdash K\chi \quad \text{and} \quad (M, w) \Vdash \neg K\chi.$$

Equivalently, a formula  $\chi$  can become an abductive problem at  $(M, w)$  if and only if it is not known at such stage but will be known after observing  $\psi$ , that is, if and only if

$$(M, w) \Vdash \neg K\chi \wedge [\psi!]K\chi.$$

Note again how the definition of an abductive problem is relative to an agent's information at some given stage (the one represented by the pointed model).

There are two points worth emphasizing. First, note again how the definition distinguishes between the formula that becomes the abductive problem,  $\chi$ , and the formula whose observation triggers the abductive problem,  $\psi$ . Although these two formulas are typically understood to be the same ( $\chi$  becomes an abductive problem after being observed), the choice in this contribution is to distinguish between them. One reason for this is technical: Here the idea is that the agent will look for explanations of formulas that she could not have known before the observation but knows afterward. However, stating this as *the agent knows  $\chi$  after observing it* is restrictive in the DEL setting as not every formula satisfies this condition. This is because the underlying EL framework is powerful enough to talk

about the knowledge an agent has not only about facts but also about her own knowledge, and so there are formulas expressing situations such as *it is raining and you do not know it* ( $r \wedge \neg Kr$ ), which can be observed but are not known afterward (now you know that it is raining!). Another reason is, as stated earlier, generality: The described agent will be able to look for explanations not only of the formulas she can observe, but also of those that she can come to know through an observation. Still, this choice does not imply that the observed formula and the one that becomes an abductive problem are unrelated: In order for the agent to know  $\chi$  after observing  $\psi$ , she must have known  $\psi \rightarrow [\psi!] \chi$  before the action. This is nothing but the *reduction axiom* for the knowledge modality in *public announcement logic*

$$[\psi!]K\chi \leftrightarrow (\psi \rightarrow K(\psi \rightarrow [\psi!] \chi)).$$

Second, the requirements Definition 13.9 asks for  $\chi$  to be an abductive problem are not exactly the ones stated in the previous section: The sentence *there is no implication  $\eta \rightarrow \chi$  such that, before  $\chi$  became an abductive problem, the agent knew both the implication and its antecedent* has been replaced by *the agent did not know  $\chi$  before  $\chi$  became an abductive problem*. The reason is that, in DEL, the agent's knowledge and beliefs are closed under logical consequence (still, small variations of the EL framework allows the representation of nonideal agents and their abductive reasoning; see Sect. 13.8), and in such setting the two statements are equivalent: If there is an  $\eta$  such that the agent knew  $\eta \rightarrow \chi$  and  $\eta$  before  $\chi$  became an abductive problem, then clearly she knew  $\chi$  too, and if she knew  $\chi$ , then there was a  $\eta$  such that  $\eta \rightarrow \chi$  and  $\eta$  were both known, namely  $\chi$  itself. In fact, the restatement of the requirement emphasizes that it is the observation of  $\psi$  what causes the agent to know  $\chi$  and hence what creates the abductive problem.

It is worthwhile to highlight how, although the definition of an abductive problem was given in terms of the agent's knowledge, it can also be given in terms of her beliefs: It also makes sense for her to look for explanations of what she has come to believe!

“A formula  $\chi$  is said to be an abductive problem at  $(M_{\psi \uparrow}, w)$  if and only if  $(M_{\psi \uparrow}, w) \Vdash B\chi$  and  $(M, w) \Vdash \neg B\chi$ .”

With this definition, a formula is a problem if it is believed now but was not believe before a belief revision with  $\psi$ . But not only that. The agent can also face abductive problems that combine knowledge and beliefs. For example, she can face an abductive problem with  $\chi$  if she does not know the formula at some stage but believes it after a belief revision with  $\psi$ :

“A formula  $\chi$  is said to be an abductive problem at  $(M_{\psi \uparrow}, w)$  if and only if  $(M_{\psi \uparrow}, w) \Vdash B\chi$  and  $(M, w) \Vdash \neg K\chi$ .”

The stated definition allows to describe several forms of abductive problems, all of which differ in the strength of the attachment of the agent to the problematic  $\chi$  (known, strongly believed, safely believed, believed, etc.) *after* the epistemic action (update, upgrade) and the strength of her attachment to the formula *before* the action.

### 13.4.2 Classifying Problems

As mentioned, some approaches classify an abductive problem  $\chi$  according to whether  $\chi$  or  $\neg\chi$  follows from the theory: If neither  $\chi$  nor  $\neg\chi$  follows, then  $\chi$  is called a *novel* abductive problem; if  $\chi$  does not follow but  $\neg\chi$  does, then  $\chi$  is called an *anomalous* abductive problem. Given the requirement *the agent did not know  $\chi$  before  $\chi$  became an abductive problem* ( $\neg K\chi$ ) in Definition 13.9, one could suggest *the agent knew  $\neg\chi$*  ( $K\neg\chi$ ) as an alternative, but since the definition also asks for  $\chi$  to be known after the observation in order to be an abductive problem, such suggestion turns out to be too strong for propositional formulas: If  $\neg\chi$  is propositional and the agent knows it at some stage, then every epistemic possibility satisfies  $\neg\chi$ . Thus, since no *epistemic* action can change the (propositional) formula's truth value, the only way for the agent to know  $\chi$  afterward is for the action to eliminate every epistemic possibility, making  $K\varphi$  true for *every* formula  $\varphi$  and thus turning the agent inconsistent. But even though it is not possible to classify abductive problems in terms of the knowledge the agent had about the formula before the observation, it is still possible (and more reasonable) to classify them by using weaker notions, such as beliefs. Here is one possibility.

#### Definition 13.10 Expected, novel and anomalous problems

Suppose  $\chi$  is an abductive problem at  $(M_{\psi \uparrow}, w)$ . Then  $\chi$  is said to be:

- An *expected* abductive problem if and only if  $(M, w) \Vdash B\chi$
- An *novel* abductive problem if and only if  $(M, w) \Vdash \neg B\chi \wedge \neg B\neg\chi$
- An *anomalous* abductive problem if and only if  $(M, w) \Vdash B\neg\chi$ .

Many people would not call the first case an abductive problem: The observation is a confirmation rather than a surprise, and thus it does not need to trigger any further epistemic action. Nevertheless, the case shows how this proposal allows for such situations to be considered. In fact, the classification can be refined by considering further attitudes, such as the *safe beliefs* of [13.34] or the *strong beliefs* of [13.42] (both definable in  $\mathcal{L}$ ).

### 13.4.3 Abductive Solutions

An abductive solution is now to be defined. Here is a version that uses only the notion of knowledge.

#### Definition 13.11 Abductive solution

Let  $(M, w)$  be a pointed plausibility model, and consider  $(M_{\psi \uparrow}, w)$ , the pointed plausibility model that results from observing  $\psi$  at  $(M, w)$ .

If at  $(M_{\psi \uparrow}, w)$  the formula  $\chi$  is an abductive problem, then  $\eta$  is an abductive solution if and only if the agent knew that  $\eta$  implied  $\chi$  before the observation, that is, if and only if

$$(M, w) \Vdash K(\eta \rightarrow \chi).$$

Equivalently, if at  $(M, w)$  the formula  $\chi$  can become an abductive problem, then  $\eta$  will be an abductive solution if and only if the agent knows that  $\eta$  implies  $\chi$ , that is, if and only if

$$(M, w) \Vdash K(\eta \rightarrow \chi).$$

Just as in the case of abductive problem, it is also possible to define an abductive solution in terms of weaker notions as beliefs. For example, while a very strict agent would accept  $\eta$  as explanation only when  $\eta \rightarrow \chi$  was known, a less strict agent could accept it when such implication was only believed.

It is worth emphasizing that, in the stated definition, a solution for a problem  $\chi$  (at some  $M_{\psi \uparrow}$ ) is a formula  $\eta$  such that  $\eta \rightarrow \chi$  is known not when the abductive problem has arisen (at  $M_{\psi \uparrow}$ ) but rather at the stage immediately before (at  $M$ ). This is because an explanation is a piece of information that would have allowed the agent to predict the surprising observation. In fact, if

an abductive solution for a problem  $\chi$  were defined as a formula  $\eta$  such that  $\eta \rightarrow \chi$  is known once  $\chi$  is an abductive problem (at  $M_{\psi!}$ ), then every formula  $\varphi$  would be a solution since (at  $M_{\psi!}$ )  $K\chi$  would be the case (because  $\chi$  is an abductive problem) and hence so would be  $K(\varphi \rightarrow \chi)$  for every formula  $\varphi$ .

Observe also how, again in the stated definition, if  $\eta$  is a solution for the abductive problem  $\chi$  (at some  $M_{\psi!}$ ), then  $\eta$  could not be known before the observation that triggered the problem (at  $M$ ). Otherwise, both  $K(\eta \rightarrow \chi)$  and  $K\eta$  would be the case at such stage ( $M$ ) and hence, by the closure under logical consequence of knowledge in EL, so would be  $K\chi$ , contradicting the fact that  $\chi$  is an abductive problem.

#### Proposition 13.1

Let  $\chi$  be an abductive problem and  $\eta$  be one of its abductive solutions, both at  $(M_{\psi!}, w)$ . Then,  $(M, w) \Vdash \neg K\eta$ .

### 13.4.4 Classifying Solutions

It is common in the literature to classify abductive solutions according to their properties (Chap. 10). For example (Definitions 13.2 and 13.3; again, see Chap. 10), given a surprising observation  $\chi$ , an abductive solution  $\eta$  is said to be:

- *Plain* when it is a solution
- *Consistent* when it does not contradict the agent's information
- *Explanatory* when it does not explain  $\chi$  by itself.

Similar properties can be described in the present setting. To begin with, the *plain* property simply states that  $\eta$  is an abductive solution; a definition that has been already provided (Definition 13.11).

For the *consistency* property, the intuitive idea is for the solution to be compatible with the agent's information. To this end, consider the following definition.

#### Definition 13.12 Consistent solution

Let  $\chi$  be an abductive problem and  $\eta$  be one of its abductive solutions, both at  $(M_{\psi!}, w)$ . It is said that  $\eta$  is a *consistent* solution if and only if the agent considers it possible at  $(M_{\psi!}, w)$ , that is, if and only if

$$(M_{\psi!}, w) \Vdash \hat{K}\eta.$$

Thus, a solution is consistent when it is epistemically possible. Note how this requirement is given in terms of the agent's information *after* the epistemic action that triggered the abductive problem, and not

before it. In fact, there are formulas that, in a given situation, are solutions according to the stated definition, and yet not epistemically possible once the abductive problem has been raised.

#### Fact 13.1

Not every abductive solution is consistent.

*Proof:* Let  $\eta$  and  $\chi$  be propositional formulas, and take a model  $M$  in which the agent considers at least one  $(\neg\eta \wedge \neg\chi)$ -world to be epistemically possible, with the rest of the epistemic possibilities being  $(\neg\eta \wedge \chi)$ -worlds. After observing  $\chi$ ,  $\neg\chi$ -worlds will be discarded and there will be only  $(\neg\eta \wedge \chi)$ -worlds left, thus making  $\chi$  itself an abductive problem (it is not known at  $M$  but it will be known at  $M_{\chi!}$ ) and  $\eta$  an abductive solution (every epistemic possibility at  $M$  satisfies  $\eta \rightarrow \chi$ , so the agent knows this implication). Nevertheless, there are no  $\eta$ -worlds at  $M_{\chi!}$ , and therefore  $\hat{K}\eta$  is false at such stage. ■

The *explanatory* property is interesting. The idea in the classic setting is to avoid solutions that imply the problematic  $\chi$  per se, such as  $\chi$  itself or any formula logically equivalent to it. In the current epistemic setting, this idea can be understood in a different way: A solution  $\eta$  is explanatory when the acceptance of  $\eta$  (which, as discussed, will be modeled via belief revision; see Sect. 13.6) changes the agent's information, that is, when the agent's information is different from  $(M_{\psi!}, w)$  to  $((M_{\psi!})_{\eta\uparrow}, w)$  (the model that results after integrating the solution  $\eta$ ). This assertion could be formalized by stating that the agent's information is the same in two pointed models if and only if the agent has the same knowledge in both, but this would be insufficient: The model operation representing an act of belief revision (the upgrade of Definition 13.8) is devised to change only the agent's beliefs (although certain knowledge, such as knowledge about beliefs, might also change). A second attempt would be to state that the agent's information is the same in two pointed models if and only if they coincide in the agent's knowledge and beliefs, but the mentioned operation can change a model without changing the agent's beliefs.

Within the current *modal* epistemic logic framework, a more natural way of specifying the idea of an agent having the same information in two models is via the notion of bisimulation.

#### Definition 13.13 Bisimulation

Let  $P$  be a set of atomic propositions and let  $M = \langle W, \leq, V \rangle$  and  $M' = \langle W', \leq', V' \rangle$  be two plausibility models based on this set. A nonempty relation  $Z \subseteq (W \times W')$  is called a *bisimulation* between  $M$  and  $M'$  (notation:  $M \leftrightarrow_Z M'$ ) if and only if, for every  $(w, w') \in Z$ :

- $V(w) = V'(w')$ , that is,  $w$  and  $w'$  satisfy the same atomic propositions
- If there is a  $u \in W$  such that  $w \leq u$ , then there is a  $u' \in W'$  such that  $w' \leq' u'$  and  $Zuu'$
- If there is a  $u' \in W'$  such that  $w' \leq' u'$ , then there is a  $u \in W$  such that  $w \leq u$  and  $Zuu'$ .

Two models  $M$  and  $M'$  are *bisimilar* (notation:  $M \leftrightarrow M'$ ) when there is a bisimulation between them, and two pointed models  $(M, w)$  and  $(M', w')$  are bisimilar (notation:  $(M, w) \leftrightarrow (M', w')$ ) when there is a bisimulation between  $M$  and  $M'$  containing the pair  $(w, w')$ .

This notion is significant because, under image-finiteness (a plausibility model is *image-finite* if and only if every world can  $\leq$ -see only a finite number of worlds), it characterizes modal equivalence, that is, it characterises models that satisfy exactly the same formulas in the modal language.

**Theorem 13.1**

Let  $P$  be a set of atomic propositions and let  $M = \langle W, \leq, V \rangle$  and  $M' = \langle W', \leq', V' \rangle$  be two image-finite plausibility models. Then  $(M, w) \leftrightarrow (M', w')$  if and only if, for every formula  $\varphi \in \mathcal{L}$ ,  $(M, w) \models \varphi$  iff  $(M', w') \models \varphi$ .

Now it is possible to state a formal definition of what it means for a solution to be explanatory.

**Definition 13.14 Explanatory solution**

Let  $\chi$  be an abductive problem and  $\eta$  be one of its abductive solutions, both at  $(M_{\psi!}, w)$ . It is said that  $\eta$  is an *explanatory* solution if and only if its acceptance changes the agent’s information, that is, if and

only if there is *no* bisimulation between  $(M_{\psi!}, w)$  and  $((M_{\psi!})_{\eta\uparrow}, w)$ .

This definition, devised in order to avoid solutions that explain the abductive problem per se, has pleasant side effects. In the abductive reasoning literature, a solution is called *trivial* when it is logically equivalent to the abductive problem  $\chi$  (i.e., when it is not explanatory) or when it is a contradiction (to the agent’s knowledge, or a logical contradiction). Under the given definition, every trivial solution is not explanatory: *Accepting any such solution will not change the agent’s information*. The reason is that, in both cases, the upgrade operation *will not make any change in the model*: In the first case because, after the observation, the agent knows the abductive problem formula, and hence every epistemically possible world satisfies it (as well as every formula logically equivalent to the problem); in the second case because *no* epistemically possible world satisfies it. In this way, this framework characterizes trivial solutions not in terms of their form, as is typically done, but rather in terms of their effect: *Accepting them will not give the agent any new information*.

In particular, this shows how the act of incorporating a contradictory explanation will not make the agent *collapse* and turn into someone that knows and believes everything, as happens in traditional approaches; thus, a logic of formal inconsistency (e.g., [13.43]; see also Chap. 15) is not strictly necessary. This is a consequence of two simple but powerful ideas:

1. Distinguishing an agent’s different epistemic attitudes
2. Assimilating an abductive solution not as knowledge, but rather as a belief.

## 13.5 Selecting the Best Explanation

Finding suitable and reasonable criteria for selecting the best explanation is a fundamental problem in abductive reasoning [13.32, 33], and in fact many authors consider this to be the heart of the subject. The so-called *thesis of purpose*, stated in [13.33], establishes that the aim of scientific abduction is:

1. To generate new hypotheses
2. To select hypotheses for further examination and testing.

Hence a central issue in scientific abduction is to provide methods for selecting. Because the true state of the world is unknown, selecting the best explanation requires more than just consistency with the available

information, and there are many proposals of what these extra criteria should be.

Some approaches are based on probabilistic measurements [13.44–46]. Even Sherlock Holmes advised that, in order to evaluate explanations, one should “balance probabilities and choose the most likely” (*The Hound of the Baskervilles*), but unfortunately explanations rarely come equipped with probabilities.

In abductive logic programming, a common strategy is to look for abductive solutions at the *dead ends* of prolog proofs [13.47]. Sound and complete procedures can be defined also by using stable models and answer sets [13.48, 49]. Apart from selection criteria based on consistency and integrity constraints, it is common to

start with a set of *abducible* predicates and select explanations built only from ground atoms using them (see Chap. 10 for more details on abductive logic programming).

There are also approaches that use logical criteria, but beyond the already mentioned requisites to avoid triviality, the definition of suitable criteria is still an open problem. One of the most pursued ideas is that of minimality, a concept that can be understood syntactically (e.g., [13.3] and [13.5] look for literals), semantically (a minimal explanation is equivalent to any other explanation it implies [13.1]), with respect to the set of possible explanations (the best explanation is the weakest, i.e., the one that is implied by the rest of them), and even with respect to the current information (the best explanation is the one that disrupt less the current information).

In fact, most logical criteria are based on restrictions on the logical form of the solutions but, as mentioned in [13.1], finer criteria to select between two equally valid solutions require contextual aspects. With this idea in mind some approaches have proposed to use an ordering among formulas [13.10, 50, 51] or among full theories (i.e., possible worlds [13.52, 53]). In particular, for the latter, a common option is the use of *preferential models* (e.g., [13.54]) in which preferential criteria for selecting the best explanation are regarded as qualitative properties that are beyond the pure causal or deductive relationship between the abductive problem and its abductive solution. But these preference criteria are normally treated as an external device, which works on top of the logical or deductive part of the explanatory mechanism, and thus it has been criticized because it seems to fall outside a logical framework.

The epistemic approach of this proposal provides with an interesting alternative. The concepts of an abductive problem and an abductive solution have been defined in terms of the agent's epistemic attitudes, so it is natural to use such attitudes as a criterion for selecting the best explanation. Consider, for instance, the following elaboration of an example presented in Chap. 10.

“Mary and Gaby arrive late to Mary's apartment; the light switch is pressed but the light does not turn on. Knowing that the apartment is old, Mary assumes a failure in the electric line as the explanation for the light not turning on. Gaby, on the other hand, does not have any information about the apartment, so she explains the light not turning on by assuming that the bulb is burned out.”

After pressing the switch, both Mary and Gaby observe that the light does not turn on. There are several explanations for this: It is possible that the electric line failed, as Mary assumed, but it can also be the case

that the bulb is burned out, as Gaby thinks, and it is even possible that the switch is faulty. Then, why do they choose a different explanation? The reason is that, though they both observe that the light does not turn on, they have different background information: Mary knows that the apartment is old, and hence she considers a failure in the electric line more likely than any other explanation, but Gaby does not have that piece of information, so for her a burned out bulb explains the lack of light better.

The example shows that, even when facing the same surprising observation (the light does not turn on), agents with different knowledge and beliefs may choose a different best explanation: While Mary assumes that the electric line has failed, Gaby thinks that the bulb is burned out. Both explanations are equally *logical* since either a failure on the electric line or else a burned out bulb is enough to explain why the light does not turn on. What makes Mary to choose the first and Gaby the second is that they have different knowledge and different beliefs. This suggests first, that, instead of looking for criteria to select *the* best explanation, the goal should be a criteria to select *the agent's* best explanation.

But there is more. The explanation an agent will choose for a given abductive problem depends not only on how the problematic formula could have been predicted, but also on what the agent herself knows and what she considers more likely to be the case. It could be argued that this criterion is not *logical in the classical sense* because it is not based exclusively on the *deductive* relationship between the observed fact and the different ways in which it could have been derived. Nevertheless, it is *logical in a broader sense* since it does depend on the agent's information: her knowledge and her beliefs. In particular, in the plausibility models framework, the agent's knowledge and beliefs are defined in terms of a plausibility relation among epistemic possibilities, so it is natural to use precisely this relation as a criterion for selecting each agent's best explanation(s).

This section presents a straightforward use of this idea. It discusses how the plausibility order among epistemic possibilities can be lifted to a plausibility order among formulas, thus providing a natural criterion to select the agent's best explanation. A generalization of this idea that works instead with all explanations will be discussed later (Sect. 13.7).

### 13.5.1 Ordering Explanations

A plausibility model provides an ordering among possible worlds. This order can be lifted to get an ordering among set of worlds, that is, an ordering among formulas of the language (with each formula seen as the set

of those worlds that make it true). The different ways in which such ordering can be defined has been studied in *preference logic* (see [13.55–57] or, for a more detailed exposition [13.58, Chap. 3.3]); this section recalls the main ideas, showing how they can be applied to the task of selecting the best explanation in abductive reasoning.

In general, an ordering among objects can be lifted to an ordering among sets of such objects in different ways. For example, one can say that the lifted ordering puts the set of objects satisfying the property  $\psi$  (the set of  $\psi$ -objects) over the set of objects satisfying the property  $\varphi$  (the set of  $\varphi$ -objects) when *there is a  $\psi$ -object that the original ordering among objects places above some  $\varphi$ -object* (a  $\exists\exists$  preference of  $\psi$  over  $\varphi$ ; see below). But one can be more drastic and say that the set of  $\psi$ -objects is above the set of  $\varphi$ -ones when the original ordering places *every  $\psi$ -object above every  $\varphi$ -one* (a  $\forall\forall$  preference of  $\psi$  over  $\varphi$ ). This quantification combination gives raise to the following possibilities

$\varphi \leq_{\exists\exists} \psi$	iff	<i>there is a <math>\varphi</math>-object <math>w</math> and there is a <math>\psi</math>-object <math>u</math> such that <math>w \leq u</math></i>
$\varphi \leq_{\forall\exists} \psi$	iff	<i>for every <math>\varphi</math>-object <math>w</math> there is a <math>\psi</math>-object <math>u</math> such that <math>w \leq u</math></i>
$\varphi \leq_{\forall\forall} \psi$	iff	<i><math>w \leq u</math> for every <math>\varphi</math>-object <math>w</math> and every <math>\psi</math>-object <math>u</math></i>
$\varphi \leq_{\exists\forall} \psi$	iff	<i>there is a <math>\varphi</math>-object <math>w</math> such that <math>w \leq u</math> for every <math>\psi</math>-object <math>u</math></i>

The first two orderings can be defined within the language  $\mathcal{L}$

$$\begin{aligned}\varphi \leq_{\exists\exists} \psi &:= \langle \sim \rangle (\varphi \wedge \langle \leq \rangle \psi) \\ \varphi \leq_{\forall\exists} \psi &:= [\sim] (\varphi \rightarrow \langle \leq \rangle \psi).\end{aligned}$$

The first formula indicates that there is a  $\psi$ -world that is at least as plausible as a  $\varphi$ -one,  $\varphi \leq_{\exists\exists} \psi$ , exactly when there is an epistemic possibility that satisfies  $\varphi$  and that can see an at least as plausible  $\psi$ -world. The second one only changes the first quantification (turning, accordingly, the conjunction into an implication): For every  $\varphi$ -world there is a  $\psi$ -world that is at least as plausible.

The last two orderings are not immediate. Given the formulas for the previous two orderings, one could propose  $[\sim](\varphi \rightarrow \langle \leq \rangle \psi)$  for the  $\forall\forall$  case, but this formula is not correct: It states that every world that is at least as plausible as any  $\varphi$ -world satisfies  $\psi$ , but it does not guarantee that *every  $\psi$ -world is indeed above every  $\varphi$ -world*:

1. There might be a  $\psi$ -world incomparable to some  $\varphi$ -one, and even if all worlds are comparable

2. There might be a  $\psi$ -world strictly below a  $\varphi$ -one ( $<$ , the strict version of  $\leq$ , is defined as  $w < u$  if and only if  $w \leq u$  and *not*  $u \leq w$ ).

The plausibility order is locally connected (i. e., inside each epistemic partition, every world is comparable to each other) so (1) cannot occur. Thus, a formula defining  $\leq_{\forall\forall}$  only needs to guarantee that no  $\psi$ -world is *strictly* below a  $\varphi$ -one; in other words, it needs to express that, given any  $\psi$ -world, every world that is strictly more plausible satisfies  $\neg\varphi$ . Such formula can be easily stated in a language that extends  $\mathcal{L}$  with a standard modality for the relation  $<$

$$\varphi \leq_{\forall\forall} \psi := [\sim](\psi \rightarrow \langle < \rangle \neg\varphi).$$

Finally, the  $\exists\forall$  ordering presents a similar situation. Following the first two cases one could propose  $\langle \sim \rangle (\varphi \wedge \langle < \rangle \psi)$ , but such formula is not appropriate, even in the current full-comparability case: It holds even when there are  $\psi$ -worlds below the chosen  $\varphi$ -one. In order to guarantee the existence of a  $\varphi$ -world that is at most as plausible as every  $\psi$ -world, the formula should state that every world that is strictly *less* plausible than the  $\varphi$ -world satisfies  $\neg\psi$ . Extending the language again, this time with a modality for  $>$ , makes such formula straightforward

$$\varphi \leq_{\exists\forall} \psi := \langle \sim \rangle (\varphi \wedge \langle > \rangle \neg\psi).$$

All in all, the important fact is that among these four orderings on sets of worlds (i. e., formulas), two are definable within  $\mathcal{L}$  and the other two only need simple extensions. This shows how the plausibility order among worlds that defines the agent's knowledge and beliefs (Sect. 13.3.1) also defines plausibility orderings among formulas (sets of worlds), and hence provides a criterion for selecting the best abductive solution *for a given agent*. It will now be shown how this criterion can be used, and how it leads to situations in which agents with different knowledge and beliefs choose different best explanations.

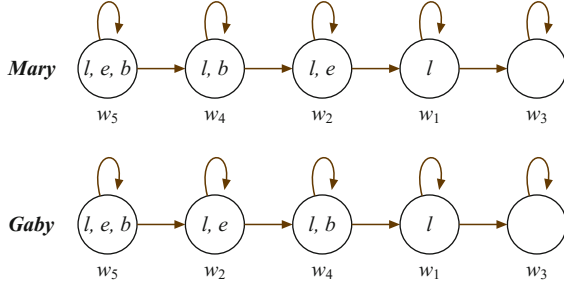
### Example 13.5

Recall Mary and Gaby's example. Both observe that after pressing the switch the light does not turn on, but each one of them chooses a different explanation: While Mary assumes that the electric line failed, Gaby thinks that the bulb is burned out. As it has been argued, the reason why they choose different explanations is that they have different knowledge and beliefs. Here is a formalization of the situation.

The following plausibility models show Mary and Gaby's knowledge and beliefs before pressing the



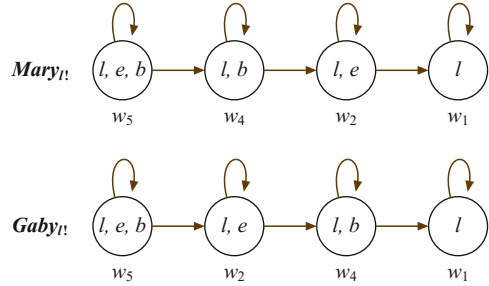
switch. They involve tree atomic propositions:  $l$  standing for lack of light,  $e$  standing for a failure in the electric line and  $b$  standing for a burned out bulb. Again, each possible world has indicated within it exactly those atomic propositions that are true in each one of them, and the arrows represent the plausibility relation (transitive arrows are omitted).



Observe how both Mary and Gaby know that both a failure on the electric line and a burned out bulb imply lack of light (both  $e \rightarrow l$  and  $b \rightarrow l$  hold in every world). In fact, the only difference in the models is the plausibility order between worlds  $w_2$  and  $w_4$ . Mary knows that the apartment is old so she considers a failure on the line ( $e$ ) more likely than a burned out bulb ( $b$ ), and hence the situation where the electric line fails but the bulb is not burned out ( $w_2$ ) is more likely than its opposite ( $w_4$ ). Gaby, on the other hand, does not know anything about the apartment, and hence for her a burned out bulb with a working electric line ( $w_4$ ) is more plausible than a working bulb and a failing electric line ( $w_2$ ). It is also assumed that, for both of them, the most likely possibility is the one in which everything works correctly ( $w_1$ ) and the least plausible case is the one in which everything fails ( $w_5$ ).

After they both observe that pressing the switch does not turn on the light, the unique world where  $l$  is

not the case,  $w_3$ , is eliminated, thus producing the following models.



As a result of the observation, Mary and Gaby know that there is no light ( $Kl$  holds in both models), something that they did not know before. Thus, following Definition 13.9, both have an abductive problem with  $l$ .

According to Definition 13.11, both  $e$  and  $b$  are abductive solutions for the abductive problem  $l$  for both Mary and Gaby: Both formulas are the antecedent of implications that have  $l$  as a consequent and that were known before the observation. So, how can each girl choose her own best explanation? For Mary, the unique ordering that puts  $b$  above  $e$  is the weakest one,  $\exists\exists$  (there is a  $b$ -world,  $w_4$ , at least as plausible as a  $e$ -one,  $w_5$ ). Nevertheless, from her point of view,  $e$  is above  $b$  not only in the weak  $\exists\exists$  way ( $w_2$  is at least as plausible as  $w_4$ ) but also in the stronger  $\forall\exists$  way (every  $b$ -world has a  $e$ -world that is at least as plausible as it). Thus, one can say that  $e$  is a more plausible explanation from Mary's perspective. In Gaby's case something analogous happens:  $b$  is above  $e$  not only in the weak  $\exists\exists$  way ( $w_4$  is at least as plausible as  $w_2$ ) but also in the strong  $\forall\exists$  way. Hence, it can be said that, for Gaby,  $b$  is the best explanation.

### 13.6 Integrating the Best Solution

Once the agent has selected the best explanation for her, she can incorporate it into her information. As discussed in Sect. 13.2, even though the nonmonotonic nature of abductive reasoning indicates that an abductive solution should not be assimilated as knowledge, the richness of the present framework allows the possibility to integrate it as a part of the agent's beliefs. Here is a modality describing such action.

**Definition 13.15 Modality for abductive reasoning**  
Let  $(M, w)$  be a pointed plausibility model and consider again  $(M_{\psi!}, w)$ , the pointed plausibility model

that results from observing  $\psi$  at  $(M, w)$ . Every pair of formulas  $\eta$  and  $\chi$  in  $\mathcal{L}$  define an existential modality  $\langle \text{Abd}_{\eta}^{\chi} \rangle \varphi$ , read as *the agent can perform an abductive step for  $\chi$  with  $\eta$  after which  $\varphi$  is the case*, and whose semantic interpretation is as follows

$$\begin{aligned}
 (M_{\psi!}, w) \models \langle \text{Abd}_{\eta}^{\chi} \rangle \varphi & \\
 \text{iff} & \\
 (1) (M_{\psi!}, w) \models K\chi \text{ and } (M, w) \models \neg K\chi, & \\
 (2) (M, w) \models K(\eta \rightarrow \chi), \text{ and} & \\
 (3) ((M_{\psi!})_{\eta \uparrow}, w) \models \varphi. &
 \end{aligned}$$

Equivalently,  $\langle \text{Abd}_\eta^x \rangle \varphi$ 's semantic interpretation can be defined as

$$\begin{aligned} (M_{\psi!}, w) \Vdash \langle \text{Abd}_\eta^x \rangle \varphi \\ \text{iff} \\ (M, w) \Vdash \neg K\chi \wedge K(\eta \rightarrow \chi) \wedge [\psi!](K\chi \wedge [\eta \uparrow]\varphi). \end{aligned}$$

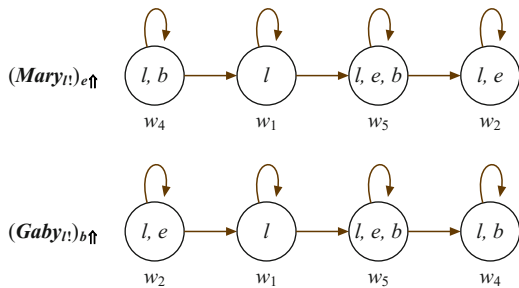
The definition states that  $\langle \text{Abd}_\eta^x \rangle \varphi$  is true at  $(M_{\psi!}, w)$  if and only if:

1.  $\chi$  is an abductive problem at  $(M_{\psi!}, w)$
2.  $\eta$  is an abductive solution also at  $(M_{\psi!}, w)$
3. An upgrade (Definition 13.8) with  $\eta$  will make  $\varphi$  true.

The last part makes precise the idea of how an agent should incorporate the selected explanation: It cannot be incorporated as knowledge, but it can be incorporated as a belief.

**Example 13.6**

Returning to Example 13.5, once Mary and Gaby have selected their respective best explanation, they can perform an abductive step. In Mary's case, worlds satisfying  $e$  ( $w_5$  and  $w_2$ ) will become more plausible than worlds that do not satisfy it ( $w_4$  and  $w_1$ ); in Gaby's case, worlds satisfying  $b$  ( $w_5$  and  $w_4$ ) will become more plausible than worlds that do not satisfy it ( $w_2$  and  $w_1$ ). Applying these upgrades to the models  $Mary_{\eta!}$  and  $Gaby_{\eta!}$  produces the following models.

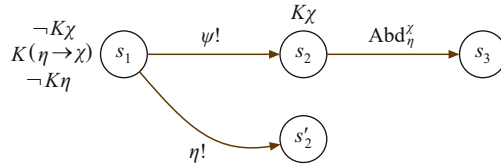


As a result of the abductive step, each agent believes her own explanation: Mary believes that the electric line has failed ( $e$  is true in her unique most plausible world  $w_2$ ), and Gaby believes that the bulb is burned out ( $b$  is true in her unique most plausible world  $w_4$ ). That is, for every  $w \in \{w_1, w_2, w_4, w_5\}$ ,

$$\begin{aligned} (Mary_{\eta!}, w) \Vdash \langle \text{Abd}_e^l \rangle Be \\ (Gaby_{\eta!}, w) \Vdash \langle \text{Abd}_b^l \rangle Bb. \end{aligned}$$

**13.6.1 Abduction in a Picture, Once Again**

The definitions that have been provided allow more precision in the diagram of abductive reasoning presented in Sect. 13.2.5. Here is the updated version for the case in which the definitions are given just in terms of the agent's knowledge. Note how the *inferring*  $\chi$  step has been dropped, as it is not needed in an omniscient setting such as DEL. Again, circles represent the agent's epistemic states (i. e., full plausibility models) and arrows are labeled with the operations that modify the agent's information.



Again, the upper path represents what really happened. After observing  $\psi$ , the agent reaches the epistemic state  $s_2$  in which she knows  $\chi$ . But before the observation, at  $s_1$ , she did not know  $\chi$ , and thus this formula is an abductive problem at  $s_2$ . Observe how  $\eta \rightarrow \chi$  was known at  $s_1$ : hence,  $\eta$  is an abductive solution at  $s_2$  and the agent can perform an abductive step with it to reach state  $s_3$ . This abductive solution  $\eta$  would have helped the agent to infer (and hence to come to know)  $\chi$ , and the lower path represents this alternative situation. In general, it cannot be guaranteed that the agent would have known  $\chi$  (or even  $\eta$ ) at state  $s'_2$ : these formulas could have had epistemic modalities, and hence the observation could have changed their truth value. However, if both formulas are propositional,  $K\chi$  and  $K\eta$  hold at  $s'_2$ .

**13.6.2 Further Classification**

Section 13.4.4 presented an epistemic version of the the common classification of abductive solutions. But the current DEL setting allows further possibilities and hence a finer classification. For example, here are two straightforward ideas. First, a solution  $\eta$  has been defined as the antecedent of an implication that has  $\chi$  as a consequent and that was known *before* the epistemic action that triggered the problem. Nevertheless, given that both formulas might contain epistemic operators, the agent can go from knowing the implication to not knowing it. Second, it has been stated that the agent incorporates the selected explanation via a belief revision (i. e., an upgrade). Nevertheless, since the solution might contain epistemic operators, the upgrade does not guarantee that the agent will believe the solution after the operation.

**Definition 13.16 Adequate solution and successful solution**

Let the formula  $\eta$  be an abductive solution for the abductive problem  $\chi$  at  $(M_{\psi!}, w)$ . Then:

- $\eta$  is an *adequate* solution if and only if the agent still knows  $\eta \rightarrow \chi$  at  $(M_{\psi!}, w)$ , that is, if and only if  $(M_{\psi!}, w) \Vdash K(\eta \rightarrow \chi)$ .
- $\eta$  is a *successful* solution if and only if it is believed after the abductive step, that is, if and only if  $(M_{\psi!}, w) \Vdash \langle \text{Abd}_{\eta}^{\chi} \rangle B\eta$ .

Here it is a result about the adequacy property.

**Proposition 13.2**

Every abductive solution is *adequate*.

*Proof:* More precisely, suppose that at  $(M_{\psi!}, w)$  the formula  $\chi$  is an abductive problem and  $\eta$  is one of its abductive solutions. Since  $\chi$  is an abductive problem,  $(M_{\psi!}, w) \Vdash K\chi$  and hence  $(M_{\psi!}, w) \Vdash K(\eta \rightarrow \chi)$ . ■

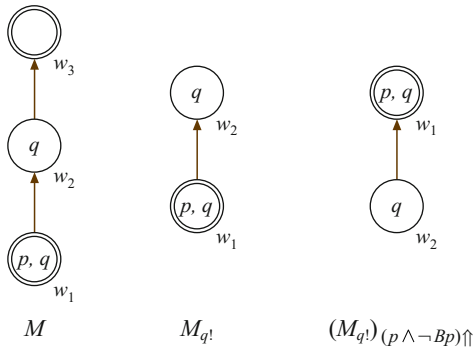
Given this result, this property is of little interest in the current setting. However, it becomes interesting in settings with nonomniscient agents. In such frameworks, it is possible for the agent not to know  $\eta \rightarrow \chi$  even when she knows  $\chi$  and she knew  $\eta \rightarrow \chi$  before.

Here it is another result, now about the property of being a successful solution.

**Fact 13.2**

Not every abductive solution is successful.

*Proof:* Let  $P = \{p, q\}$  be the set of atomic propositions, and consider the pointed plausibility models below (reflexive and transitive arrows omitted) in which the evaluation points are double circled.



Observe how  $q$  is an abductive problem at  $M_{q!}$  since it is not known at  $M$  (there is an epistemically possible world where  $q$  fails, namely,  $w_3$ ) but it is known at  $M_{q!}$ . Observe also how  $p \wedge \neg Bp$  is an abductive solution since  $K((p \wedge \neg Bp) \rightarrow q)$  holds at  $M$  (it is true at  $w_1$  and  $w_2$  because  $q$  is true in those worlds, and also true at  $w_3$  because  $p \wedge \neg Bp$  fails in this world). Furthermore,  $p \wedge \neg Bp$  is a consistent solution since it is epistemically possible in  $M_{q!}$  ( $p$  and  $\neg Bp$  are both true at  $w_1$ , the latter because there is a most plausible world,  $w_2$ , where  $p$  is not the case, and hence the agent does not believe  $p$ ). Nevertheless, after an upgrade with  $p \wedge \neg Bp$  this very formula is not believed. It fails at the unique most plausible world  $w_1$  because  $\neg Bp$  fails at it: the most plausible world ( $w_1$  itself) satisfies  $p$  and hence the agent now believes  $p$ , that is,  $Bp$  is the case. ■

Nevertheless, if a propositional solution  $\eta$  is also consistent, then it is successful.

**Proposition 13.3**

Suppose that at  $(M_{\psi!}, w)$  the formula  $\eta$  is an abductive solution for the abductive problem  $\chi$ . If  $\eta$  is a propositional and consistent solution, then it is successful.

*Proof:* If  $\eta$  is a consistent solution, then at  $(M_{\psi!}, w)$  there is at least one epistemically possible  $\eta$ -world. Therefore, an upgrade with  $\eta$  will put worlds that satisfied  $\eta$  in  $(M_{\psi!}, w)$  on top of the plausibility order. Now,  $\eta$  is propositional, and hence its truth value depends only on the valuation of each possible world; since the upgrade operation does not affect the valuation, then any world satisfying  $\eta$  in  $M_{\psi!}$  will still satisfy it in  $(M_{\psi!})_{\eta\uparrow}$ . Hence, after the operation, the most plausible worlds will satisfy  $\eta$ , and thus  $((M_{\psi!})_{\eta\uparrow}, w) \Vdash B\eta$  will be the case. This, together with the fact that at  $M_{\psi!}$  the formula  $\chi$  is an abductive problem and the formula  $\eta$  is an abductive solution, yield  $(M_{\psi!}, w) \Vdash \langle \text{Abd}_{\eta}^{\chi} \rangle B\eta$ . ■

It has been already stated that a solution is explanatory when it changes the agent's information. A further classification of abductive solutions can be provided according to *how much* they change the agent's information, that is, according to the attitude of the agent toward the solution *before* it was incorporated.

**Definition 13.17**

Suppose that  $\chi$  is an abductive problem at  $(M_{\psi!}, w)$ . An explanatory abductive solution  $\eta$  is:

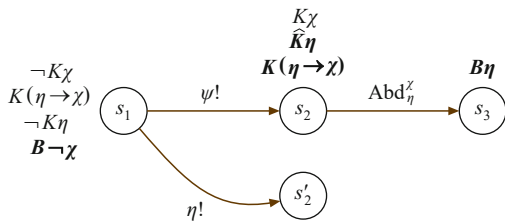
- *Weakly explanatory* when  $(M_{\psi!}, w) \Vdash B\eta$

- Neutral when  $(M_{\psi!}, w) \models \neg B\eta \wedge \neg B\neg\eta$
- Strongly explanatory when  $(M_{\psi!}, w) \models B\neg\eta$ .

Again, there are more possibilities if further epistemic attitudes are considered.

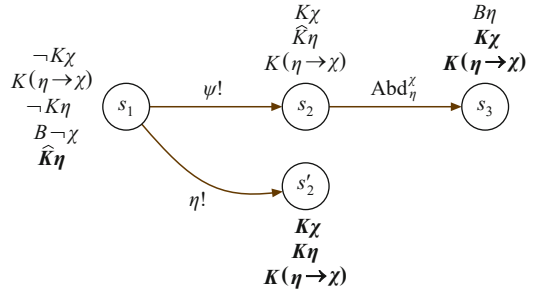
### 13.6.3 Properties in a Picture

Consider an anomalous abductive problem  $\chi$  (i. e.,  $B\neg\chi$  holds at  $s_1$ ) whose abductive solution  $\eta$  is consistent ( $\hat{K}\eta$  holds at  $s_2$ ) and successful ( $B\eta$  holds at  $s_3$ ), recalling also that every solution is adequate (so  $K(\eta \rightarrow \chi)$  holds at  $s_2$ ). This extends the diagram of Sect. 13.6.1 in the following way.



Moreover, consider the case in which both  $\chi$  and  $\eta$  are propositional, the typical case in abductive reasoning in which the agent looks for explanations of facts, and not of her own (or, in a multiagent setting, of other agents') epistemic state. First, in such case,  $\eta$  should be an epistemic possibility not only at  $s_2$  but also at  $s_1$ . But not only that; it is possible now to state the effects of the abductive step at  $s_2$  (the agent will believe  $\eta$  and will still know  $\eta \rightarrow \chi$ ) and of the hypothetical announcement of

$\eta$  at  $s_1$  (she would have known both  $\eta$  and  $\chi$ , and she would have still known  $\eta \rightarrow \chi$ ). Therefore,



This diagram beautifully illustrates what lies behind this proposal's understanding of abductive reasoning. In the propositional case, if  $\eta$  is a consistent and successful abductive solution for the abductive problem  $\chi$ , then, after abductive reasoning, the agent will know  $\chi$  and will believe  $\eta$ . In fact, when the observed formula  $\psi$  is actually the same  $\chi$  that becomes an abductive problem, the epistemic effect of abductive reasoning, from knowledge to beliefs, can be described with the following validity [13.59],

$$K(\eta \rightarrow \chi) \rightarrow [\chi!](K\chi \rightarrow \langle \text{Abd}_\eta^\chi \rangle B\eta) .$$

What makes  $\eta$  a reasonable solution is the existence of an *alternative reality* in which she observed  $\eta$  and, thanks to that, came to know  $\chi$ . Similar diagrams can be obtained for the cases in which the definitions of an abductive problem and an abductive solution are given in terms of epistemic attitudes other than knowledge.

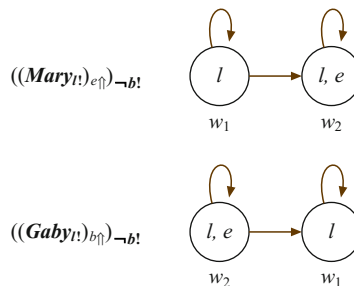
## 13.7 Working with the Explanations

The reason why abductive solutions are incorporated as beliefs and not as knowledge is because the selected explanation (in fact, *any* explanation) is just a hypothesis, subject to change in light of further information. Consider the following continuation of the Mary and Gaby's situation.

### Example 13.7

After their respective abductive steps (models  $(Mary!_l)_{e\uparrow}$  and  $(Gaby!_l)_{b\uparrow}$  of Example 13.6), Mary and Gaby take a closer look at the bulb and observe that it is not burned out ( $\neg b$ ). Semantically this is simply an observation operation that eliminates  $w_4$  and  $w_5$ , exactly those epistemic possibilities where the bulb is burned out (i. e., where  $b$  holds). The resulting models

are the following.



This observation does not affect Mary's explanation: She still believes that the electric line has failed ( $e$  is true in her unique most plausible world  $w_2$ ). But Gaby's

case is different: She does not have an explanation for  $l$  anymore. Although she knows it ( $Kl$  holds at the model on the bottom, that is,  $l$  is true in every epistemic possibility), she neither knows nor believes the antecedent of a known implication with  $l$  as a consequent (besides, of course, the trivial ones); she needs to perform a further abductive step in order to explain it.

There is, nevertheless, a way to avoid the extra abductive reasoning step. Recall that after applying the defined upgrade operation (Definition 13.8), all the worlds satisfying the given formula become more plausible than the ones that do not satisfy it, *and within the two zones the old ordering remains*. If the *lifted* worlds are not those that satisfy the agent's most plausible explanation but rather those that satisfy *at least one* of her explanations, the resulting model will have two layers: the lower one with worlds that do not satisfy any explanation, and the upper one with worlds that satisfy at least one. But inside the upper layer the old ordering will remain. In other words, the most plausible worlds in the resulting model (i. e., the most plausible ones in the upper layer) will be the ones that satisfy at least one explanation and that were already more plausible than the rest. Thus, with respect to the most plausible explanation, the same result is achieved: After such upgrade, roughly, the agent will believe the explanation that was the most plausible for her.

The difference with respect to the approach of the previous section is that the worlds that appear below the most plausible ones are not arbitrary. Worlds on the second best layer satisfy already some explanation; an explanation that was not chosen because it was not the most plausible one. Then, if further observations make the original best explanation obsolete, once that the correspondent (and now also obsolete) worlds have been discarded, the ones that will be at the top of the plausibility ordering will be the previously second best. Thus, an explanation will be already present and no further abductive steps will be needed.

### 13.7.1 A Modality

The idea just described is formalized now by introducing a modality that, given an abductive problem  $\chi$ , upgrades those worlds that satisfy at least one of its abductive explanations.

#### Definition 13.18 Modality for formula-based abduction

Let  $(M, w)$  be a pointed plausibility model and consider again  $(M_{\psi!}, w)$ , the pointed plausibility model that results from observing  $\psi$  at  $(M, w)$ . Every formula  $\chi$  in

$\mathcal{L}$  defines an existential modality of the form  $\langle \text{Abd } \chi \rangle \varphi$ , read as *the agent can perform a complete abductive step for  $\chi$  after which  $\varphi$  is the case*, and whose semantic interpretation is as follows

$$\begin{aligned} (M_{\psi!}, w) \Vdash \langle \text{Abd } \chi \rangle \varphi \\ \text{iff} \\ (1) (M_{\psi!}, w) \Vdash K\chi \text{ and } (M, w) \Vdash \neg K\chi, \\ (2) ((M_{\psi!})_{(\bigvee \Sigma_\chi) \uparrow}, w) \Vdash \varphi, \end{aligned}$$

where  $\Sigma_\chi$  is the set of abductive solutions for  $\chi$ , that is,

$$\Sigma_\chi := \{\eta \mid (M, w) \Vdash K(\eta \rightarrow \chi)\}.$$

Equivalently,  $\langle \text{Abd } \chi \rangle \varphi$ 's semantic interpretation can be defined as

$$\begin{aligned} (M_{\psi!}, w) \Vdash \langle \text{Abd } \chi \rangle \varphi \\ \text{iff} \\ (M, w) \Vdash \neg K\chi \wedge [\psi!](K\chi \wedge [\bigvee \Sigma_\chi \uparrow] \varphi). \end{aligned}$$

The correspondent universal modality,  $[\text{Abd } \chi]$ , is defined as usual.

The definition states that  $\langle \text{Abd } \chi \rangle \varphi$  is true at  $(M_{\psi!}, w)$  if and only if (1)  $\chi$  is an abductive problem at  $(M_{\psi!}, w)$ , and (2) an upgrade with  $\bigvee \Sigma_\chi$  will make  $\varphi$  true. The last part makes precise the idea of working with all the solutions:  $\Sigma_\chi$  contains all abductive solutions for  $\chi$ , so  $\bigvee \Sigma_\chi$  is a disjunction characterising those worlds that satisfy at least one of them and hence an upgrade with it will move such worlds to the topmost layer. But inside this layer, the former plausibility order will persist, and hence worlds at the top of it will be precisely those that satisfy at least one solution for  $\chi$  and, among them, were already the most plausible ones.

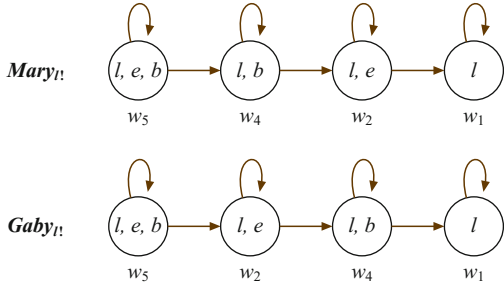
#### Remark 13.1

The set  $\Sigma_\chi$  contains, among others,  $\chi$ ,  $\chi \wedge \chi$ , and so on, and hence  $\bigvee \Sigma_\chi$  is an infinite disjunction. Syntactic restrictions can be imposed in order to avoid such situations (e.g., asking for solutions that are also minimal conjunctions of literals). Another possibility, closer to the semantic spirit of this approach, is to work with finite plausibility models, and then look for solutions among the formulas that characterize each possible world.

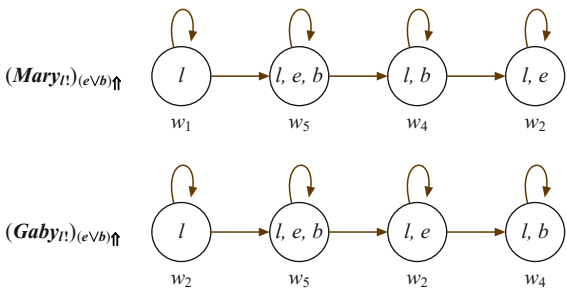
The following example shows how this new operation allows the agent to have ready another explanation in case the initially *best* one turns out to be incorrect.

**Example 13.8**

Let us go back to Mary and Gaby’s example all the way to the stage after which they have observed that the light does not turn on (models  $Mary_{!}$  and  $Gaby_{!}$  of Example 13.5, repeated here).



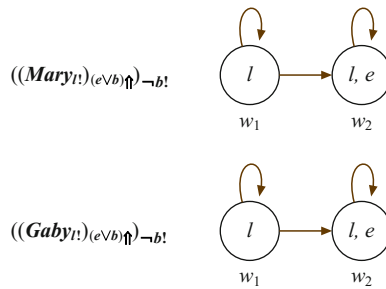
Suppose that, instead of selecting their respective most plausible explanation and assimilating it (as they did in Example 13.5), Mary and Gaby work with all their explanations: Instead of an upgrade with  $e$  for Mary and an upgrade with  $b$  for Gaby, both of them perform an upgrade with  $e \vee b$ . This produces the following models.



The worlds satisfying  $e \vee b$  ( $w_2$ ,  $w_4$ , and  $w_5$ ) have been upgraded. As a result of this, both Mary and Gaby have an explanation for  $l$ , but each one of them has *her own* explanation: While Mary believes that the electric line

has failed ( $e$  is the case in the most plausible world at the model on the top), Gaby believes that the bulb is burned out ( $b$  holds in the most plausible world at the model on the bottom).

So far the result of the upgrade is, with respect to Mary and Gaby beliefs, exactly the same as with the previous proposal where only worlds that satisfy the most plausible explanation are upgraded (in both cases,  $w_2$  and  $w_4$  are Mary’s and Gaby’s most plausible worlds, respectively). But note what happens now when they both observe that the bulb is in fact not burned out ( $\neg b$ ): Such action produces the following situation.



Again, the observation does not affect Mary’s explanation ( $e$  still holds in the most plausible world at model on the top), but it does change Gaby’s since her previous explanation  $b$  is not possible anymore. The difference is that now she does not need to perform an extra abductive step because she has already another explanation: She now believes that the electric line has failed ( $e$  holds in the most plausible world at model on the bottom).

Thus, after an upgrade with all explanations, what the agent will be lead to believe depends on her plausibility order, just as with the first proposal. Nevertheless, if further information invalidates such best explanation, the agent will believe the next to best one without the need of further abductive steps.

**13.8 A Brief Exploration to Nonideal Agents**

As most (if not all) proposals for representing a given phenomena, the presented epistemic and dynamic approach to abduction has made some assumptions for the sake of simplicity. One of the most important of these is the fact that agents whose information is represented within the plausibility framework are *ideal*: Their knowledge and beliefs are closed under logical consequence. This supposition is not exclusive of this approach; the classic logical definitions of abductive reasoning assume not only that the given set of formulas  $\Phi$ , the theory, is closed under logical con-

sequence, but also that  $\vdash$  is the logical consequence relation.

The present proposal highlights the epistemic nature of abductive reasoning, and so it is natural to ask how such reasoning process works for a different kind of agents, in particular, for those whose information does not need to have *ideal* properties and thus are, in that sense, closer to real computational agents with limited resources (and also closer to us human beings). This final section briefly discusses some ideas; further developments in this direction can be found in [13.60].

### 13.8.1 Considering Inference

Suppose Karl is in his dining room and sees smoke coming out of the kitchen. This seems unjustified at first, but then he realises that the chicken he placed on the fire has been there for a long time. Initially Karl did not have any explanation for the smoke, but after a moment he realized that such event was actually not surprising at all.

This case is different from the discussed ones because Karl is not an ideal agent: He does not have at hand all logical consequences of his information, and therefore he did not realize that the information he had before seeing the smoke was enough to predict it (i. e., to infer that there would be smoke). Described in more technical terms, seeing the smoke raised an abductive problem for Karl, but such problem arose because he did not have, at the time of the observation, all the logical consequences of the information he actually had (otherwise there would have been no abductive problem at all). Accordingly, in such case the abductive solution is not necessarily a piece of information that would have allowed Karl to predict the smoke; it might be a simple inference step that made *explicit* what was only *implicit* before.

This shows not only how agents whose information is not closed under logical consequence can face at least a new kind of abductive problem, but also how such problems give rise to a different kind of solutions.

### 13.8.2 Different Reasoning Abilities

In the previous example, the abductive solution was a simple inference step because Karl had the needed

reasoning tools to infer *there is smoke in the kitchen from the chicken has been on the fire for a long time*. But what if that was not the case? That is, what if, besides not having at hand all the logical consequences of his information, Karl did not have the required reasoning tools to infer some of them?

In such new situation, Karl faces again an abductive problem, but this time of a different nature. The surprising observation could have been predicted in the sense that it is a logical consequence of Karl's information *the chicken has been on the fire for a long time*, just as in the initial version of this example. The difference is that such observation is not something that Karl could have predicted by himself: He did not have the needed tools. One can say that, even though *there is smoke in the kitchen* is *objectively* derivable from the initial information, it is not *subjectively* derivable in the sense that Karl could not have done it. To put it in other words, besides not having at hand all the logical consequences of her actual information, Karl might not even be able to reach them.

Accordingly, the simple inference step of before cannot be a solution to the problem now, as Karl does not have the needed tools to perform it. One possible solution is, as in the traditional case, a piece of information that would have allowed Karl to predict the smoke from some other previously known fact, but a more interesting one is some reasoning tool that would have helped him to predict the fire from the known fact *the chicken has been on the fire for a long time*.

New cases arise when further kinds of agents are considered. A systematic study of such cases can be found in [13.61].

## 13.9 Conclusions

This chapter has proposed an epistemic and dynamic approach to abductive reasoning, understanding this form of reasoning as a process that:

1. Is triggered by an epistemic action through which the agent comes to know or believe certain  $\chi$  that otherwise she could not have been able to know or believe
2. Looks for explanations for  $\chi$  in the set of formulas that could have helped the agent to come to know or believe  $\chi$
3. Incorporates the chosen explanation as a part of the agent's beliefs.

Besides providing formal definitions of what an abductive problem and an abductive solution are in terms

of an agent's knowledge and beliefs, the present proposal has discussed:

1. A classification of abductive problems in terms of both how convinced the agent is of the problematic formula after the observation (she *knows* it, or just *believes* it) and how plausible the formula was *before* the epistemic action that triggered the problem
2. A classification of abductive solutions based not only on their deductive relation with the abductive problem or their syntactic form, but also in terms of both their plausibility *before* the problem was raised and the way it will affect the agent's information once they are incorporated
3. A new perspective that looks not for *the best* explanation but rather for *the agent's best* explanation,

and the possibility to carry out this search in terms of which explanations are more likely from the agent's point of view, that is, in terms of the agent's beliefs

4. The possibility of integrating the chosen solution into the agent's information as part of her beliefs, which allows not only to identify trivial solutions because of their effect rather than their form, but also to revise and eventually discard solutions that become obsolete in the light of further information.

Crucial for all these contributions has been the use of plausibility models and, in general, the DEL guidelines, which puts emphasis in the representation of both epistemic attitudes and the actions that affect them.

It is worthwhile to compare, albeit briefly, the present proposal to other epistemic approaches to abductive reasoning. Besides immediate differences in the respective semantic models (while other approaches follow the Alchourrón–Gärdenfors–Makinson (AGM) belief revision, using a set of formulas for representing the agent's information, here possible worlds are used), there are two main points that distinguish the presented ideas from other proposals. First, here several epistemic attitudes are taken into account, thus making a clear difference between what the agent holds with full certainty (knowledge) and what she considers very likely but still cannot guarantee (beliefs); this allows to distinguish between the certainty of both the previous information and the surprising observation, and the mere plausibility of the chosen solution (recall the validity  $K(\eta \rightarrow \chi) \rightarrow [\chi!](K\chi \rightarrow \langle \text{Abd}_\eta^x \rangle B\eta)$ , briefly discussed at the end of Sect. 13.6). Second, this approach goes one step further by making explicit the different stages of the abductive process, thus making also explicit the epistemic actions involved. This highlights the importance of actions such as *belief revision*, commonly understood in epistemic approaches to abduction as the one *triggered* by the abductive problem [13.12, 62], and also such as *observation*, understood here as the one that *triggers* the abductive process.

This chapter presents only the first steps toward a proper study of abductive reasoning from an epistemic and dynamic perspective, and several of the current proposals can be refined. For example, the specific definition of an abductive problem (Definition 13.9) relies on the fact that, within the DEL framework, agents are logically omniscient. As it has been hinted at in Sect. 13.8, in a nonomniscient DEL setting [13.35, 63] the ideas discussed in Sect. 13.2 would produce a different formal definition (which, incidentally, would allow to classify abductive problems and abductive solutions according to some *derivability* criteria). Moreover, it would be possible to analyze the full abductive picture presented in Sect. 13.2.1, which requires inference steps

in the alternative reality path. These extensions are relevant: They would allow a better understanding of the abductive process as performed by *real* agents.

But it is also possible to do more than just follow the traditional research lines in abductive reasoning, and here are two interesting possibilities (whose development exceeds the limits of this chapter). First, the DEL framework allows multiagent scenarios in which abductive problems would arise in the context of a community of agents. In such setting, further to the public observation and revision used here, actions that affect the knowledge and beliefs of different agents in different ways are possible. For example, an agent may be privately informed about  $\psi$ : If this raises an abductive problem  $\chi$  for her and another agent has private information about  $\eta \rightarrow \chi$ , they can interact to obtain the abductive solution  $\eta$ . Second, the DEL framework deals with high-order knowledge, thus allowing to study cases in which an agent, instead of looking for an explanation of a fact, looks for an explanation of her own epistemic state. Interestingly, explanations might involve epistemic actions as well as the lack of them.

According to those considerations, this logical approach takes into account the dynamics aspects of logical information processing, and one of them is abductive inference, one of the most important forms of inference in scientific practices. The aforementioned multiagent scenarios allow to model concrete practices, particularly those that develop a methodology based on observation, verification, and systematic formulation of provisional hypotheses, such as in empirical sciences, social sciences, and clinical diagnosis. The epistemological repercussions of this DEL approach is given by the conceptual resources that it offers, useful to model several aspects of explanatory processes. If known theories of belief revision, at the last resort, say nothing about context of discovery, by means of DEL the accessibility of this context to rational epistemological and logical analysis is extended, further on classical logical treatment of abduction. From the perspective of game theoretic semantics, for example, now it is easier to determine what rules are strategic and what are operatories when abductive steps were given. But applications should also be considered to tackle certain philosophical problems. For example, abductive scenarios within multiagent settings can be used to study the implications of different forms of communication within scientific communities.

**Acknowledgments.** The first author acknowledges the support of the project *Logics of discovery, heuristics and creativity in the sciences* (PAPIIT, IN400514-3), granted by the National Autonomous University of Mexico (UNAM).



## References

- 13.1 A. Aliseda: *Abductive Reasoning. Logical Investigations into Discovery and Explanation*, Synthese Library, Vol. 330 (Springer, Dordrecht 2006)
- 13.2 A.C. Kakas, R.A. Kowalski, F. Toni: Abductive logic programming, *J. Logic Comput.* **2**(6), 719–770 (1992)
- 13.3 M.C. Mayer, F. Pirri: First order abduction via tableau and sequent calculi, *Log. J. IGPL* **1**(1), 99–117 (1993)
- 13.4 M.C. Mayer, F. Pirri: Propositional abduction in modal logic, *Logic J. IGPL* **3**(6), 907–919 (1995)
- 13.5 A.L. Reyes-Cabello, A. Aliseda, Á. Nepomuceno-Fernández: Towards abductive reasoning in first-order logic, *Logic J. IGPL* **14**(2), 287–304 (2006)
- 13.6 S. Klarman, U. Eudriss, S. Schlobar: ABox abduction in the description logic ACC, *J. Autom. Reason.* **46**, 43–80 (2011)
- 13.7 J. Lobo, C. Uzcátegui: Abductive consequence relations, *Artif. Intell.* **89**(1/2), 149–171 (1997)
- 13.8 A. Aliseda: Mathematical reasoning vs. abductive reasoning: A structural approach, *Synthese* **134**(1/2), 25–44 (2003)
- 13.9 B. Walliser, D. Zwirn, H. Zwirn: Abductive logics in a belief revision framework, *J. Log. Lang. Info.* **14**(1), 87–117 (2004)
- 13.10 H.J. Levesque: A knowledge-level account of abduction, *Proc. 11th Intl. Joint Conf. on Artif. Intell.*, ed. by N.S. Sridharan (Morgan Kaufmann, Burlington 1989), pp. 1061–1067, Detroit 1989
- 13.11 C. Boutilier, V. Becher: Abduction as belief revision, *Artif. Intell.* **77**(1), 43–94 (1995)
- 13.12 A. Aliseda: Abduction as epistemic change: A Peircean model in artificial intelligence. In: *Abduction and Induction: Essays on Their Relation and Integration*, Applied Logic, ed. by P.A. Flach, A.C. Kakas (Kluwer, Dordrecht 2000) pp. 45–58
- 13.13 L. Magnani: *Abductive Cognition: The Epistemological and Eco-Cognitive Dimensions of Hypothetical Reasoning*, Cognitive Systems Monographs, Vol. 3 (Springer, Heidelberg 2009)
- 13.14 D. Gabbay, J. Woods (Eds.): *The Reach of Abduction: Insight and Trial, A Practical Logic of Cognitive Systems*, Vol. 2 (Elsevier, Amsterdam 2005)
- 13.15 J. Woods: Cognitive economics and the logic of abduction, *Rev. Symb. Log.* **5**(1), 148–161 (2012)
- 13.16 J. Hintikka: *Knowledge and Belief: An Introduction to the Logic of the Two Notions* (Cornell Univ. Press, Ithaca 1962)
- 13.17 R. Fagin, J.Y. Halpern, Y. Moses, M.Y. Vardi: *Reasoning About Knowledge* (MIT Press, Cambridge 1995)
- 13.18 H. van Ditmarsch, W. van der Hoek, B. Kooi: *Dynamic Epistemic Logic*, Synthese Library, Vol. 337 (Springer, Dordrecht 2007)
- 13.19 J. van Benthem: *Logical Dynamics of Information and Interaction* (Cambridge Univ. Press, Cambridge 2011)
- 13.20 P.A. Flach, A.C. Kakas: *Abduction and Induction: Essays on their Relation and Integration*, Applied Logic (Kluwer, Dordrecht 2000)
- 13.21 F. Soler-Toscano, D. Fernández-Duque, Á. Nepomuceno-Fernández: A modal framework for modeling abductive reasoning, *Log. J. IGPL* **20**(2), 438–444 (2012)
- 13.22 M.E. Quilici-Gonzalez, W.F.G. Haselager: Creativity: Surprise and abductive reasoning, *Semiotica* **153**(1–4), 325–342 (2005)
- 13.23 J. van Benthem, F.R. Velázquez-Quesada: The dynamics of awareness, *Synthese (Knowl., Rationality and Action)* **177**, 5–27 (2010)
- 13.24 B. Hill: Awareness dynamics, *J. Phil. Log.* **39**(2), 113–137 (2010)
- 13.25 F.R. Velázquez-Quesada, F. Soler-Toscano, Á. Nepomuceno-Fernández: An epistemic and dynamic approach to abductive reasoning: Abductive problem and abductive solution, *J. Appl. Log.* **11**(4), 505–522 (2013)
- 13.26 Á. Nepomuceno-Fernández, F. Soler-Toscano, F.R. Velázquez-Quesada: An epistemic and dynamic approach to abductive reasoning: Selecting the best explanation, *Log. J. IGPL* **21**(6), 943–961 (2013)
- 13.27 F. Soler-Toscano, F.R. Velázquez-Quesada: A dynamic-epistemic approach to abductive reasoning. In: *Logic of Knowledge. Theory and Applications*, Dialogues and the Games of Logic. A Philosophical Perspective, Vol. 3, ed. by C. Barés Gómez, S. Magnier, F.J. Salguero (College Publications, London 2012) pp. 47–78
- 13.28 C.S. Peirce: *The Essential Peirce. Selected Philosophical Writings (1893–1913)*, Vol. 2 (Indiana Univ., Bloomington, Indianapolis 1998), ed. by N. Houser
- 13.29 C.S. Peirce: *The Essential Peirce. Selected Philosophical Writings (1867–1893)*, Vol. 1 (Indiana Univ., Bloomington, Indianapolis 1992), ed. by N. Houser, C. Kloesel
- 13.30 E. Lorini, C. Castelfranchi: The cognitive structure of surprise: Looking for basic principles, *Topoi* **26**(1), 133–149 (2007)
- 13.31 G.H. Harman: The inference to the best explanation, *Phil. Rev.* **74**(1), 88–95 (1965)
- 13.32 P. Lipton: *Inference to the Best Explanation* (Routledge, London, New York 2004)
- 13.33 J. Hintikka: What is abduction? The fundamental problem of contemporary epistemology, *Trans. C.S. Peirce Soc.* **34**(3), 503–533 (1998)
- 13.34 A. Baltag, S. Smets: A qualitative theory of dynamic interactive belief revision. In: *Logic and the Foundations of Game and Decision Theory (LOFT)*, Texts in Logic and Games, Vol. 3, ed. by G. Bonanno, W. van der Hoek, M. Wooldridge (Amsterdam Univ. Press, Amsterdam 2008) pp. 13–60
- 13.35 F.R. Velázquez-Quesada: Dynamic epistemic logic for implicit and explicit beliefs, *J. Log. Lang. Info.* **23**(2), 107–140 (2014)
- 13.36 C. Boutilier: Unifying default reasoning and belief revision in a modal framework, *Artif. Intell.* **68**(1), 33–85 (1994)

- 13.37 R. Stalnaker: On logics of knowledge and belief, *Phil. Stud.* **128**(1), 169–199 (2006)
- 13.38 J.A. Plaza: Logics of public communications, *Proc. 4th Intl. Symp. Methodol. Intell. Sys.*, ed. by M.L. Emrich, M.S. Pfeifer, M. Hadzikadic, Z.W. Ras (North-Holland, Amsterdam 1989) pp. 201–216
- 13.39 J. Gerbrandy, W. Groeneveld: Reasoning about information change, *J. Log. Lang. Info.* **6**(2), 147–196 (1997)
- 13.40 J. van Benthem: Dynamic logic for belief revision, *J. Appl. Non-Class. Log.* **17**(2), 129–155 (2007)
- 13.41 H. van Ditmarsch: Prolegomena to dynamic logic for belief revision, *Synthese* **147**(2), 229–275 (2005)
- 13.42 A. Baltag, S. Smets: Learning by questions and answers from belief-revision cycles to doxastic fixed points. In: *Logic, Language, Information and Computation*, ed. by H. Ono, M. Kanazawa, R. de Queiroz (Springer, Berlin, Heidelberg 2009) pp. 124–139
- 13.43 W.A. Carnielli: Surviving abduction, *Log. J. IGPL* **14**(2), 237–256 (2006)
- 13.44 J. Pearl: *Probabilistic Reasoning in Intelligent Systems – Networks of Plausible Inference* (Morgan Kaufmann, San Francisco 1989)
- 13.45 D. Poole: Probabilistic horn abduction and bayesian networks, *Artif. Intell.* **64**(1), 81–129 (1993)
- 13.46 D. Dubois, A. Gilio, G. Kern-Isberner: Probabilistic abduction without priors, *Intl. J. Approx. Reason.* **47**(3), 333–351 (2008)
- 13.47 M. Denecker, D. De Schreye: SLDNFA: An abductive procedure for normal abductive programs, *Proc. Intl. Joint Conf. Symp. Log. Program.*, ed. by K.R. Apt (MIT Press, Washington 1992) pp. 686–700
- 13.48 A.C. Kakas, P. Mancarella: Generalized stable models: A semantics for abduction, *Proc. 9th Eur. Conf. Artif. Intell. ECAI '90*, ed. by L.C. Aiello (Pitman, Stockholm 1990) pp. 385–391
- 13.49 F. Lin, J.-H. You: Abduction in logic programming: A new definition and an abductive procedure based on rewriting, *Proc. 17th Int. Joint Conf. Artif. Intell., IJCAI*, ed. by B. Nebel (Morgan Kaufmann, Seattle 2001) pp. 655–666
- 13.50 M.C. Mayer, F. Pirri: Abduction is not deduction-in-reverse, *Log. J. IGPL* **4**(1), 95–108 (1996)
- 13.51 P. Gärdenfors, D. Makinson: Nonmonotonic inference based on expectations, *Artif. Intell.* **65**(2), 197–245 (1994)
- 13.52 R. Pino-Pérez, C. Uzcátegui: Jumping to explanations versus jumping to conclusions, *Artif. Intell.* **111**(1/2), 131–169 (1999)
- 13.53 R. Pino-Pérez, C. Uzcátegui: Preferences and explanations, *Artif. Intell.* **149**(1), 1–30 (2003)
- 13.54 D. Makinson: Bridges between classical and non-monotonic logic, *Log. J. IGPL* **11**(1), 69–96 (2003)
- 13.55 J. van Benthem, S. van Otterloo, O. Roy: Preference logic, conditionals and solution concepts in games. In: *Modality Matters: Twenty-Five Essays in Honour of Krister Segerberg*, Uppsala Philosophical Studies, ed. by H. Lagerlund, S. Lindström, R. Sliwinski (Univ. Uppsala, Uppsala 2006) pp. 61–76
- 13.56 P. Girard: *Modal Logic for Belief and Preference Change*, Ph.D. Thesis (Stanford Univ., Stanford 2008)
- 13.57 J. van Benthem, P. Girard, O. Roy: Everything else being equal: A modal logic for ceteris paribus preferences, *J. Phil. Log.* **38**(1), 83–125 (2009)
- 13.58 F. Liu: *Reasoning about Preference Dynamics*, Synthese Library, Vol. 354 (Springer, Heidelberg 2011)
- 13.59 F.R. Velázquez-Quesada: Reasoning processes as epistemic dynamics, *Axiomathes* **25**(1), 41–60 (2015)
- 13.60 F. Soler-Toscano, F.R. Velázquez-Quesada: Generation and selection of abductive explanations for non-omniscient agents, *J. Log. Lang. Info.* **23**(2), 141–168 (2014)
- 13.61 F. Soler-Toscano, F.R. Velázquez-Quesada: Abduction for (non-omniscient) agents, *Workshop Proc. MALLOW 2010*, Vol. 627, ed. by O. Boissier, A. El Fallah Seghrouchni, S. Hassas, N. Maudet (CEUR, Lyon 2010), [www.ceur-ws.org/vol-627/lrba\\_4.pdf](http://www.ceur-ws.org/vol-627/lrba_4.pdf)
- 13.62 J. van Benthem: Abduction at the interface of logic and philosophy of science, *Theoria* **22**(3), 271–273 (2009)
- 13.63 F.R. Velázquez-Quesada: Explicit and implicit knowledge in neighbourhood models. In: *Logic, Rationality, and Interaction – Proc. 4th Int. Workshop LORI 2013, Hangzhou*, Lecture Notes in Computer Science, Vol. 8196, ed. by D. Grossi, O. Roy, H. Huang (Springer, Berlin, Heidelberg 2013) pp. 239–252

# 14. Argumentation and Abduction in Dialogical Logic

Cristina Barés Gómez, Matthieu Fontaine

This chapter advocates for a reconciliation of argumentation theory and formal logic in an agent-centered theory of reasoning, that is, a theory in which inferences are studied as human activities. First, arguments in favor of a divorce between the two fields are presented. Those arguments are not so controversial. However, rather than forcing a radical separation, they urge logicians to rethink the object of their studies. Arguments cannot be analyzed as objects independent from human activity, whether it is dealt with deductive or nondeductive reasoning. The present analysis naturally takes place in the context of dialogical logic in which the proof process and the semantics are conceived in terms of argumentative games, which involve the agents, their commitments and their actions. This work focuses first on deductive reasoning and then takes abduction as a case of nondeductive reasoning. By relying on some relevant ideas of the Gabbay-Woods (GW) schema of abduction and Aliseda's approach, a new dialogical explanation of abduction in terms of *concession-problem* is proposed. This notion of *concession problem* will be defined thereafter. With respect to the topics of the model-based sciences, the question of the specificity of the speech act by means of which a hypothesis is conjectured is set more specifically.

14.1	<b>Reasoning as a Human Activity</b> .....	295
14.2	<b>Logic and Argumentation: The Divorce</b> .....	297
14.3	<b>Logic and Argumentation: A Reconciliation</b> .....	299
14.3.1	What is Dialogical Logic? .....	299
14.3.2	Particle Rules .....	300
14.3.3	Structural Rules .....	301
14.3.4	Winning Strategy and Validity .....	302
14.4	<b>Beyond Deductive Inference: Abduction</b> .....	303
14.4.1	The GW Model of Abduction .....	303
14.5	<b>Abduction in Dialogical Logic</b> .....	306
14.5.1	Triggering .....	306
14.5.2	Guessing .....	308
14.5.3	Committing .....	309
14.6	<b>Hypothesis: What Kind of Speech Act?</b> .....	310
14.7	<b>Conclusions</b> .....	312
	<b>References</b> .....	312

## 14.1 Reasoning as a Human Activity

In this chapter, it is argued against the radical dissociation of formal logic and argumentation advocated by *Toulmin* [14.1] and *Perelman and Olbrechts-Tyteca* [14.2]. It is proposed to bring formal logic and argumentation together in the field of dialectical interaction in which the *human being* and the *action* of the agent are given a central role. In this contribution, a unified theory of reasoning is thus advocated, the key concept of which is not something as a *universal logic* but rather the notion of *commitment*, that is, what a speaker is ready to defend on uttering a sentence or

in making use of a particular argument. Indeed, from the perspective of dialectical interactions, the crucial question is: *What are we committed to when we utter a sentence in a dialectical interaction?* In other words, when an agent performs a claim, it is never for free, and further justifications may be demanded for by the speaker's argumentative partners. The commitment to providing further justifications precisely constitutes the ground in order to distinguish between various kinds of speech acts relevant for the specification of different forms of reasonings. This study takes place within

the dialogical framework in which the proof is conceived in terms of a dialectical process. The specificity of deductive and abductive reasonings is clarified by identifying different kinds of speech acts specific to each of these forms of reasoning. The aim is to show that abductive dialogues involve specific speech acts, namely certain conjectural claims that differ from usual assertions and questions of deductive dialogues. A more exhaustive study of commitment and its role in the definition of different kinds of speech acts in dialogical interaction which can also be found in *Walton and Krabbe* [14.3]. However, this study focuses here on some aspects of commitment related to assertions and questions in deductive dialogues and considers how to extend the picture to abductive dialogues. The context of this study is first explained. In order to defend a practical logic to study the fallacies, *Woods* [14.4] identifies what he calls *third-way reasoning*, which operates beyond the usual standards of deduction and induction. According to *Woods*, logicians have missed the target concerning the study of fallacies because they have failed to invoke the right standards of reasoning. The mistake is linked to an ostracism with respect to the human being when the task should be to describe reasoning. Indeed, in most logical studies of reasoning, the human being has simply been left out of the story! In *Woods*' own words, "there are no people in the models of mainstream mathematical logic" [14.4, p. 12].

*Toulmin* [14.1] also reports that logicians left out the human being while they were modeling reasoning. As a solution, he urges for the divorce between logic and argumentation by claiming that logic was too narrow to study argumentation. *Toulmin* was right in thinking that formal logicians had forgotten the human being. He was wrong in thinking that the solution was to dissociate logic and argumentation. Independently of how some logicians might have led their investigations, the point of view endorsed in this chapter is that an agent-centered logic (that is, a logic built around human activity) is possible. Logic and argumentation should again be brought together. Human beings play a fundamental role in third-way reasoning as well as in deductive reasoning.

A study centered on the role of the agent constitutes the condition of possibility of a unified theory of reasoning, that is, a theory in which logic and argumentation are analyzed together. What is to be considered is not a mere relation of consequence-having but a relation of consequence drawing. As stressed by *Woods* [14.4], while the former is a mere relation between propositions, the latter is to be linked with agent-based inferences, that is, actions by means of which an agent draws conclusions. The latter is the ba-

sis of what has been called an *agent-centered logic*. The position defended in this chapter, which is perhaps stronger than that of *Woods*, is that focusing on a consequence-having relation is also a mistake with respect to deductive reasoning. Reasoning, in general, must be studied in a general framework in which particular attention is paid to the action of the agents and their commitments.

More precisely, it is argued that deductive as well as nondeductive reasoning should be understood within argumentative practices, taking into account the interaction between agents. This can be achieved by means of dialogical logic, a semantics based on argumentative practices and presented as a game between a proponent of a thesis and an opponent to this thesis. More precisely, dialogical logic is grounded on speech acts and commitments related to these speech acts. That is, a dialogue is a sequence of speech acts, questions and assertions, in order to justify or challenge an initial thesis. Moreover, utterances are not free of further justifications: When we utter something, we are committed to providing justification of what we are saying. This is the basis of the rules which say how to challenge and how to defend an utterance. Deductive validity is thus conceived in terms of strategy by means of which a proponent of a thesis defends her initial claim against every attack of her opponent.

However, this is just deduction! Is it possible to generalize the picture to nondeductive reasoning? To answer this question, abductive reasoning will be considered as a case of nondeductive reasoning. A relevant conceptual question is therefore the following: What is the difference and the specificity of abduction with respect to other inference kinds? If it makes sense to talk about abduction as a third kind of inference, it is because it is neither a deductive nor an inductive inference.

According to *Gabbay and Woods* [14.5, p. 192], "[w]hereas deduction is truth-preserving and induction is probability-enhancing, abduction is ignorance-preserving." An abduction is triggered by an ignorance problem that arises when a fact cannot be explained by the current knowledge of an agent. The inability to solve an ignorance problem is a cause of discomfort, which *Gabbay and Woods* [14.5, p. 190] call a *cognitive irritant*. Such an unpleasant situation is sometimes overcome by conjecturing a hypothesis on the basis of which further actions are made possible. Even if such a conjecture allows the agent to overcome the irritant situation, it does not constitute a solution to the ignorance problem: It is only a defeasible hypothesis. This precisely grasps the specificity of abduction.

Rather than an explanation in terms of *ignorance problem*, the specificity of abduction is set from a di-

logical perspective in terms of *concession problem*. A *concession problem* is overcome by a conjecture on the basis of which the dialogue is continued. In contrast with the usual deductive dialogues, such a conjecture is settled in a new kind of move allowed by an additional rule. The difficulty is thus to specify which kind of speech act is at stake while performing such conjectural moves. Indeed, under the view endorsed in this chapter, conjectural moves are performed by means of speech acts which are neither assertions nor questions of usual deductive dialogues.

Reasons why Toulmin argues in favor of a radical separation between formal logic and argumentation are given in the first section. Although it is true that some aspects of argumentation such as the role of the agent, the dynamic of the contexts, and the defeasibility are to be taken into account, it is not a reason to conclude that formal logic and argumentation should be studied

separately. First, it is not true that those aspects are completely missing in formal logic. It is shown in the rest of this section that numerous formal logics deal with these aspects, although they have yet to be brought all together. Second, in this contribution, it is thought that even deduction is to be understood within argumentative practices. Hence, the dialogical framework is introduced in the third section, where it is come back to the key concept of commitment. It is also shown how dialogical logic enables to grasp the central role of the agent as well as the dynamics of the contexts in terms of a pluralist attitude. After having presented abductive reasoning in the fourth section, the scene for a dialogical understanding of abduction is set in the fifth section. All the details of dialogical pluralism, dynamics of contexts, and dialogical defeasibility cannot be given here. However, the relevant related works on each of these points will be systematically mentioned.

## 14.2 Logic and Argumentation: The Divorce

Heavy criticism against the formal logic approach to natural human reasoning has been raised by theoreticians of argumentation who have stressed the importance of the context, the plausibility and the defeasibility of arguments, the commitments and the actions of the agents, and so on. Some of the most virulent of these theoreticians were perhaps *Toulmin* [14.1] (see also [14.6] for recent studies about Toulmin Model) and *Perelmann* and *Olbrechts-Tyteca* [14.2]. This chapter focuses on Toulmin, who defined a model of argumentation based on the analysis of microarguments. This model will be called the *Toulmin Model* of argumentation whose general idea is that some data leads to the claim (or conclusion). The data is supported by a war-

rant. The whole process is qualified by an adverb such as *plausibly*, *probably*, or *necessarily*, that may be rebutted. An important insight of Toulmin’s work was to emphasize the role of the agent and the persuasive feature of argument. Arguments are used to persuade someone to believe something. An agent puts forward an argument in order to defend a thesis and the inferences are defeasible, that is, they might be rebutted when new information is encountered. Schematically, the Toulmin Model may be represented as in Fig. 14.1.

This schema represents the process that consists in defending a claim against a challenger. First, the agent asserts a *Claim* (C) and then defends this claim by appealing to relevant available facts, the so-called *Data*

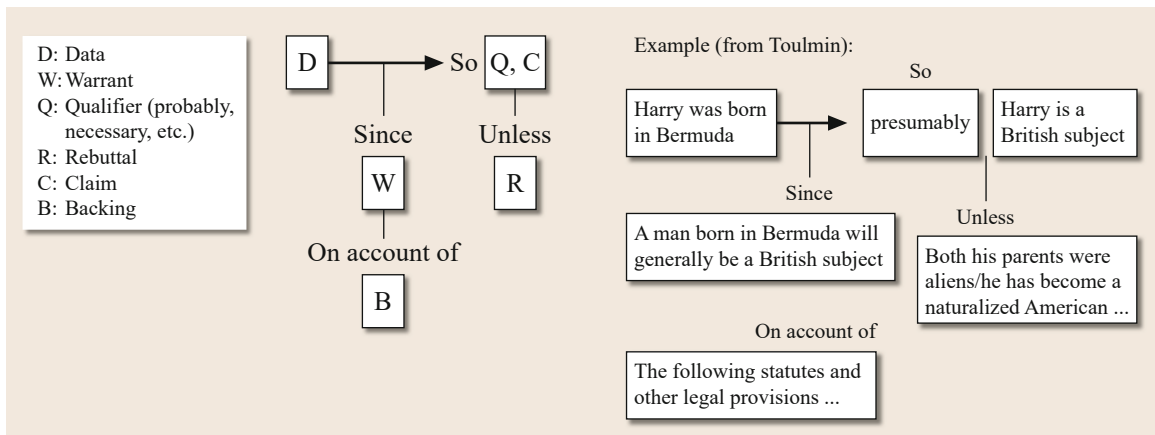


Fig. 14.1 Toulmin Model

(D). Next, the challenger may ask for the bearing of the data and this is exactly what is called the *Warrant* (W). The warrant influences the degree of force on the conclusion it justifies and this is signaled by the qualifying of the conclusion with the *Qualifier* (Q): *necessarily*, *probably*, or *presumably*. The qualifier *presumably* renders the argument defeasible, and the condition of *Rebuttal* (R) should be specified. The process ends in a question that consists in asking what is thought about the general acceptability of the argument: what Toulmin calls *Backing* (B). In different fields, warrant and backing might be of different kinds.

According to the Toulmin Model, an argument is assumed to be used by a practical agent. Inferences are not conceived in terms of the relationship between propositions independent from any act. And the act of inferring is linked with the agent who expresses a claim, by means of which a commitment to a thesis is in fact expressed. The underlying methodological thesis is that the study of reasoning must be related to real-life reasoning. This kind of reasoning is never perfect (as in an ideal model of formal logic) because we never have all the information needed to defend a claim and we might always find a rebuttal that changes it. Hence, the right standard of a good argument cannot be the deductive standard of validity. An argument succeeds or fails only in relation to an agent's target. Toulmin's schema enriches the traditional premises–conclusion relationship of the deductive reasoning model of arguments by distinguishing additional elements, such as warrant, backing, and rebuttal. It is an interesting fact that the Toulmin Model and argumentation theory call up not only the matter of the burden of proof, but also the matter of the burden of questioning, which is of importance for the beginning of the process. A consequence of this action- and agent-centered analysis is that an account of the defeasibility of reasoning is now required. The fact that none of these features appeared in formal logic constituted the core of Toulmin's criticism, that led him to consider argumentation theory and formal logic as radically different disciplines.

There is nothing really controversial in Toulmin's critics of formal logic or in his model of argumentation. Nevertheless, following *van Benthem* [14.7], in this chapter, it is believed him to be wrong in pronouncing the divorce of argumentation theory and formal logic. Indeed, it might be true that classical formal logic is insufficient to deal with reasoning as a human activity. Classical formal logic is not the only way to do logic, however. Although Toulmin's work has the virtue of emphasizing the role of human being, the defeasible feature of everyday life reasoning, and the dynamic of argumentative contexts, it is worth noting that those features were not completely lacking in formal logic.

With respect to the agent, intuitionism initiated by the Dutch mathematician *Brouwer* [14.8] is motivated by the need of taking the importance of the agent into account. More recently, agent-centered dialogical logic initiated by *Lorenzen* [14.9, 10] conceives the notion of proof itself in terms of interactions between agents. It remains true that further efforts are still required to deal with nondeductive reasoning. Before pronouncing the divorce between logic and argumentation theory, it should be recognized that many logicians had already widened the range of argumentative schemas in formal logic, by adding the agent, thinking otherwise the premises–conclusion relation, and defining several kinds of consequence relations.

Certain cognate aspects of reasoning must be grasped. For example, inference is a process which involves a flow of information, changes of belief, knowledge or even desires. Logicians have to take the “dynamic turn”, in the words of *Gochet* [14.11]. That is why the agent has sometimes been introduced explicitly in the object language in order to express intentional relations by means of specific operators. The enterprise does not always head in the same direction as an agent-centered analysis (in fact it almost never goes in that direction) but the enterprise does provide new tools on how to implement the agent in the study of reasoning. *Hintikka's* explicit epistemic logic [14.12], and more recently *Priest's* intentional logic [14.13], among others, define useful tools to describe the intentional states of an agent. In addition, dynamic approaches, such as the AGM-Belief Revision Theory [14.14] (and see Chap. 10), are meant to give an account of how to incorporate new pieces of beliefs into an agent's belief set, conceived as a set of sentences. In the same spirit, dynamic studies coming from natural language semantics [14.15–18] and dynamic epistemic logic [14.17–19] add operators to deal with the flow of information and the transmission of information between groups of agents. The study of dynamic inferences is not restricted to model theory and to the change in information. From a pluralistic point of view, a change of logic might occur with respect to a given context of argumentation. For example, dialogical logic is a pluralist enterprise in which the context of argumentation is defined by means of rules governing the general organization of a dialectical game (more precision on this point below). Although this fails to provide Toulmin with an answer to each critic he addresses on formal logic, it does reveal how formal studies are sufficiently rich to consider the possibility of a more practical logic in which reasoning is conceived as a human activity.

Another aspect of argumentation stressed by Toulmin is the imperfect feature of human reasoning, which

he deals with by means of the notion of *rebuttal*. Thinking of reasoning as defeasible means that an agent never draws conclusions definitively, that is, whatever she infers from a given base of information might be revised when faced with new information. In other words, the conclusions drawn by an agent might be defeated. It is worth noting that defeasibility does not need to be studied in the context of nonmonotonic logics. If nonmonotonic reasoning is defeasible, the converse does not hold. Interesting ways of defeasible cases come from the context [14.20]. What characterizes defeasible reasoning is the possibility to defeat, or to change a previously drawn consequence. Again, this feature cannot be claimed to be completely missing in formal logic. Indeed, defeasible reasoning has been studied from various perspectives [14.21]. As already mentioned, one well-known approach is the epistemic approach such as that in the context of belief revision theory. The formal epistemology of *Pollock* [14.22], who differentiates between fundamental knowledge and inferred knowledges, provides another example. In this theory, inferred knowledge is precisely a knowledge which might be defeated. Another approach is centered on the notion of logical consequence, that is, dealing with defeasibility in the context of nonmonotonic logics (Chap. 10). Some of the most important proposals are Default Logic by *Reiter* [14.23] and Circumscription by *McCarthy* [14.24]. In both of these frameworks, the conclusion follows defeasibly or nonmonotonically from a set of premises, just in the case that it will hold in almost all models that verify the premises. (For a relevant survey, see for example [14.25–27]. See also the third-way reasoning in [14.4].) It is also important to mention *Batens'* adaptive logic [14.28], a formal logic in which the application of inference

rules may be subject to conditions with respect to the context of the proof (e.g., in the context of contradictory premises, disjunctive syllogism might be rejected).

The three main aspects of the Toulmin Model of argumentation that have been highlighted are the central role of the agent, the dynamics related to the action and changes of contexts, and defeasibility. In what follows, it will be argued for a reconciliation of formal logic and argumentation, and deduction will be also defined in argumentative practices. Note that it is not the purpose of this chapter to deal exhaustively with all the relevant aspects of argumentation. Indeed, every facet of the dialogical pluralism (although the general principles are explained and relevant related works are mentioned) or defeasible reasoning cannot be presented here. The designation *defeasible reasoning* gathers together aspects of default logic, nonmonotonic logics, truth maintenance systems, defeasible inheritance logics, autoepistemic logics, circumscription logics, logic programming systems, preferential reasoning logics, abductive logics, theory revision logics, belief change logics, and so on. In fact, all of this relates to what is called by *Woods* the *third-way reasoning* [14.4]. Various systematic approaches to defeasible argumentation that make use of formal tools originating from computational sciences and artificial intelligence can be found in [14.29].

The main thesis of this contribution is that a unified study of reasoning may be achieved by focusing on the key notion of commitment in argumentative interaction. Indeed, this notion forms the basis for a distinction between various kinds of speech acts that are significant for the specification of different kinds of reasonings, such as deduction and abduction.

## 14.3 Logic and Argumentation: A Reconciliation

It is true that a study of logic that is not centered on human activity is not sufficient to deal with reasoning in general. However, it is a mistake to conclude that the divorce between argumentation and logic is to be pronounced. Logic and argumentation must be brought together within a general framework in which a consequence–drawing relation, conceived as a human practice, is taken into account, and not a consequence–having relation conceived as a mere calculus between propositions. In the rest of this section, deduction is modeled inside an argumentation theory and the standard (deductive) dialogical logic is defined by giving the rules for the propositional level.

### 14.3.1 What is Dialogical Logic?

The dialogical logic referred to in this chapter has its roots in the works of *Lorenzen* and *Lorenz* [14.10], and more recently in *Rahman* [14.30] and his collaborators (see for example *Rahman* and *Keiff* [14.31], *Fontaine* and *Redmond* [14.32], and *Clerbout* [14.33]). Different works on dialogical logic have also been developed by *Barth* and *Krabbe* [14.34, 35], among others, and by the Pragma-dialecticians from an informal point of view [14.36, 37].

Dialogical logic is considered to be an alternative semantics, that is neither a model-theoretic semantics

nor a proof-theoretic semantics, and is grounded in the argumentative practices. It is a semantics based on the “meaning is use” of *Wittgenstein* [14.38, p. 43] and the description of specific language games governed by the rules defined below. Although it was first developed to deal with intuitionist logic, it has since then taken a pluralist turn. Indeed, different kinds of rules enable a sharp distinction between different semantic levels and this enables the definition of a wider range of logics in a unified framework.

Roughly speaking, dialogical logic is a framework in which the proof process is conceived as a dialectical game between two players: the *Proponent* of a thesis and the *Opponent*. The Proponent utters an initial thesis and tries to defend it against challenges performed by the Opponent, who criticizes the thesis. The two players make moves alternately. Those moves consist of specific speech acts by means of which they perform challenges and defences. A thesis is valid if and only if the Proponent is able to defend it against every attack of the Opponent. In order to criticize an assertion of her argumentation partner, a move in which a formula has been uttered has to be challenged with respect to its main connective. Such sequences of utterances, challenges, and defences are regulated by the particle rules by means of which the local meaning of the logical constant is given. In addition, structural rules give the general organization of the dialogue and determine the global level of semantics. In fact, these structural rules dictate how the particle rules may be applied and allow to define different games for different contexts of argumentation, for different underlying logics. In the following sections, the particle rules are given, then the structural rules, and finally, the notion of winning strategy which is necessary for the definition of the dialogical notion of validity is presented.

### 14.3.2 Particle Rules

In a dialogical games, moves are of two different kinds: challenges and defences (plus the utterance of the initial thesis as a special move), and are performed by means of two kinds of speech acts: assertive utterances and interrogative utterances. Note that challenges are not necessarily performed by means of interrogative utterances (as shown later below). An utterance is challenged with respect to its main connective. How to challenge and how to defend an utterance is prescribed by the particle rules, which therefore gives the local meaning of logical constants. More precisely, particle rules are abstract descriptions of how an assertion may be challenged and defended with respect to its main connective. They are abstract because they are not related to any specific context of argumentation and are

defined independently of the identity of **P** and **O** (hence they are defined making use of player variables **X** and **Y**). It is fundamental that when agents perform utterances, they are committed to justify their claims. This commitment is essential in the characterization of different kinds of speech acts and in giving the meaning of what is said.

The language used to define the rules of dialogical logic is defined as follows. Let *L* be the language of standard propositional logic:

- Two labels, **O** and **P**, stand for the players of the game: the Opponent and the Proponent, respectively.
- To define particle rules, variables **X** and **Y** are required, with  $X \neq Y$ , that hold for players (regardless of their identity with **O** or **P**).
- Force symbols, ! and ?, are used to specify the kind of speech act at stake: ! for declarative utterances, and ? for interrogative utterances.
- The conjunction can be indexed yielding  $\wedge_i$ , where  $i \in \{1, 2\}$ , such that  $\wedge_1$  stands for the first conjunct, and  $\wedge_2$  the second.
- $r := n$  indicates the rank chosen by the player at the beginning of a dialogue, as pointed out by the rule [SR0]. For example,  $n := 1$  means that the rank is 1. (The notion of rank is explained and defined in Sect. 14.3.3)

A *move* is an expression of the form  $X-f-e$ , where **X** is a player variable, *f* a force symbol, and *e* is either a well-formed formula of *L* or a question of the form  $? \vee$  or  $? \wedge_i$ . Note that the dash – has no meaning, it is used only in order to distinguish in a clear way the element of a dialogical expression. A sequence of such moves will be called a *play*, and a sequence of plays a (dialogical) *game*.

Particle rules (Table 14.1) are abstract descriptions that consist of sequences of moves such that the first member of the sequence is an assertive utterance, the second says how to challenge that utterance with respect to its main connective, and the third says how to answer the challenge.

Rules are *abstract descriptions* that are formulated by making use of variables **X** and **Y** (and not **O** and **P**).

**Table 14.1** Particle rules

Assertion	Challenge	Defence
$X-!-\varphi \wedge \psi$	$Y-?- \wedge_1$ or $Y-?- \wedge_2$ ,	$X-!-\varphi$ or $X-!-\psi$ respectively
$X-!-\varphi \vee \psi$	$Y-?-?\vee$	$X-!-\varphi$ or $X-!-\psi$
$X-!-\neg\varphi$	$Y-!-\varphi$	No defence
$X-!-\varphi \rightarrow \psi$	$Y-!-\varphi$	$X-!-\psi$



They are independent of any specific context of argumentation. They are the same no matter the presupposed logic and are applied in the same way by both players. The formulation of particle rules is symmetric.

Symmetry is an essential feature of dialogical particle rules and this is the reason why dialogical logic is immune to trivializing connectives such as *Prior's* tonk [14.39], even if there is no reference to any model or to any truth condition. *Rahman* et al. [14.40] and *Rahman* [14.41] show that defining a rule for a tonk operator would lead to a formulation of particle rule which is not symmetric. This would involve player-dependent rules, which is not possible in dialogical logic because, at the local level, the identity of the players has not been yet defined. As rightly stressed by *Clerbout* [14.33], it does not even make sense to talk of Opponent and Proponent at the local level. Indeed, the identity of the players is defined at the level of structural rules, when it is said, for example, that the Proponent is the player who utters the initial thesis.

Note how commitment is essential to the meaning of an assertion. An agent, on uttering a conjunction, is committed to give a justification for both of the conjuncts. Hence the challenger has the choice of which subformula to defend. That is, if **X** utters  $\varphi \wedge \psi$ , **Y** challenges this move by asking either  $?\wedge_1$  (the first conjunct) or  $?\wedge_2$  (the second conjunct). In the case of a disjunction, it is the defender (**X**) who chooses. Indeed, an agent uttering a disjunction is committed to give a justification for (at least) one of the disjuncts, that is, **Y** asks  $?\vee$  and **X** chooses to answer either  $\varphi$  or  $\psi$ .

Note that a challenge on a negation cannot be answered. The challenge consists in a switch in the burden of the proof: If a player **X** utters a formula  $\neg\varphi$ , a player **Y** challenges that formula uttering  $\varphi$  and has to defend it thereafter. For the conditional, **Y** takes the burden of the proof of the antecedent. It might be said that when an agent **X** utters a conditional  $\varphi \rightarrow \psi$ , then **X** is committed to justifying  $\psi$  with the proviso that the argumentation partner **Y** concedes  $\varphi$ .

### 14.3.3 Structural Rules

Now structural rules are needed in order to define the general organization of a dialogue by explaining how to apply the particle rules, that is, how to start a dialogue, who has to play, when, who wins, and so on. The global level of meaning is defined by these rules, that is, a level of meaning that arises from the application of the particle rules in specific contexts of argumentation.

#### [SR0] [Starting Rule]

Let  $\varphi$  be a complex formula. Every dialogical game

$D(\varphi)$  starts with the assertion of  $\varphi$  by **P** ( $\varphi$  is called the *initial thesis*). **O** and **P** then choose a positive integer called *repetition rank*.

#### [SR1-c] [Classical Gameplay Rule]

After the ranks have been chosen, moves are alternately performed by **O** and **P** and every move is either a challenge or a defence. Let  $n$  be the repetition rank of a player  $X$ : When it is  $X$ 's turn to play,  $X$  can challenge a preceding utterance or defend herself against a preceding challenge at most  $n$  times by the application of particle rules.

#### [SR1-i] [Intuitionistic Gameplay Rule]

After the ranks have been chosen, moves are alternately performed by **O** and **P** and every move is either a challenge or a defence. Let  $n$  be the repetition rank of a player  $X$ : When it is  $X$ 's turn to play,  $X$  can challenge a preceding utterance or defend herself against *the last challenge which has not yet been defended*, at most  $n$  times by the application of particle rules.

#### [SR2] [Formal Rule]

**P** is not allowed to utter an atomic formula unless **O** uttered the same atomic formula before. Atomic formulae cannot be challenged.

#### [SR3] [Winning Rule]

A player **X** wins the game if and only if the game is finished and **X** made the last move. It is said that a game is finished if and only if there are no more moves allowed according to the particle rules.

The first rule [SR0] sets the identity of the players by claiming that the Proponent is the one who utters the initial thesis and introduces asymmetry. Once the initial thesis is uttered, the players have to choose a rank of repetition. That rank of repetition prevents them from infinitely repeating the same moves. In fact, they indicate how many times a player can challenge or defend a formula. For example, if a player chooses rank 1, then this player is allowed to challenge a formula at most once. Ranks are used to ensure that every game ends after a finite number of moves. Rules [SR1-c] and [SR1-i] regulate the gameplay and distinguish classical from intuitionistic games. Note that a game is never played with both of them. The classical rule [SR1-c] does not impose any restriction with respect to the defences. While playing with the intuitionistic rule [SR1-i], it is forbidden to defend the same move twice or to give a defence against a challenge that is not the last one. This is related to the intuitionistic requirement of having a direct justification for the uttered formula.

The formal rule, [SR2], might be understood as a rule that prevents the Proponent from making any supposition which might be used to win. Without that rule, dialogues would be trivial and the Proponent would al-

ways be in a situation to win. Finally, the winning rule, [SR3], gives the conditions of victory.

### 14.3.4 Winning Strategy and Validity

Hitherto, nothing has been said about the notion of validity. In dialogical logic, validity is not defined in terms of truth preservation but rather in terms of winning strategy. It is said that a player has a winning strategy if and only if she is able to win regardless of the moves and the choices made by her argumentation partners. This leads to the strategic level which is not involved at the level of particle and structural rules. Indeed, nothing in those rules indicate how to play strategically and in no way do they indicate how to win; neither do they prevent anybody from playing badly. Note then that it is not one play of the game which is to be taken into account to determinate the validity of a formula: The validity of a formula is determined by the existence of a winning strategy.

Now, it is reasonable to ask whether a generally good strategy exists. First a comment about the choice of rank. As explained by Clerbout [14.33,42], it is sufficient to consider the case in which the Opponent chooses rank 1 and the Proponent rank 2 in order to obtain a significant range of winning strategies to deal with deductive validity. Second, trained dialogicians know in fact that the best way to play is always to let the Opponent choose first when it is possible and thereafter to repeat the same choices. This is the well-known copy-cat strategy based on a clever use of the formal rule.

An illustration of a dialogue is given in Table 14.2 by taking the elimination of double negation principle  $\neg\neg p \rightarrow p$  as an example. In Table 14.2, the moves of the players are written down in the column **O** for the **O**-moves, and in the column **P** for the **P**-moves. The number of a move is indicated in the outer column whereas those of the challenges moves are indicated in the inner columns. The game runs by applying the classical rule [SR1-c].

At move 0, **P** states the initial thesis. At move 1, **O** chooses rank 1 and **P** chooses rank 2. At move 3, **O** challenges the initial thesis uttering the antecedent of the conditional, namely  $\neg\neg p$ . **P** cannot answer immediately by giving the consequent  $p$  because **P** cannot utter an atomic formula. Therefore, at move 4, **P** challenges the double negation  $\neg\neg p$  by uttering  $\neg p$ . No defence is allowed and **O** has to counter-attack by uttering  $p$ . **P** uses that concession to answer to the attack 3 at move 6. Again, **P** wins. However, this game has been

Table 14.2 Dialogue 1

	<b>O</b>		<b>P</b>		
			$\neg\neg p \rightarrow p$	0	
1	$r := 1$		$r := 2$	2	
3	$\neg\neg p$	0	$p$	6	
	---		3	$\neg p$	4
see o 5	$p$	4	---		

played with classical rule [SR1-c]. If it had been played with the intuitionistic rule [SR1-i], **P** would have lost. **P** could not have performed move 6 because the last challenge of **O** is 5, not 3 ([SR2]). Thus the dialogue would have ended at move 5 with a victory by **O**.

These two different possible gameplays illustrate the difference between classical and intuitionistic negation. Quine’s claim “change of logic, change of subject” [14.43, pp. 80–94] must be thought otherwise. Indeed, the dialogical setting displays that negation has the same local meaning in every logic, and its global meaning is changing according to its use in different contexts of argumentation. Both the semantic levels are significant in fully defining the meaning of an expression.

Beyond the classical and intuitionistic logics, the sharp distinction between the particle rules and the structural rules allows a development of dialogical logic as a pluralistic tool. The pluralistic aspect of dialogical logic allows us to deal with various kinds of argumentation contexts and their dynamics, the importance of which has been stressed by argumentation theoreticians. Indeed, more expressive languages may be introduced by means of the introduction of new symbols, the (local) meaning of which will be given by a particle rule. A language may be used in different contexts of argumentation, with various underlying logics. Dialogically, this means that a language may be used in different kinds of games distinguished by their structural rules.

As stated earlier, it is not the purpose of this contribution to present all the varieties of dialogical logics which nevertheless should be taken into account in order to deal with the contextual aspect of argumentation. More details on first-order dialogical logic are to be found in Clerbout [14.42]. With different structural rules it is also possible to define a dialogical free logics as in Rahman et al. [14.44], Fontaine and Redmond’s paper in [14.45] and an application to the logic of fiction is to be found in Fontaine [14.46]. For the introduction of modal operators (and explicit contexts of argumentation) and their use in different modal frames, see [14.47] and [14.31].

## 14.4 Beyond Deductive Inference: Abduction

Within dialogical logic, an analysis of the relation of consequence–drawing in terms of argumentative games in which the action of the agents and their interactions are taken into account has been proposed. Until now, it has been focused on deductive reasoning. However, if the aim is to bring together logic and argumentation, it is necessary to extend the dialogical approach to nondeductive reasoning. This is performed by taking abduction as a case of nondeductive reasoning. After having defined the conception of abduction that will be defended here, the basis for abductive dialogues will be described. While relying on existing proposals in this field, the aim is to offer a new and different understanding of abduction in the context of a dialogical interaction.

As explained by *Magnani* [14.48], the *knowledge assimilation theory* of *Kowalski* [14.49], in which the assimilation of new information into a knowledge base is described, might explain the role of the agent in abduction in terms of the generation of hypotheses. *Aliseda* also explores an epistemic study of abduction in which a more important role would be given to the agent. She defines abduction in terms of epistemic changes in the context of Belief Revision Theory [14.50, pp. 179] (see also Chap. 10). The role of the agents might be strengthened by developing that approach in the context of dialogical logic, and more precisely in the context of the dialogical approach to belief revision of *Fiutek* [14.51]. In a similar way, *Nepomuceno* et al. [14.52] (see also Chap. 13 by *Nepomuceno* et al.) define abduction in the context of dynamic epistemic logic and its public announcement operator. This might be dialogically understood on the basis of *Magnier* [14.53]. However, this is not the path followed in this contribution, because the agent would be introduced into the language and abduction would still be understood in terms of consequence–having relation, despite some kind of interaction in a dialogical reconstruction. Moreover, an epistemic understanding of abduction would lead to consider hypothetical *abductive solutions* as new pieces of knowledge; something that is not defended in this chapter, as clarified in the following.

Essentially, the challenge consists in explaining what is specific to abduction in a dialogue. As shown below, while studying abduction, the concepts of *abductive problem* and *abductive solution* are fundamental (Chap. 10). In order to define dialogues based on these concepts, a new kind of move performed by means of a specific type of speech act is needed. Therefore, the problem is to clarify this type of speech act and the rules which govern it. Again, the key question

is related to commitment: What are we committed to when we state an *abductive problem* or an *abductive solution*? The purpose is to understand abduction in terms of consequence–drawing and to study the key step of such an inference in terms of interactions in relation to the question of commitment. Therefore, although there exists different approaches to abduction, in this chapter, the GW schema (following *Gabbay* and *Woods* [14.5]), in which a central role is given to the agent, constitutes a landmark. This contribution will also rely on *Aliseda*'s insights [14.50] when a dialogical reconstruction of abduction is proposed, thereby benefitting from her clear and formal systematization of this kind of inference.

### 14.4.1 The GW Model of Abduction

What is characteristic to abduction and is not characteristic to other reasoning kinds, such as deduction and induction? When is an abduction triggered? Why does an agent begin an abductive process? How does an agent draw abductive conclusions and what is the (cognitive or epistemic) status of those conclusions? According to *Gabbay* and *Woods* [14.5, 54], and more recently *Woods* [14.4], abduction is first to be understood as an inference triggered by an ignorance problem and, second, the relation between the premises and the conclusion is to be understood as an ignorance–preserving relation.

Abduction is an inference triggered in response to an ignorance problem, in particular, there is an ignorance problem when, with respect to a (surprising) fact or state of affairs, there is a question (a problem),  $Q$ , we cannot answer with our present knowledge. We assume that there is a sentence  $\alpha$  such that if we knew it, it would help us to answer  $Q$ . With respect to such a  $Q$ , three situations are possible:

- Subduance, that is, new knowledge removes ignorance (e.g., by discovering an empirical explanation)
- Surrender, that is, we give up and do not look for an answer
- Abduction, that is, we set a hypothesis as a basis of new actions.

Abduction is thus an inference by means of which we do not solve the ignorance problem, but we overcome it in a certain way by setting a hypothesis. This hypothesis can then be released in further reasoning, something which allows for specific kinds of actions. In *Woods*' words, abduction “is a response that offers the agent a reasoned basis for new action in the presence of that ignorance” [14.4, p. 368]. Therefore, what

must be grasped here is that the conclusion of an abduction is not (necessarily) a true sentence or a new piece of knowledge; it is a hypothesis that can be used in further reasoning. The ignorance contained at the level of the premises is inherited by the conclusion. What is specific in the relation between premises and conclusions here is not a gain of knowledge, but rather an ignorance-preserving relation.

For reasons of clarity, the GW schema is formally presented following Woods' latest version in [14.4, p. 369]. Let  $T$  be an agent's epistemic state at a specific time,  $K$  the agent's knowledge base at that time,  $K^*$  an immediate successor base of  $K$ ,  $R$  an attainment relation for  $T$  (that is,  $R(K, T)$  means that the knowledge-base  $K$  is sufficient to reach the target  $T$ ),  $\rightsquigarrow$  a symbol denoting the subjunctive conditional connective, for which no particular formal interpretation is assumed, and  $K(H)$  the revision of  $K$  upon the addition of  $H$ .  $C(H)$  denotes the conjecture of  $H$  and  $H^c$  its activation. Let  $T!Q(\alpha)$  denote the setting of  $T$  as an epistemic target with respect to an unanswered question  $Q$  to which, if known,  $\alpha$  would be the answer. According to the GW schema, the general structure of abduction is as follows:

1.  $T!Q(\alpha)$
2.  $\neg(R(K, T))$  [fact]
3.  $\neg(R(K^*, T))$  [fact]
4.  $H \notin K$  [fact]
5.  $H \notin K^*$  [fact]
6.  $\neg R(H, T)$  [fact]
7.  $\neg R(K(H), T)$  [fact]
8.  $H \rightsquigarrow R(K(H), T)$  [fact]
9.  $H$  meets further conditions  $S_1, \dots, S_n$  [fact]
10. Therefore,  $C(H)$  [subconclusion, 1-7]
11. Therefore,  $H^c$  [conclusion, 1-8]

The aim, here, is to characterize what is specific to abductive inference, by taking into account what triggers such an inference, and to describe the subsequent process. At the beginning, a cognitive target  $T!Q(\alpha)$  is set (1): something we aim to reach in response to an ignorance problem. The ignorance problem triggers an abduction because it is a cognitive irritant, that is, it places us in an unpleasant situation of lack of knowledge which can be overcome by action and reasoning.

Step (2)  $\neg(R(K, T))$  says that the current knowledge is insufficient to attain the cognitive target. This is essential if we face an ignorance problem. Step (3),  $\neg(R(K^*, T))$ , says that there is no immediate successor of  $K$  by means of which the target would be attained. This is a crucial step. If there were such a  $K^*$ , we would just extend our knowledge by adding new information and would refrain from triggering anything such as an abduction. This would be subduance, that is, new knowledge would remove the initial ignorance.

If there is no  $K$  or  $K^*$  relating to the cognitive target, a hypothesis  $H$  is sought by the agent in order to set a plausible solution to the ignorance problem. Such a hypothesis is not knowledge, it is a hypothesis. This is represented in steps (4) and (5). Since it is only a hypothesis, it cannot relate to the cognitive target either, because it is not a solution. Even combined with the knowledge set, the cognitive target is not attained. This is expressed in steps (6) and (7).

What is the purpose of the hypothesis  $H$  if it does not solve the problem? In step (8), it is settled as a hypothesis that *subjunctively* relates to the cognitive target in combination with our knowledge base. What does this mean that it *subjunctively* relates to the cognitive target? This is how Gabbay and Woods understand Peirce's, *hence*, in the schema laid down by Peirce [14.55, 5.189] (see also Chap. 10 for the original formulation). It means that it is not a true sentence, it is not a piece of knowledge either, but if it were, it would give an acceptable solution to the cognitive problem. As in step (9), some additional conditions should be added for the acceptability of  $H$ .

Having set hypothesis  $H$  as a *subjunctive* solution of the cognitive problem, abduction first consists in concluding that we are right in conjecturing that hypothesis. This is the first subconclusion at step (10).  $C(H)$  means that the hypothesis  $H$  is conjectured. It is important to notice here that abduction does not end at step (10). Indeed, by taking seriously the fact that abduction is triggered by a cognitive problem, we trigger an abduction not to conjecture a hypothesis, but in order to find a possibility of further actions despite the lack of knowledge. Therefore, the abduction should not end before step (11), that is, when the conjecture is released and when the hypothesis is used in further reasoning as a basis for new action.  $H^c$  represents the hypothesis released in a further reasoning, that is, in a reasoning in which we act on the hypothesis  $H$  and the superscript  $c$  indicates the conjectural origin of the hypothesis. Following Woods [14.4, p. 371] an inference that ends at step (10) will be called a *partial abduction*, and an inference continuing with step (11) a *full abduction*.

For the purpose of clarity, in step (10) we face two possibilities. First, we do not test the hypothesis but we use it in a further reasoning (as in step (11)). This is precisely what is called *full abduction*. Second, we test the hypothesis, by empirical methods, for example. This presents us with three possibilities. First, the hypothesis is confirmed and we obtain a new piece of knowledge; this would lead to a situation similar to the  $K^*$  situation above. In this case, no full abduction is triggered, that is, we do not act on the hypothesis in an ignorance-preserving way. In fact, we would end with

new knowledge and this is subduance, or hypothetico-deductive reasoning, induction or even a mix, but this is not abduction. Second, we do not have confirmation and give up on the hypothesis: again this would end in a partial abduction. Third, we do not have confirmation but we continue with the hypothesis and perform a full abduction.

This is what leads Woods [14.4] to claim that abduction should not be understood as an inference to the best explanation (that would consist of the first part of the abduction schema, with respect to certain aspects), but rather as an inference *from* the best explanation. That is, we opt for a hypothesis and make an inference by activating that hypothesis. The following example of daily reasoning illustrates this point:

*Shahid* and *Ángel* are in Mexico City at the *Baranca del Muerto* underground station (Fig. 14.2). They want to go to a conference near *Universidad*. On their map, the new line (dotted line on the map) between *Mixcoac* and *Zapata* is missing. They do not know the existence of this line and decide to travel first to *Tacubaya*. During the trip to *Tacubaya*, they see a workmate disembarking the train at *Mixcoac*. They think their workmate will be late and they proceed to change in *Tacubaya*. There, they board another train to *Centro Médico* where they will change again to go to *Universidad*. When they arrive at the conference they are surprised to see their workmate already there. The fact to be disclosed is now that there is a faster way to go to *Universidad* which cannot be explained on the basis of the information contained on the incomplete map.

With respect to the previously detailed GW schema, step (1)  $T!Q(\alpha)$  is such that  $Q$  is the question of knowing how their workmate might have arrived so early. The cognitive target  $T$  would be a situation in which an  $\alpha$  is known such that  $\alpha$  would be the answer to that question. With respect to step (2), their knowledge base is insufficient to answer the question because their map does not show any another way to reach *Universidad* ( $\neg(R(K, T))$ ). In step (3), they receive no further knowledge (e.g., an updated map) to answer the question ( $\neg(R(K*, T))$ ). There are three possibilities: First, they do not care and follow the same trip as the day before (surrender). Second, they search for more information and obtain an updated map in which the line between *Mixcoac* and *Universidad* appears (subduance). Note that in this last case, no abduction is triggered, a new piece of information is added to the knowledge base (such that the new knowledge-base  $K^*$  explains why the workmate went faster the day before  $\neg R(K*, T)$ ). Third, they perform an abduction. That is, they conjecture the existence of the line and, therefore, they can leave half an hour later the following day. The exist-

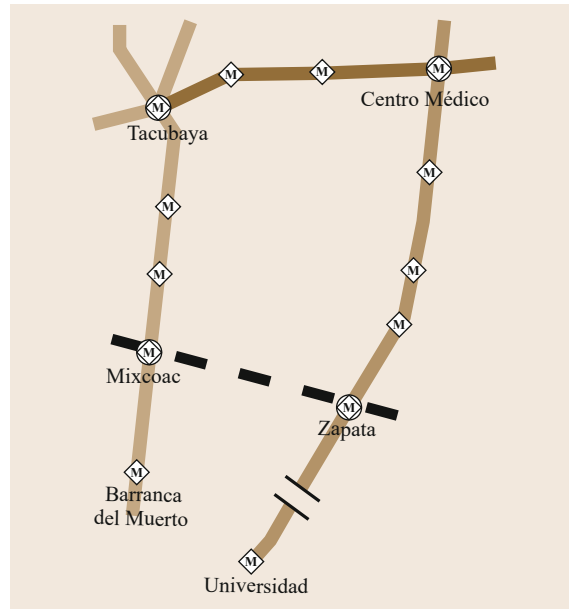


Fig. 14.2 México City Metro

tence of such a line is a hypothesis,  $H$ , and is such that  $H \notin K$  (step (4)) and  $H \notin K^*$  (step (5)), and is not part of any knowledge set. Therefore, step (6) holds because  $H$  is not an established fact and does not relate to the target ( $\neg R(H, T)$ ). Moreover, step (7) also holds because even combined with their knowledge-base  $K$ , it does not relate to the cognitive target ( $\neg R(K(H), T)$ ). Step (8) is crucial because  $H$  only subjunctively relates to the cognitive target, that is, the effective existence of another line might be such that, when added to the knowledge-base  $K$ , it would allow the cognitive target  $T$  to be reached. However,  $H$  is only a hypothesis and without further information, it does not constitute an  $\alpha$  answer to  $Q$  that would relate to the cognitive target  $T$ . If  $H$  meets further conditions ( $S_1, \dots, S_n$ ) it might be considered as a good or plausible explanation, perhaps the *best* one, as expressed in step (9), and that hypothesis would be conjectured as in step (10) ( $C(H)$ ). The following day, *Sahid* and *Ángel* stop at *Mixcoac* as if they knew the existence of this line, but in fact they do not. That is, step (11), they release the conjectured hypothesis ( $H^C$ ) and act upon it despite their persisting ignorance with respect to the genuine explanation of the initial problem.

The fact that an epistemic view of the abductive inference thus described would not grasp the specificity of abduction has to be emphasized. Indeed, the new hypothesis is not to be considered as a new piece of knowledge or belief. It might be accepted as an abductive conclusion and as a good explanation without being believed or accepted as *the* good explanation. What is

characteristic of an abduction is the conjectural aspect of its conclusion and the activation of the hypothesis in further reasoning. What is essential to an abduction is

that the cognitive target is not attained by a definitive solution of the initial problem at the level of the conclusion.

## 14.5 Abduction in Dialogical Logic

It is time to propose a dialogical understanding of abduction and its basic concepts. In previous work, Keiff [14.56, pp. 200ff] defines *abductive problems* in the context of substructural dialogues, namely dialogues in which optimal rules to defend a thesis are sought. Roughly speaking, in the context of substructural dialogue for modal logic for example, the Proponent is allowed to conjecture the accessibility relations that are needed for the defence of her thesis, while this kind of move is forbidden in modal dialogical logic. The framework now outlined is rather different and more general, even if some basic problems remain similar.

Although some given explanations have their roots in the GW model and Aliseda's characterization of abduction, the aim is not to give a faithful dialogical formalization of these approaches. The point is rather to identify the general features of abductive dialogues following these three main questions:

- How can a surprising fact, an *abductive problem* and (or) an ignorance problem be characterized?
- How can the guessing step, in which a hypothetical explanation is conjectured, be characterized?
- How can the ignorance-preserving feature of abduction be characterized?

The first question relates to the conditions under which an *abductive problem* may be stated in a game, that is, the *triggering* of an abductive dialogue must be described. The second question relates to the possibility of conjecturing an explicative hypothesis during the dialogue, that is, the act of *guessing* specific to an abductive dialogue must be described. The third question relates to the conjectural status of the explicative hypothesis in a dialogue, that is, the question of the *commitment* must be asked. Indeed, what are we committed to when we conjecture a hypothesis? This last question is more complicated and involves in-depth considerations about the defeasible aspect of conjectures. In this chapter, this difficulty will be explained in terms of *not-conceded preservation*, that is, a hypothesis that has not been conceded by the opponent, remains not conceded at the end of an abductive dialogue even if the Proponent has conjectured it. In fact, it is the general point of this contribution to propose a dialogical explanation of abduction in terms of what is called a *concession problem*.

The leading idea is to allow the Proponent to claim *I am facing an abductive problem* when this agent has no winning strategy for a thesis  $\Phi$  given some shared knowledge or an accepted theory  $\Theta$ . As explained below, the theory  $\Theta$  is represented as a set of initial concessions of the Opponent.  $\Phi$  is the thesis that the Proponent cannot explain on the basis of the initial concessions, that is,  $\Phi$  holds for a *surprising* fact. If the Proponent is able to justify which kind of *abductive problem* she is being faced, then she triggers a subdialogue in which she is allowed to conjecture a hypothetical explanation  $\alpha$ .

The Proponent then has to show that this conjecture is such that if it had been conceded, it would have enabled her to explain the surprising fact expressed by  $\Phi$ . However, the hypothetical explanation  $\alpha$  remains not-conceded at the end of the dialogue and the Proponent only *subjunctively* and *defeasibly* wins. To parallel the explanation to the GW model, let the target  $T$  of the Proponent be the situation in which she wins the dialogue, the question  $Q$  be a set of challenges she is unable to answer, and  $\alpha$  that which would enable her to defend herself against those challenges. When the Proponent conjectures a hypothesis  $\alpha$ , the target is only subjunctively attained. In other words, the Proponent is in a situation, such that *if  $\alpha$  had been conceded, she would have explained  $\Phi$* . The explicative feature of the conjectured  $\alpha$  may thus be interpreted in the spirit of the *subjunctive attainment relation* ( $\rightsquigarrow$ ) of the GW model of abduction. The situation also parallels the  $\neg R(H, T)$  and  $\neg R(K(H), T)$  of the GW model because the Proponent does not actually reach the target. Note that this process, which will be described in this section, is simply partial abduction. In order to attain a characterization of full abduction, it should be explained how to release the conjectured hypothesis in a further dialogue and how to act upon it. This requires in-depth considerations about the specificity of the speech act by means of which such a hypothesis is conjectured. This issue will be dealt with in Sect. 14.6.

### 14.5.1 Triggering

The triggering of an abductive dialogue is characterized within what is called by Rahman and Keiff [14.31] and Rahman and Tulenheimo [14.57] *material dialogues*,

**Table 14.3** Dialogue 2

	O		P	
$\Theta_1$	$A \rightarrow B$			0
$\Theta_2$	$B \rightarrow (C \wedge \neg E)$		$C \wedge \neg E$	
$\Theta_3$	$D \rightarrow (C \wedge \neg E)$			
1	$m := 1$		$n := 2$	2
3	$? \wedge_1$			

that is, dialogues with the standard rules plus initial concessions of the Opponent. For example, let  $\Theta$  be a theory or a knowledge base consisting of several sentences, and  $\Phi$  be the initial thesis. A material dialogue begins with **O** conceding all the formulae contained in  $\Theta$  and with the initial thesis uttered by **P**. The dialogue then runs as usual. The following example, with  $\Theta = \{A \rightarrow B, B \rightarrow (C \wedge \neg E), D \rightarrow (C \wedge \neg E)\}$  and  $\Phi = C \wedge \neg E$ , is given in Table 14.3.

In Dialogue 2 (Table 14.3), **O** concedes the formulae of  $\Theta$  numbered  $\Theta_1, \dots, \Theta_n$ . **P** states the initial thesis  $C \wedge \neg E$ . Both choose a rank. At move 3, **O** challenges the initial thesis by asking for the first conjunct. **P** cannot answer because the required  $C$  is an atomic formula. There is no winning strategy for **P** for  $\Phi$ , given the initial concessions  $\Theta_1, \dots, \Theta_n$ . If  $\Theta_1, \dots, \Theta_n$  is interpreted as the shared (assumed) knowledge and  $\Phi$  as holding for a fact, then it might be said that  $\Phi$  represents a surprising fact which cannot be explained by the current knowledge. The idea is now that in such a situation, **P** has to be allowed to claim that she is facing an *abductive problem*. In fact, as it is clear in the dialogue (Table 14.3), an *abductive problem* is triggered by a *concession problem*: **P** cannot explain her thesis on the basis of the concessions.

The notion of *abductive problem* is dialogically defined following Aliseda's [14.50, p.47] definitions of abductive novelty and abductive anomaly (see also Chap. 10): **P** will now be allowed to claim *I am facing an abductive novelty* or *I am facing an abductive anomaly*, as in the rules [SR-AN] and [SR-AA] below. In the first case, **P** is committed to show that neither  $\Phi$  nor  $\neg\Phi$  is entailed by  $\Theta$ . In other words, **P** has no winning strategy for  $\Phi$  nor for  $\neg\Phi$ , given  $\Theta$ . In the second case, **P** is committed to show that  $\Phi$  is not entailed by  $\Theta$  while  $\neg\Phi$  is. That is, **P** has a winning strategy for  $\neg\Phi$  but not for  $\Phi$  given  $\Theta$ . Here, the technical difficulty is that the Proponent would need a *losing strategy* in order to justify she is facing an abductive novelty or an abductive anomaly. How strange such a game would be!

**Table 14.4** Particle rule for  $\mathcal{F}$  operator

Assertion	Challenge	Defence
$X \text{--}! \text{--} \mathcal{F}_\Theta \Phi \text{--} d_1$	$Y \text{--}? \mathcal{F}_\Theta \text{--} d_{1,i}$	$X \text{--}! \text{--} \neg(\Theta \rightarrow \Phi) \text{--} d_{1,i}$
	$Y$ opens a subdialogue $d_{1,i}$ .	

The difficulty is easily overcome by making use of the *attackability* operator  $\mathcal{F}$  introduced by Rahman and Rückert [14.58] in their dialogical connexive logic. This  $\mathcal{F}$  operator allows the Proponent to claim that under some conditions, the formula in the scope of that operator cannot be defended. Here a subscript is used in order to apply this operator in material dialogues. That is, if **X** says  $\mathcal{F}_\Theta \Phi$ , then she is claiming that  $\Phi$  may be attacked given the premises  $\Theta$ . The rule is stated in Table 14.4.

Note that the rule is formulated with an indication for what is called *section of a dialogue*, i. e. main dialogue ( $d_1$ ) and subdialogue ( $d_{1,i}$ ), in accordance with the following definitions:

**[D2] [Main Dialogue  $d_1$ ]**

The first section of a game in which **P** defends an initial thesis is called *main dialogue*  $d_1$ .

**[D3] [Subdialogue]**

The subdialogue triggered by the challenge of an  $\mathcal{F}$  operator in a section  $d_1$  or the subdialogue triggered by an *AS-challenge* is called *subdialogue*  $d_{1,i}$ . Note that this rule allows subdialogues of subdialogues (i. e., a subdialogue  $d_{1,i,i}$  triggered in a subdialogue  $d_{1,i}$ , for example in the case of an *AS-challenge*, as defined in the next section).

The main idea is that **X** justifies  $\mathcal{F}_\Theta \Phi$  by showing that  $\Phi$  cannot be justified, given  $\Theta$ . This supposes a device that allows switches of the burden of proof, in addition to the particle rule for the  $\mathcal{F}$  operator. This presupposes a generalization of the formal rule by means of a structural rule which says that the player who plays formally (i. e., the player who cannot introduce atomic formula) is the player who challenges an  $\mathcal{F}$  operator (or the player who defends an *AS-move*). In other words, the argumentation partner who challenges a formula such as  $\mathcal{F}_\Theta \Phi$  will have to take the burden of the proof by defending  $\Phi$  under the formal restriction.

**[SR2.1] [Formal Restriction]**

Let  $d_n$  be a section of a dialogue (main dialogue or subsection): If **X** plays under formal restriction in  $d_n$ , then **X** is not allowed to utter an atomic formula unless **Y** uttered the same atomic formula before in the same section  $d_n$ .

The rule governing the application of the formal restriction is now defined as follows:

**Table 14.5** Dialogue 2.1

O		$d_1$		P	
...	...	...	...	...	...
3	? $\wedge_1$				
				$\mathcal{F}_\Theta(C \wedge \neg E) \wedge \mathcal{F}_\Theta \neg(C \wedge \neg E)$	4
5	? $\wedge_1$	4		$\mathcal{F}_\Theta(C \wedge \neg E)$	6
$d_{1,1}$					
7	? $\mathcal{F}_\Theta$	6		$\neg(\Theta \rightarrow (C \wedge \neg E))$	8
9	$\Theta \rightarrow (C \wedge \neg E)$	8		---	
11	$C \wedge \neg E$		9	$\Theta$	10
...	...	...	...	...	...

**[SR2.2] [Application of the Formal Restriction]**

The application of the formal restriction is regulated by the following conditions:

1. In the main dialogue  $d_1$ , if  $\mathbf{X} = \mathbf{P}$ , then  $\mathbf{X}$  plays formally.
2. If  $\mathbf{X}$  opens a subdialogue  $d_{1,i}$  by challenging an  $\mathcal{F}$  operator, then  $\mathbf{X}$  plays formally.
3. If  $\mathbf{X}$  opens a subdialogue  $d_{1,i}$  by challenging an AS-move, then  $\mathbf{Y}$  plays formally.

Now, structural rules that allow the Proponent to claim she is facing an *abductive problem* are added. She has the choice between the two kinds of *abductive problems* previously defined:

**[SR-AN] [Utterance of Abductive Novelty]**

When  $\mathbf{P}$  loses a game playing deductively (i. e., with the standard rules of dialogical logic), then  $\mathbf{P}$  is allowed to claim that she is facing an abductive novelty by saying  $\mathcal{F}_\Theta \Phi \wedge \mathcal{F}_\Theta \neg \Phi$ .

In the same way, the *Proponent* is allowed to choose between the claim that she is facing an abductive novelty or an abductive anomaly. The following rule is therefore added:

**[SR-AA] [Utterance of Abductive Anomaly]**

When  $\mathbf{P}$  loses a game playing deductively (i. e., with the standard rules of dialogical logic), then  $\mathbf{P}$  is allowed to claim that she is facing an abductive novelty by saying  $\mathcal{F}_\Theta \Phi \wedge \neg \Phi$ . Note here that the second conjunct does not trigger any subdialogue  $d_{i,i}$  and that  $\neg \Phi$  has to be defended in the same context as the initial thesis, namely when the concession of  $\Theta$  is given by  $\mathbf{O}$ .

Without going into excessive details, this point is explained with an example based on Dialogue 2, (Table 14.3).

In Dialogue 2.1 (Table 14.5),  $\mathbf{P}$  loses and now claims she is facing an abductive novelty at move 4.  $\mathbf{O}$  challenges that move at move 5,  $\mathbf{P}$  answers by giving the first conjunct.  $\mathbf{O}$  then challenges the  $\mathcal{F}$  operator

at move 7 by opening a subdialogue in which she now plays formally.  $\mathbf{P}$  answers by saying  $\neg(\Theta \rightarrow (C \wedge \neg E))$ , where  $\Theta$  is the same set of initial concessions as before, and the dialogue runs as usual. It is easy to verify that  $\mathbf{O}$  will lose (for the same reasons  $\mathbf{P}$  lost in Dialogue 2, Table 14.3). In the same way,  $\mathbf{O}$  will lose even if she challenges the second conjunct in move 4, and  $\mathbf{P}$  will have justified that she was facing an abductive novelty.

**14.5.2 Guessing**

After having shown she was facing an *abductive problem*, the Proponent has to guess what is missing in order to solve it. The Proponent has to be allowed to conjecture a hypothetical *abductive solution*. Such a conjecture is made by means of a subjunctive speech act, that is, a speech act that does not rely on the concessions of the Opponent. In fact, such a move should be rendered as a feasible move. Note that it will not be explained how an *abductive solution* is generated or chosen. Instead, the present proposal will be to describe the conditions under which such a conjectural move is performed.

This contribution relies again on *Aliseda* [14.50] who proposes a calculus for abduction, based on the semantic trees of deductive logic, but with some nuances with respect to the status of the abductive explanation. Roughly speaking, the idea is to construct the tree for  $\Theta \rightarrow \Phi$  and identify its open branches together with such formulae that may close those branches (in a consistent way). Several formulae may do the job. These formulae are called by *Aliseda* *abductive solutions* to *abductive problem* consisting of the pair  $\Theta, \Phi$ . Therefore, an abduction consists in guessing (or discovering) what the possible *abductive solutions* are. To put it in *Aliseda's* own words, consistent abductions are those formulae which “if they had been in the theory before, they would have closed those branches that remain open after  $\neg \Phi$  is incorporated into the tableau” [14.50, p. 110].



**Table 14.6** AS-challenge 1

Utterance	Challenge	Defence
$X-!-SA : \alpha - d_{1,i}$	$Y-?-?Plain_{AS} - d_{1,i,i}$	$X-!- (\Theta \wedge \alpha) \rightarrow \Phi - d_{1,i,i}$
	Y opens a subdialogue $d_{1,i,i}$	

**Table 14.7** AS-challenge 2

Utterance	Challenge	Defence
$X-!-SA : \alpha - d_{i,i}$	$Y-?-?Explanatory_{SA} - d_{i,i,i}$	$X-!- ((\Theta \wedge \alpha) \rightarrow \Phi)$ $\wedge (\neg(\Theta \rightarrow \Phi) \vee (\alpha \rightarrow \Phi)) - d_{i,i,i}$
	Y opens a subdialogue $d_{1,i,i}$	

Even if, from a dialogical viewpoint, it is not looked for any *true* formula, what is an *abductive solution* may be defined following a similar process. Indeed, after having shown that she was facing an *abductive problem*, the Proponent should be allowed to put forward a hypothesis. What the Proponent has to look for is a formula, not conceded by the Opponent, that enables her to win the dialogue previously lost. Therefore, a rule that allows the Proponent to conjecture the hypothesis of an explanation called *abductive solution* is added:

#### [SR-SA] [Abductive Solution Rule]

When the Proponent has won the subdialogue triggered by the challenge of the  $\mathcal{F}$  operator, whether it be novelty or anomaly, the Opponent is allowed to ask her ?AS (i. e., she claims *do you have an abductive solution to propose?*). If so, the Proponent answers  $AS : \alpha$  (i. e., she claims  $\alpha$  is my abductive solution).

What does it mean that  $\alpha$  is an *abductive solution* for the Proponent, and why is that *abductive solution* the conjecture of a hypothesis? In fact, this move consists in claiming that there is a plausible explanation to the surprising fact  $\Phi$  given  $\Theta$ . This specific move,  $AS : \alpha$ , is the move that forces to reconsider dialogical games to fit in with abductive reasoning. Indeed, it may consist of the utterance by the Proponent of an atomic formula not previously conceded by the Opponent. Nevertheless, the introduction of this new piece of information is to be understood as a subjunctive explanation. That is, the Proponent introduces  $\alpha$  as she would say *if you had conceded me  $\alpha$ , I would have been able to explain  $\Phi$* . In no way is  $\alpha$  introduced as an **O**-concession to be incorporated into  $\Theta$  or into a  $\Theta'$ , a successor of  $\Theta$  containing the initial concessions and the other concessions made during the dialogue.  $\alpha$  is a new formula that may be used in further reasoning, but only temporarily, and that temporarily nature requires further justification. Indeed, as a hypothesis,  $\alpha$  is defeasible, that is, it is a conclusion *faute de mieux* guessed by the Proponent. If it is shown later that this is not a good explanation or if a counter-example is encountered, then  $\alpha$  will be defeated and removed. Its

conditions of use are not the same as the usual assertions of the standard dialogical logic because it is subject to further justification, no matter whether it is an atomic or a complex formula: Would it be a new kind of utterance?

### 14.5.3 Committing

The dialogues defined here only describe a partial abduction, that is, an *abductive problem* is set and a plausible answer is guessed. However, in order to characterize a full abduction, it should be explained how the conjecture might be released in a further dialogue and how the players might act upon it. As already explained, Gabbay and Woods characterize abduction as an ignorance-preserving inference. It has been shown that abductive dialogues are not-conceded-preserving: The explicative conjecture remains not-conceded and the Proponent only gives a subjunctive explanation for the surprising fact. The difficulty at this point involves the clarification of the commitment carried by such conjectural moves, which are rather different from the usual assertions.

Even if the question of the commitment of the conjectural move is very complex (it might even vary according to the argumentation contexts), a rule to deal with the consequence requirement of the type of abduction called “plain abduction” by *Aliseda* [14.50, Part II] can be defined. In a dialogue, this consists in adding the possibility of a challenge on the  $AS$ -move, called an  $AS$ -challenge. The Opponent makes the request to justify that it is sufficient to consider the conjunction of  $\Theta$  and  $\alpha$  to derive  $\Phi$  by means of the rule in Table 14.6.

Under this rule, the challenger opens a subdialogue in which the defender will have to defend the condition  $(\Theta \wedge \alpha) \rightarrow \Phi$ . The act of  $Y$  opening a subdialogue means that  $X$  will play under formal restriction. The formal restriction is applied in accordance with the rules [SR2.1] and [SR2.2] given earlier. More kinds of such challenges should be defined to complete the picture. Adding the explanatory character of  $\alpha$  might also be required. Thus, the possibility to chose another attack against an  $AS$ -move is offered to  $Y$  (Table 14.7).

Other requirements, such as consistency ( $\Theta, \alpha \not\vdash \perp$ ), minimality and so on (Chap. 10), might be added in the same way. It would also be possible to rely on these rules in order to deal with the defeasibility of AS-moves. Indeed, if a player is not able to answer the AS-challenges performed by her argumentation partner, then her conjectural move should be removed and considered as null. In the same way, if some counterexamples or a better explanation are found, the AS-moves should also be cancelled. However, defeasibility is a very wide topic and cannot be dealt with in detail in this chapter. A nonmonotonic account of abduction that makes use of adaptive logic is given by *Meheus and Batens* [14.59] and *Beirlaen and Aliseda* [14.60] (see also Chap. 12 by *Gauderis*). For a dialogical study of defeasible reasoning, see the work of *Nzokou* [14.61]. For a nonmonotonic treatment of inconsistencies in the context of an adaptive dialogical logic, see *Rahman and Van Bendegem* [14.62].

What has been characterized in this section is only partial abduction. In order to attain a full abduction, the framework over which dialogues are obtained and in which the hypothesis  $\alpha$  is released in the defence of

another thesis *as if* it had been conceded, should be developed. Indeed, full abductive dialogue should be not-conceded-preserving, that is, the agents act upon the hypothesis although it has not been conceded by the Opponent. This is the dialogical understanding of ignorance-preservation in the GW model of abduction defended in this contribution. In the GW schema, it was said that neither  $R(K(H), T)$  nor  $R(K * (H), T)$  were the case. Here, this parallels the fact that **P** does not actually attain the target. **P** only encounters something similar to a subjunctive winning strategy, a strategy which would lead to the victory if **O** had conceded  $\alpha$ ; similarly in the GW model, it is only a subjunctive attainment relation expressed by  $H \rightsquigarrow R(K(H), T)$ . Now, the challenge faced in order to complete the picture and to define the conditions of use of a hypothetical explanation  $\alpha$  in a full abduction, consists in providing an in-depth analysis of the commitment carried by such a conjecture. This relates to the following question: What kind of speech act is at stake when a hypothesis is conjectured? Without a precise answer to this question, no precise rule of victory for abductive dialogues can be yet formulated.

## 14.6 Hypothesis: What Kind of Speech Act?

In the previous section, a new kind of move specific to abductive dialogues, the so-called AS-move, by means of which a hypothetical *abductive solution* is conjectured, has been introduced. Such a move is considered as a subjunctive move, that is, a move stated hypothetically with an assumption such as *if you had conceded me  $\alpha$ , I would have been able to justify  $\Phi$* . The conditions under which it is possible to conjecture an *abductive solution* and how such a hypothetical *abductive solution* might be challenged have been clarified. However, by means of what kind of speech act is an AS-move performed? What kind of speech act is the conjecture of a hypothetical *abductive solution* if  $\alpha$  can be used in the defence of another thesis (in a full abduction)?

An epistemic explanation might have seemed attractive, relying for example on the notion of subjunctive knowledge defined by *Rückert* [14.63]. Subjunctive knowledge is defined in a modal frame as the knowledge people of another world would have about the actual world. Abduction might thus be thought of in terms of subjunctive epistemic change, namely if some people of another world had the knowledge of what is expressed by the hypothesis, they would be able to explain a surprising fact in the actual world. This would smartly explain the subjunctive status of the

explanatory relation conjectured in a hypothetical *abductive solution*. However, it would have ended up in an account explicitly involving the epistemic states of the agents instead of taking into account their actions. Moreover, such an account would yield an excessively strong commitment on the part of the agent with respect to the belief or the knowledge of the truth of the hypothetical *abductive solution*. However, as explained earlier, this is not necessary. An *abductive solution* can be conjectured as being plausible without any commitment to the belief of the truth of what is expressed.

This last point brings back the problem of the status of an AS-move. Is it an assertive speech act? How could it be? An assertive speech act is usually characterized by the commitment (of the speaker) to its truth. In his theory of speech acts, *Searle* [14.64, p. 12] defines the class of assertive speech act as follows:

“The point or purpose of the members of the assertive class is to commit the speaker (in varying degrees) to something being the case, to the truth of the expressed proposition.”

Although, in dialogical logic, the commitment to the truth is irrelevant in the characterization of an assertion, assertion can be thought of in terms of com-

mitment to justify what is said (by defending it against further challenges or by relying on the concessions of the Opponent). What about the AS-move? It is conjectured and might be released in another dialogue without being conceded by the Opponent or fully justified by the Proponent. Therefore, is the conjecture of a hypothesis an assertive utterance in the dialogical sense of the term? In Searle's terms, is conjecturing an assertive act? It seems that it cannot be. Answering these questions is crucial if the aim is to succeed in introducing the AS-move defeasibly and to release the conjectures in further reasoning in the same way as in the GW schema of abduction in a dialogical framework.

If the speech act, by means of which a hypothetical *abductive solution* is conjectured, is not an assertion, would it be a commissive speech act? Beyond the question of the commitment to the truth or to belief, or even to the acceptance of what is uttered, an *abductive solution* commits the speaker to a subsequent series of actions. First, the speaker is committed to answer the AS-challenges. Second, the use of the hypothesis in a full abduction without knowing whether it is true or not, might be seen as a peculiar kind of commitment. Does such a peculiar commitment relate to what Searle has called the commissive speech acts? More precisely, Searle defines the commissives as "those illocutionary acts whose point is to commit the speaker (again in varying degree) to some future course of action" [14.64, p. 14]. In the dialogical approach, which has been outlined in the previous section, the underlying idea is that the Proponent conjectures a hypothetical *abductive solution* which is such that, if it had been conceded, it would have explained the surprising fact. However, this should not be the end of the story because the aim would be to release the hypothesis in a further reasoning: in another dialogue in which the Proponent defends another thesis by acting on the hypothesis at stake. That is why the commitment carried out by the speech act, by means of which an AS-move is performed, indicates that it could be understood in terms of a commissive speech act. In addition to further justification, it also commits the agent to further dialectical actions. Nevertheless the commissives are usually speech acts in which the agents commit themselves to an action over which they have full control. That is to say, the commissive speech act commits to something that depends only on the agents, as it happens in the case of promises and oaths. However, the agent who performs an abduction does not have full control of the explanatory force of an *abductive solution*. Indeed, while in the first case

the failure of the promise is dependent upon the agent herself, the failure of an abductive explanation includes a wider range of factors which do not exclusively depend on the agent activity. So, it does not seem that the speech act by means of which an AS-move is performed, is a commissive speech act.

If it consists in neither an assertive nor a commissive act, would a conjectural move be a fictional speech act? Indeed, according to Searle [14.65], fictional discourse is not composed of genuine assertions but instead of pretended assertions. The point is that in fiction, even if the author is not committed to the truth of what she says, she does not have the intention to lie. Therefore, the author does not tell the truth but neither is the author lying. The author tells a story doing *as if* she were asserting. When a player performs an AS-move and uses it in a further reasoning, she does *as if* it were conceded. She does not have to believe what she says but neither is she trying to mislead the interlocutor. However, beyond the fact that Searle's theory of fictionality is not shared by this contribution, it is thought that abduction has a practical dimension, which is not necessary to the fictional discourse. Hence, in this chapter, it is not believed that the hypothetical speech act should be explained in terms of fictional discourse. Moreover, what is to be explained while studying fiction is its double aspect, the fact that while we know it is not true we react to such a discourse without experiencing any kind of cognitive dissonance. Also there is no such tension to be explained in the conjecture of a hypothetical *abductive solution*. (For more details on these points, see [14.66–68].)

An alternative, though tentative solution, would be to reconsider the taxonomy of speech acts. For example, Bach [14.69] defines the wider category of constative, which the conjecturing act would be part of. Other inspirations might be found in the work of Barés Gómez [14.70] who distinguishes between different kinds of assertions in natural language (assertive paradigm, negative paradigm, and evidentiality) by making use of Dynamic Epistemic Logic and by focusing on the transmission of information. These different kinds of assertions might also be understood as different types (talking thus about hypothetical judgement); see the recent work of Rahman and Clerbout [14.71], on Constructive Type Theory in the context of dialogical logic follows in this respect. The question is left as a challenge for further investigations. Is a hypothetical speech act a particular kind of assertive or commissive act? Is it a mix of both? Is it a completely new kind of speech act?

## 14.7 Conclusions

In this chapter, it is first advocated for a reconciliation of argumentation theory and formal logic in an agent-centered theory of reasoning, that is, a theory in which inferences are studied in terms of human activities. More precisely, the dialogical approach to logic, in which reasoning is studied through a dialectical interaction between the Proponent of a thesis and the Opponent of it, is defended. In this context, the necessity of taking into account, not only the actions of the agents, but also the importance of the notion of commitment is stressed. Beginning with deductive dialogues, the picture has been extended to abduction, which is considered as a case of nondeductive reasoning.

The starting point to deal with abduction is the agent-centered analysis of the GW model. While Gabbay and Woods identify abduction as an ignorance-preserving inference triggered by an ignorance problem, abductive dialogues have been defined here as not-conceded-preserving dialogues triggered by a *concession problem*. The specificity of abductive dialogues has been identified at the level of the so-called AS-moves by means of which hypothetical *abductive solutions* are conjectured. To allow such moves, new rules have been put forward. The challenge for dialogicians now consists in exploring the release of such hypotheses in further dialogues in which they remain not-conceded. However, the difficulty of defining the

nature of such a hypothetical speech act is being faced, which leads to the key question of commitment. What are we committed to when we conjecture a hypothetical explanation of a surprising fact and when we release such a hypothesis in further reasoning? A definite answer to this question is left for further investigations.

**Acknowledgments.** Research for this chapter was supported by the project *Logics of discovery, heuristics and creativity in the sciences* (PAPIIT, IN400514-3) granted by the National Autonomous University of Mexico (UNAM) and by the project *Interpretaciones alternativas de lógicas no clásicas, IALNoC* (P10-HUM-5844) granted by Junta de Andalucía (Consejería de Innovación, Ciencia y Empresas). Matthieu Fontaine is greatly indebted to the Dirección General de Asuntos del Personal Académico (UNAM) and to the Programa de Becas Posdoctorales de la Coordinación de Humanidades (UNAM). We thank Atocha Aliseda and Mathieu Beirlaen for their comments. We are also thankful to Shahid Rahman for fruitful discussions on these topics (some of his arguments were detailed in his work *What Is Wrong about Perelman-Toulmin's Opposition between Legal Reasoning and Logic?*, JURIOLOG conference, Lille, 2014 and some ideas were suggested during our talk *Transmission de l'information dans les pratiques argumentatives. Evidentialité dans une sémantique dialogique*, Ve SPS, Lille, 2014).

## References

- 14.1 S.E. Toulmin: *The Uses of Argument* (Cambridge Univ. Press, Cambridge 2003)
- 14.2 C. Perelman, L. Olbrechts-Tyteca: *Traité de l'Argumentation: La Nouvelle Rhétorique* (Presses Univ. de France, Paris 1958), in French
- 14.3 D. Walton, E. Krabbe: *Commitment in Dialogue. Basic Concepts of Interpersonal Reasoning* (SUNY Press, Albany 1995)
- 14.4 J. Woods: *Errors of Reasoning. Naturalizing the Logic of Inference* (College Publications, London 2013)
- 14.5 D. Gabbay, J. Woods: Advice on abductive logic, *Log. J. IGPL* **14**(2), 189–219 (2006)
- 14.6 D. Hitchcock, B. Verheij (Eds.): *Arguing on the Toulmin Model. New Essays in Argument Analysis and Evaluation* (Springer, Dordrecht 2006)
- 14.7 J.F.A.K. van Benthem: On logician's perspective on argumentation, *COGENCY* **1**(2), 13–25 (2009)
- 14.8 L.E.J. Brouwer: *Intuitionisme en formalisme* (Noordhoff, Groningen 1912), in Dutch
- 14.9 P. Lorenzen: *Einführung in die Operative Logik und Mathematik* (Springer, Berlin 1955), in German
- 14.10 P. Lorenzen, K. Lorenz: *Dialogische Logik* (Wissenschaftliche Buchgesellschaft, Darmstadt 1978), in German
- 14.11 P. Gochet: The dynamic turn in twentieth century logic, *Synthese* **130**(2), 175–184 (2002)
- 14.12 J. Hintikka: *Knowledge and Belief. An Introduction to the Logic of the Two Notions* (Cornell Univ. Press, Ithaca 1962)
- 14.13 G. Priest: *Towards Non-Being: the Semantics and Metaphysics of Intentionality* (Oxford Univ. Press, Oxford 2005)
- 14.14 C.E. Alchourrón, P. Gärdenfors, D. Makinson: On the logic of theory change: Partial meet contraction and revision functions, *J. Symb. Log.* **50**(2), 510–530 (1985)
- 14.15 J. Groenendijk, M. Stokhof: Dynamic predicate logic, *Linguist. Philos.* **14**(1), 39–100 (1991)
- 14.16 A. Baltag, H. van Ditmarsch, L.S. Moss: Epistemic logic and information update. In: *Handbook on the Philosophy of Information*, ed. by P. Aderiaans, J.F.A.K. van Benthem (Elsevier, Amsterdam 2008)

- 14.17 H. van Ditmarsch, W. van der Hoek, B. Kooi: *Dynamic Epistemic Logic* (Springer, Dordrecht 2008)
- 14.18 J.F.A.K. van Benthem: *Logical Dynamics of Information and Interaction* (Cambridge Univ. Press, Cambridge 2014)
- 14.19 R. Fagin, J.Y. Halpern, Y. Moses, M.Y. Vardi: *Reasoning About Knowledge* (Bradford, The MIT Press, Cambridge 2004)
- 14.20 R. Morado: Problemas filosóficos de la lógica no-monótona. In: *Filosofía de la Lógica: Enciclopedia Iberoamericana de Filosofía*, Vol. 27, ed. by R. Orayen, A. Moretti (Trotta, Madrid 2005), in Spanish
- 14.21 R. Koons: Defeasible Reasoning. The Stanford Encyclopedia of Philosophy, ed. by E. N. Zalta, Spring Ed., <http://plato.stanford.edu/archives/spr2013/entries/reasoning-defeasible> (2013)
- 14.22 J.L. Pollock: Defeasible reasoning, *Cogn. Sci.* **11**, 481–518 (1987)
- 14.23 R. Reiter: A logic for default reasoning, *Artif. Intell.* **13**, 81–137 (1980)
- 14.24 J. McCarthy: Circumscription – A form of non-monotonic reasoning, *Artif. Intell.* **13**(1–2), 27–39 (1980)
- 14.25 M.L. Ginsberg (Ed.): *Readings in Non-Monotonic Reasoning* (Morgan Kaufmann, Los Altos 1987)
- 14.26 V.W. Marek, M. Truszczyński: *Nonmonotonic Logic. Context-Dependent Reasoning* (Springer, New York 1993)
- 14.27 J.F. Horty: *Reasons as Default* (Oxford Univ. Press, Oxford 2012)
- 14.28 D. Batens: A universal logic approach to adaptive logics, *Log. Univers.* **1**, 221–242 (2007)
- 14.29 H. Prakken, G. Vreeswijk: Logics for defeasible argumentation. In: *Handbook of Philosophical Logic*, 2nd, Vol. 4, ed. by D. Gabbay, F. Guenther (Kluwer Academic, Dordrecht 2002) pp. 219–318
- 14.30 S. Rahman: *Über Dialogue, Protologische Kategorien und andere Seltenheiten* (Peter Lang, Bern 1993), in German
- 14.31 S. Rahman, L. Keiff: On how to be a dialogician. In: *Logic, Thought and Action*, Logic, Epistemology and the Unity of Science, Vol. 2, ed. by D. Vanderveken (Springer, Dordrecht 2005) pp. 359–408
- 14.32 M. Fontaine, J. Redmond: *Logique dialogique. Une Introduction* (College Publications, London 2008), in French
- 14.33 N. Clerbout: *La sémantique dialogique: Notions fondamentales et éléments de métathéorie* (College Publications, London 2014), in French
- 14.34 E.M. Barth, E. Krabbe: *From Axiom to Dialogue: A Philosophical Study of Logic and Argumentation* (de Gruyter, Berlin 1982)
- 14.35 E. Krabbe: Formal systems of dialogue rules, *Synthese* **63**, 295–328 (1985)
- 14.36 F.H. van Emereen, R. Grootendorst: *A Systematic Theory of Argumentation. The Pragma-Dialectical Approach* (Cambridge Univ. Press, Cambridge 2004)
- 14.37 F.H. van Emereen, P. Houtlosser, A.F. Snoeck Henkemans: *Argumentative Indicators in Discourse. A Pragma-Dialectical Study* (Springer, Dordrecht 2007)
- 14.38 L. Wittgenstein: *Philosophical Investigations* (Blackwell, Oxford 1953), ed. by G. Anscombe, R. Rhees
- 14.39 A. Prior: The runabout inference-ticket, *Analysis* **21**(2), 38–39 (1960)
- 14.40 S. Rahman, N. Clerbout, L. Keiff: On dialogues and natural deduction. In: *Acts of Knowledge: History, Philosophy and Logic. Essays Dedicated to Göran Sundholm*, ed. by G. Primiero, S. Rahman (College Publications, London 2009)
- 14.41 S. Rahman: Negation in the logic of first degree entailment and tonk: A dialogical study. In: *The Realism-Antirealism Debate in the Age of Alternative Logics*, ed. by S. Rahman, G. Primiero, M. Marion (Springer, Dordrecht 2010) pp. 175–201
- 14.42 N. Clerbout: First-order dialogical games and tableaux, *J. Philos. Log.* **43**(4), 785–801 (2014)
- 14.43 W.V. Quine: *Philosophy of Logic*, 2nd edn. (Harvard Univ. Press, Cambridge 1986)
- 14.44 S. Rahman, M. Rückert, H. Fischmann: On dialogues and ontology. The dialogical approach to free logic, *Log. Anal.* **160**, 357–374 (1997)
- 14.45 C. Barés Gómez, S. Magnier, F. Salguero (Eds.): *Logic of Knowledge. Theory and Applications* (College Publications, London 2012)
- 14.46 M. Fontaine, S. Rahman: Fiction, creation and fictionality – An overview, *Methodos* (2010), doi:10.4000/methodos.2343
- 14.47 S. Rahman, H. Rückert: Dialogische modallogik (für T, B, S4, und S5), *Log. Anal.* **167/168**, 243–282 (2001)
- 14.48 L. Magnani: Logic and abduction: Cognitive externalizations in demonstrative environments, *Theoria* **60**, 275–284 (2007)
- 14.49 R. Kowalski: Logic without model theory. In: *What is a Logical System?*, Studies in Logic and Computation, ed. by D. Gabbay (Oxford Univ. Press, Oxford 1994) pp. 73–106
- 14.50 A. Aliseda: *Abductive reasoning. Logical Investigations into Discovery and Explanation* (Springer, Dordrecht 2006)
- 14.51 V. Fiutek: *Playing with Knowledge and Belief*, Ph.D. (Universiteit van Amsterdam, Amsterdam 2013)
- 14.52 F.R. Velázquez-Quesada, F. Soler-Toscano, Á. Nepomuceno-Fernández: An epistemic and dynamic approach to abductive reasoning: Abductive problem and abductive solution, *J. Appl. Log.* **11**(4), 505–522 (2013)
- 14.53 S. Magnier: *Approche dialogique de la dynamique épistémique et de la condition juridique* (College Publications, London 2013), in French
- 14.54 D. Gabbay, J. Woods: *The Reach of Abduction. Insight and Trial* (Elsevier, Amsterdam 2005)
- 14.55 C.S. Peirce: *Collected Papers of Charles Sanders Peirce*, ed. by P. Hartshorne, P. Weiss, A. Burks (Harvard Univ. Press, Cambridge 1931–1958)
- 14.56 L. Keiff: Le Pluralisme Dialogique. Approches dynamiques de l'argumentation formelle, Ph.D. Thesis (Université Lille 3, Lille 2007), in French
- 14.57 S. Rahman, T. Tulenheimo: From games to dialogues and back. In: *Games: Unifying Logic, Language and Philosophy*, Vol. 15, ed. by O. Majer, A.-

- V. Pietarinen, T. Tulenheimo (Springer, Dordrecht 2009) pp. 153–208
- 14.58 S. Rahman, H. Rückert: Dialogical conxive logic, *Synthese* **127**, 105–139 (2001)
- 14.59 J. Meheus, D. Batens: A formal logic for abductive reasoning, *Log. J. IGPL* **14**, 221–236 (2006)
- 14.60 M. Beirlaen, A. Aliseda: A conditional logic for abduction, *Synthese* **191**(15), 3733–3758 (2014)
- 14.61 G. Nzokou: *Logique de l'argumentation dans les traditions orales africaines. Proverbes, Connaissance et Inférences non-monotoniques* (College Publications, London 2013), in French
- 14.62 S. Rahman, J.-P. Van Bendegem: The dialogical dynamics of adaptive paraconsistency. In: *Paraconsistency: The Logical Way to the Inconsistent*, Lecture Notes in Pure and Applied Mathematics, Vol. 228, ed. by W. Carnielli, M.E. Coniglio, I.M. Lofredo D'Ottaviano (Dekker, New York 2001) pp. 295–322
- 14.63 H. Rückert: A solution to Fitch's paradox of knowability. In: *Logic, Epistemology and the Unity of Science*, Logic, Epistemology and the Unity of Science, ed. by S. Rahman, J. Symons, D. Gabbay, J.P. Van Bendegem (Springer, Dordrecht 2004)
- 14.64 J.R. Searle: *Expression and Meaning. Studies in the Theory of Speech Acts* (Cambridge Univ. Press, Cambridge 1979)
- 14.65 J.R. Searle: The logical status of fictional discourse, *New Lit. Hist.* **6**(2), 319–332 (1975)
- 14.66 J. Woods: Fictions and their logics. In: *Handbook of Philosophy of Science*, Vol. 5, ed. by D. Gabbay, P. Thagard, J. Woods (Elsevier, Amsterdam 2007) pp. 1061–1126
- 14.67 J. Woods, J. Isenberg: Psychologizing the semantics of fiction, *Methodos* (2010), doi:[10.4000/methodos.2387](https://doi.org/10.4000/methodos.2387)
- 14.68 M. Fontaine: *Argumentation et engagement ontologique. Être, c'est être choisi* (College Publications, London 2013), in French
- 14.69 K. Bach: Speech acts. In: *Routledge Encyclopedia of Philosophy*, (Routledge, London 1998)
- 14.70 C. Barés Gómez: *Lógica dinámica epistémica para la evidencialidad negativa*, Vol. 5 (College Publications, London 2013), in Spanish
- 14.71 S. Rahman, N. Clerbout: Constructive type theory and the dialogical approach to meaning, the Baltic International Yearbook of Cognition, Log. Commun. **8**, 1–72 (2013)

# 15. Formal (In)consistency, Abduction and Modalities

Juliana Bueno-Soler, Walter Carnielli, Marcelo E. Coniglio, Abilio Rodrigues Filho

This chapter proposes a study of philosophical and technical aspects of logics of formal inconsistency (LFIs), a family of paraconsistent logics that have resources to express the notion of consistency inside the object language. This proposal starts by presenting an epistemic approach to paraconsistency according to which the acceptance of a pair of contradictory propositions  $A$  and  $\neg A$  does not imply accepting both as true. It is also shown how LFIs may be connected to the problem of abduction by means of tableaux that indicate possible solutions for abductive problems. The connection between the notions of modalities and consistency is also worked out, and some LFIs based on positive modal logics (called *anodic modal logics*), are surveyed, as well as their extensions supplied with different degrees of negations (called *catholic modal logics*). Finally, *swap structures* are explained as new and interesting semantics for the LFIs, and shown to be as a particular important case of the well-known *possible-translations semantics* (PTS).

15.1	<b>Paraconsistency</b> .....	315
15.2	<b>Logics of Formal Inconsistency</b> .....	316
15.2.1	mbC: A Minimal LFI .....	318
15.2.2	A Logic of Evidence and Truth .....	321
15.3	<b>Abduction</b> .....	322
15.3.1	mbC-Tableaux .....	324
15.3.2	Quantification .....	326
15.4	<b>Modality</b> .....	327
15.4.1	The Anodic System $K^\diamond$ .....	328
15.4.2	The Logic $mbC^\square$ .....	329
15.4.3	Extensions of $mbC^\square$ .....	330
15.5	<b>On Alternative Semantics for mbC</b> .....	331
15.6	<b>Conclusions</b> .....	333
	<b>References</b> .....	334

## 15.1 Paraconsistency

Paraconsistency is the study of logical systems in which the presence of a contradiction does not imply triviality, that is, logical systems with a nonexplosive negation  $\neg$  such that a pair of propositions  $A$  and  $\neg A$  does not (always) trivialize the system. In paraconsistent logics the *principle of explosion* does not hold

$$A, \neg A \not\vdash B. \quad (15.1)$$

But what would be the reason for devising a paraconsistent logic? Or more precisely, if avoiding contradictions is a fundamental criterion of thought and reason, what is the point of a formal system that tolerates contradictions? It will be argued here that to have available a logical formalism capable of dealing with contradictions does not imply any sympathy with the thesis that

there are true contradictions, nor that reality is, in some sense, contradictory.

It is a fact that contradictions appear in a number of real-life contexts of reasoning. From databases to scientific theories, we often have to deal with contradictory information. There are several scientific theories, however successful in their areas of knowledge, that yield contradictions, either by themselves or when combined with other efficacious and compelling theories [15.1, Chap. 5]. The presence of contradictions is not a sufficient condition for discarding interesting theories. In order to deal rationally with contradictions, explosion cannot be valid without restrictions, since triviality, that is, a circumstance such that everything holds, is obviously unacceptable. Given that, in classical logic, explosion is a valid principle of inference, the underly-

ing logic of a contradictory context of reasoning cannot be classical.

Indeed, the occurrence of contradictions in both scientific theories and everyday contexts of reasoning is being increasingly recognized. Notice that, as a general rule, these theories have been successful in describing and predicting a wide range of phenomena. The realist (and naive) assumption that scientific theories provide correct descriptions of reality would unavoidably imply that there are ontological contradictions, but this would be a careless and hasty conclusion, since these contradictions are better taken as provisional [15.2, p. 2]. If contradictions are provisional, they should not be taken as *true* contradictions. From a strictly logical point of view, the problem is how to formulate an account of logical consequence capable of identifying, in contradictory contexts, the inferences that are allowed, distinguishing them from those that must be blocked. It is clear that such an account of logical consequence must be paraconsistent.

## 15.2 Logics of Formal Inconsistency

There are two different but classically equivalent notions of consistency with respect to a deductive system  $S$  with a negation  $\sim$ .  $S$  is consistent if and only if:

1. There is a formula  $B$  such that  $\not\vdash_S B$ ;
2. There is no formula  $A$  such that  $\vdash_S A$  and  $\vdash_S \sim A$ .

What (1) says is that  $S$  is not trivial; and (2) says that  $S$  is noncontradictory. In classical logic both are (provably) equivalent.

A theory is a set of propositions (or sentences, if one prefers) closed under logical consequence. Given a set of propositions  $\Gamma$  in the language of a given logic  $L$ , let  $T = \{A: \Gamma \vdash_L A\}$  be the theory whose nonlogical axioms are the propositions of  $\Gamma$  and the underlying logic is  $L$ . Suppose the language of  $T$  has a negation  $\sim$ . We say that  $T$  is:

- *Contradictory*: if and only if there is a proposition  $A$  in the language of  $T$  such that  $T$  proves  $A$  and  $T$  proves  $\sim A$ ;
- *Trivial*: if and only if for any proposition  $A$  in the language of  $T$ ,  $T$  proves  $A$ ;
- *Explosive*: if and only if  $T$  trivializes when exposed to a pair of contradictory formulas – i. e., for all  $A$  and  $B$ ,  $T \cup \{A, \sim A\} \vdash B$ .

A theory whose underlying logic is classical is contradictory if and only if it is trivial. But it is the case precisely because such a theory is explosive, since the principle of explosion holds in classical logic. It is clear,

The question about the nature of contradictions accepted by paraconsistent logics is where a great deal of the debate on the philosophical significance of paraconsistency has been concentrated. In philosophical terminology, we say that something is ontological when it has to do with reality, with the world in the widest sense, and that something is epistemological when it has to do with knowledge and the process of its acquisition. A central question for paraconsistency is the following: Are the contradictions that paraconsistent logic deals with ontological or epistemological? Do contradictions have to do with reality proper? Or do contradictions have to do with knowledge and thought? Contradictions of the latter kind, called here epistemological contradictions, have their origin in our cognitive apparatus, in the failure of measuring instruments, in the interactions of these instruments with phenomena, in operations of thought, or even in simple mistakes that in principle could be corrected later on.

then, that *it is contradictoriness together with explosiveness that implies triviality*. The obvious move in order to deal with contradictions is thus to reject the unrestricted validity of the principle of explosion by means of adopting a *nonexplosive negation* (that is, a negation  $\neg$  such that  $A, \neg A \not\vdash B$ ). This is a necessary condition if one wants a contradictory but nontrivial theory.

For classical negation  $\sim$  the following conditions hold

$$A \wedge \sim A \vDash , \quad (15.2)$$

$$\vDash A \vee \sim A . \quad (15.3)$$

According to (15.2), there is no model  $M$  such that  $A \wedge \sim A$  holds in  $M$ . Equation (15.3) says that for every model  $M$ ,  $A \vee \sim A$  holds in  $M$ . A negation is para-complete if it disobeys (15.3), and is paraconsistent if it disobeys (15.2). Now, given classical consequence,  $A \vee \sim A$  follows from anything, and anything follows from  $A \wedge \sim A$ . From the point of view of rules of inference, the duality is not between noncontradiction and excluded middle, but rather between explosion and excluded middle.

The principle of noncontradiction is usually taken as a claim that there can be no contradictions in reality. But we may well understand the principle of explosion as a stronger way of saying precisely the same thing:  $A$  and  $\sim A$  cannot hold together, otherwise we get triviality. From the above considerations, it is clear that in



order to give a counterexample to the principle of explosion, we need a weaker negation and a semantics in which there is a model  $M$  such that  $A$  and  $\neg A$  holds in  $M$ .

In classical logic the values 0 and 1 are understood respectively as false and true, but in nonclassical logics this does not need to be the case. It is not necessary that a paraconsistent logic takes a pair of formulas  $A$  and  $\neg A$  as both true. The semantic value 1 attributed to a formula  $A$  may be read as *A is taken to be true*, *A is possibly true*, *A is probably true*, or perhaps better as *there is some evidence that A is true* in the sense of there being reasons for believing that  $A$  is true. Thus, the attribution of the value 1 to a pair of propositions  $A$  and  $\neg A$ , does not need to be understood as if both propositions are true in the sense that there is something in the world that makes them true. Rather, it is better to consider that  $A$  and  $\neg A$  are both being taken in a sense weaker than true, perhaps waiting for further investigations that will decide the issue, and discard one of them.

*Logics of formal inconsistency* are a family of paraconsistent logics that have resources to express the notion of consistency inside the object language by means of a sentential unary connective:  $\circ A$  means (informally) that *A is consistent*. As in any other paraconsistent logic, explosion does not hold in LFIs. But this is handled in a way that allows distinguishing between contradictions that can be accepted from those that cannot. In LFIs, negation is explosive only with respect to consistent formulas

$$A, \neg A \not\vdash_{\text{LFI}} B, \text{ while } \circ A, A, \neg A \vdash_{\text{LFI}} B. \quad (15.4)$$

An LFI is thus a logic that separates the propositions for which explosion holds from those for which it does not hold. The former are marked with  $\circ$ . For this reason, they are called *gently explosive*.

In the  $C_n$  hierarchy, introduced by *da Costa* in [15.3], the so-called *well-behavedness* of a formula  $A$ , in the sense that it is not the case that  $A$  and  $\neg A$  hold, is also expressed inside the object language. However, in  $C_1$ ,  $A^\circ$  is an abbreviation of  $\neg(A \wedge \neg A)$ , which makes the *well-behavedness* of a proposition  $A$  equivalent to saying that  $A$  is noncontradictory. A full hierarchy of calculi  $C_n$ , for  $n$  natural, is defined and studied in [15.3].

The first step in paraconsistency is the distinction between triviality and contradictoriness. But there is a second step, namely, the distinction between consistency and noncontradictoriness. In LFIs the consistency connective  $\circ$  is not only primitive, but it is also not always logically equivalent to noncontradiction. This is the most distinguishing feature of the logics of formal inconsistency. Once we break up the equivalence

between  $\circ A$  and  $\neg(A \wedge \neg A)$ , some very interesting developments become available. Indeed,  $\circ A$  may express notions different from consistency as freedom from contradiction.

The circumstance in which both  $A$  and  $\neg A$  receive the value 1 may be understood as the presence of simultaneous but nonconclusive evidence that  $A$  is true and  $\neg A$  is true. *Evidence for A* in the sense proposed here are *reasons to believe in A*. One may be justified in believing that  $A$  is true inasmuch one has evidence available that  $A$  is true. But of course it may be that there are also reasons for believing  $\neg A$ , and in this case the evidence is not conclusive.

Suppose that according to some empirically testable criteria, an atomic proposition  $A$  is true if and only if a condition  $c$  is fulfilled and on the other hand, there is also a condition  $d$ , independent of  $c$ , such that obtaining  $d$  implies the truth of  $\neg A$ . In some critical circumstances, it may happen that both criteria  $c$  and  $d$  are obtained [15.4, pp. 9–10]. Although  $c$  and  $d$  have been conceived initially as criteria of truth, it seems far more reasonable at this point to not draw the conclusion that  $A$  and  $\neg A$  are both true. It is better to be more careful and to take the contradiction as a provisional state, a kind of excess of information that should, at least in principle, be eliminated by means of further investigation. The criteria  $c$  and  $d$  provide reasons for believing (i. e., provide evidence) that  $A$  and  $\neg A$  are true, but do not establish conclusively that both are true. Thus, a counterexample for explosion is straightforward: there may be nonconclusive evidence for both  $A$  and  $\neg A$ , but no evidence for some  $B$ .

This intuitive interpretation for the paraconsistent negation justifies the invalidity of explosion. However, it is not possible yet to express that some proposition is true, because the notion of evidence is weaker than truth. With the help of the consistency operator this problem can be solved. The following intuitive meaning for the consistency operator is proposed:  $\circ A$  means informally that the truth value of  $A$  has been conclusively established. Now one has resources to express not only that there is evidence that  $A$  is true but also that  $A$  has been established (by whatever means) as true:  $\circ A \wedge A$ . Notice that how the truth or falsity of a proposition is established is not a concern of logic. The establishment of the truth of a given proposition  $A$  comes from outside the formal system.

A very good example of a provisional contradiction in physics, better understood in terms of conflicting evidence rather than truth, is the problem faced by Einstein just before he formulated the special theory of relativity. It is well known that there was an incompatibility between classical Newtonian mechanics and Maxwell's theory of electromagnetic field. This is a typical case of

two (supposedly) noncontradictory theories that yield contradictory results.

A friendly presentation of the problem may be found in *Einstein* [15.5] and *Feynman et al.* [15.6]. Briefly, with respect to the same hypothetical situation, with  $c$  being the velocity of light in vacuum and  $w$  the velocity of light in a particular circumstance [15.5, Sects. 6 and 7], Newtonian mechanics and Maxwell's theory provide that  $\neg(w = c)$  and  $w = c$  respectively. So, combining the two theories yields a contradiction, and if the underlying logic is classical, triviality follows.

In such a scenario, two contradictory propositions hold in the sense that both may be *proven* from theories that were supposed to be correct. This fact may be represented by the attribution of the semantic value 1 to both  $\neg(w = c)$  and  $w = c$ . But clearly, the meaning of this should not be that both are true – actually, we know it is not the case, and nobody has ever supposed that it could be the case. The meaning of the simultaneous attribution of the value 1, as we suggest, is that at that time there was evidence for both in the sense, mentioned above, of some *reasons for believing* that both were true, because there was evidence that the results yielded by both classical mechanics and theory of the electromagnetic field were true. The *contradiction* has been solved (roughly speaking) in the following way: As velocity grows, time *slows down* and *space shortens*. So, the relation between space and time that gives velocity remains the same, because both have decreased (for details, see [15.6, Sect. 15]). This is an example of what we call *epistemic contradictions*. We want to call attention to the fact that the general logical framework Einstein was working in was not classical. He had two different theories at hand: classical mechanics and the theory of electromagnetic field, which, when combined, yielded a nonexplosive contradiction. Later, according to the special theory of relativity, the *contradiction* disappeared. Although there were some reasons to believe that both  $\neg(w = c)$  and  $w = c$  were true, only one, the latter, has been established as true. The value 1 attributed to  $\neg(w = c)$  later became 0.

### 15.2.1 mbC: A Minimal LFI

Let us present *mbC*, a basic LFI. This name stands for a *minimal logic with the axiom bc1*, and *bc* stands for *basic property of consistency*. *mbC* is an extension of classical positive propositional logic (CPL+) enriched with a nonexplosive negation and a *consistency* operator, the unary operator  $\circ$ . *mbC* is interesting because it has a minimal apparatus and several technical properties that illustrate the main features of logics of formal inconsistency. As we shall see, *mbC*:

- i. Permits us to define classical negation, and thus can be seen as an extension of classical logic
- ii. Permits recovering classical consequence by means of a derivability adjustment theorem (DAT)
- iii. Distinguishes the consistency of a formula  $A$  from the noncontradiction of  $A$ , i. e.,  $\circ A$  and  $\neg(A \wedge \neg A)$  are not equivalent
- iv. Is gently explosive in the sense that it tolerates some pairs of formulas  $A$  and  $\neg A$ , while it is explosive with respect to others; and
- v. Has a sound and complete bivalued semantics.

#### The Syntax of mbC

Let  $L_1$  be a language with a denumerable set of sentential letters  $\{p_1, p_2, p_3, \dots\}$ , the set of connectives  $\{\circ, \neg, \wedge, \vee, \rightarrow\}$ , and parentheses. The consistency operator  $\circ$  is a primitive symbol and  $\neg$  is a nonexplosive negation. The set of formulas of  $L_1$  is obtained recursively in the usual way; and Roman capitals stand for metavariables for formulas of  $L_1$ .

The logic *mbC* is defined over the language  $L_1$  by the following Hilbert system:

Axiom-schemas:

- Ax.1.  $A \rightarrow (B \rightarrow A)$ ,
- Ax.2.  $(A \rightarrow (B \rightarrow C)) \rightarrow ((A \rightarrow B) \rightarrow (A \rightarrow C))$ ,
- Ax.3.  $A \rightarrow (B \rightarrow (A \wedge B))$ ,
- Ax.4.  $(A \wedge B) \rightarrow A$ ,
- Ax.5.  $(A \wedge B) \rightarrow B$ ,
- Ax.6.  $A \rightarrow (A \vee B)$ ,
- Ax.7.  $B \rightarrow (A \vee B)$ ,
- Ax.8.  $(A \rightarrow C) \rightarrow ((B \rightarrow C) \rightarrow ((A \vee B) \rightarrow C))$ ,
- Ax.9.  $A \vee (A \rightarrow B)$ ,
- Ax.10.  $A \vee \neg A$ ,
- Ax.bc1.  $\circ A \rightarrow (A \rightarrow (\neg A \rightarrow B))$ ,

Inference rule: modus ponens.

Positive classical propositional logic, CPL+, is given by axioms Ax. 1 to Ax. 9 plus modus ponens. *mbC* is thus an extension of CPL+.

The definition of a derivation of  $A$  from a set of premises  $\Gamma$  ( $\Gamma \vdash_{\text{mbC}} A$ ) is the usual one: a finite sequence of formulas  $B_1 \dots B_n$  such that  $A$  is  $B_n$  and each  $B_i$ ,  $1 \leq i \leq n$  (that is, each line of the proof) is an axiom, a formula that belongs to  $\Gamma$ , or a result of modus ponens. A theorem is a formula derived from the empty set of premises.

#### Lemma 15.1

The logic *mbC* satisfies the following properties:

- P1. Reflexivity: if  $A \in \Gamma$ , then  $\Gamma \vdash_{\text{mbC}} A$ ;

- P2. Monotonicity: if  $\Gamma \vdash_{\text{mbC}} B$ , then  $\Gamma, A \vdash_{\text{mbC}} B$ , for any  $A$ ;
- P3. Cut: if  $\Delta \vdash_{\text{mbC}} A$  and  $\Gamma, A \vdash_{\text{mbC}} B$ , then  $\Delta, \Gamma \vdash_{\text{mbC}} B$ ;
- P4. Deduction theorem: if  $\Gamma, A \vdash_{\text{mbC}} B$ , then  $\Gamma \vdash_{\text{mbC}} A \rightarrow B$ ;
- P5. Compactness: if  $\Gamma \vdash_{\text{mbC}} A$ , then there is  $\Delta \subseteq \Gamma$ ,  $\Delta$  finite,  $\Delta \vdash_{\text{mbC}} A$ .

*Proof:* The properties P1, P2, P3 and P5 come directly from the definition of  $\Gamma \vdash_{\text{mbC}} A$ . The deduction theorem comes from axioms Ax. 1 and Ax. 2 plus modus ponens. ■

Since the properties P1, P2 and P3 hold, *mbC* is thus a standard logic [15.7, p. 6]. Due to the axiom *bc1*, *mbC* is gently explosive, that is

For some  $A$  and  $B$ :

$A, \neg A \not\vdash_{\text{mbC}} B$ ,

$\circ A, A \not\vdash_{\text{mbC}} B$ ,

$\circ A, \neg A \not\vdash_{\text{mbC}} B$ ,

While for every  $A$  and  $B$ :  $\circ A, A, \neg A \vdash_{\text{mbC}} B$ .

Thus, the formal system is able to distinguish the contradictions that do not lead to explosion from those that do. The axiom *bc1* is also called the *gentle explosion law*, because it is explosive only with respect to formulas marked with  $\circ$ .

Classical logic may be recovered in *mbC* in two ways: by defining a negation that has the properties of the classical negation and by means of a derivability adjustment theorem (DAT).

#### Fact 15.1

Classical negation is definable in *mbC*.

*Proof:* We define  $\perp \stackrel{\text{def}}{=} \circ A \wedge A \wedge \neg A$  and  $\sim A \stackrel{\text{def}}{=} A \rightarrow \perp$ . Now, we get explosion,  $A \rightarrow (\sim A \rightarrow B)$ , as a theorem, in a few steps from *bc1*. From the axiom 9, excluded middle is obtained,  $A \vee \sim A$ . Classical propositional logic (CPL) is obtained by axioms 1–8 plus explosion, excluded middle and modus ponens. ■

The general purpose of a derivability adjustment theorem is to establish a relationship between two logics, *L1* and *L2*, in the sense of restoring inferences that are lacking in one of them. The basic idea is that some *information* has to be added to the premises to restore the inferences that are lacking. DATs are especially interesting because they show what is needed to restore the classical consequence in a paraconsistent scenario.

For the sake of precisely stating the DAT between *mbC* and CPL, we need to take into account the differ-

ence between the respective languages. The first step is to translate one language into another. Let  $L_2$  be a language with the set of connectives  $\{\sim, \vee, \wedge, \rightarrow\}$ . Instead of a paraconsistent negation  $\neg$ ,  $L_2$  has classical negation  $\sim$ .

#### Fact 15.2

Let  $t$  be a mapping that replaces  $\sim$  by  $\neg$ . Then, the following holds:

For all  $\Gamma$  and for all  $B$ ,  $\Gamma \cup \{B\} \subseteq L_2$ , there is a  $\Delta$ ,  $\Delta \subseteq L_1$  such that  $\Gamma \vdash_{\text{CPL}} B$  iff  $t[\Gamma], \circ \Delta \vdash_{\text{mbC}} t[B]$ , where  $\circ \Delta = \{\circ A : A \in \Delta\}$ .

*Proof:* From left to right, suppose there is a derivation  $D$  of  $\Gamma \vdash_{\text{CPL}} B$  (in the language  $L_2$  of CPL). If we simply change the classical negations  $\sim$  to  $\neg$ , such a derivation does not hold in *mbC*. We need to be concerned only with occurrences of explosion. The relevant point is that some information must be available in order to reconstruct classical reasoning. An occurrence of a line:

1.  $A \rightarrow (\sim A \rightarrow B)$

in the derivation  $D$  has to be substituted by the following lines, obtaining a derivation  $D'$ :

2.  $\circ A$

3.  $\circ A \rightarrow (A \rightarrow (\neg A \rightarrow B))$

4.  $A \rightarrow (\neg A \rightarrow B)$ .

From right to left, suppose there is a derivation  $D'$  of  $t[\Gamma], \circ \Delta \vdash_{\text{mbC}} t[B]$ . We get a derivation  $D$  of  $\Gamma \vdash_{\text{CPL}} B$  just by deleting the occurrences of  $\circ$  and changing  $\neg$  to  $\sim$ . ■

The reader should notice the difference between restoring classical consequence by means of a definition of a classical negation inside *mbC* and by means of a DAT. In the latter case, the central issue is the information that has to be available to restore classical reasoning. In each occurrence of classical explosion,  $A \rightarrow (\sim A \rightarrow B)$ , the information needed from the viewpoint of *mbC* is the consistency of  $A$ , represented by  $\circ A$ .

#### A Semantics for *mbC*

The sentential connectives of classical logic are truth functional. This means that the truth-value of a molecular formula is functionally determined by its structure and by the truth-values of its components, which reduce to the truth-values of the atomic formulas. Truth functionality as a property of the semantics of certain logics is a mathematical rendering of the *principle of compositionality*, which says that the meaning of a complex expression is functionally determined by the meanings of its constituent expressions and the rules used

to combine them. This principle is also called Frege's principle, since it has been traced back to Frege. The truth-value of molecular formulas may be determined by using the familiar matrices (truth-tables) that any logic student is familiar with. These matrices have only two values (*true* and *false*, or 1 and 0) in the case of classical propositional logic, but the idea can be generalized to any number of *truth-values*.

A logic can be *truth functional* even if it is characterized semantically by a finite, or even by an infinite, number of *truth-values*. Indeed, in most many-valued logics the *truth-value* of a molecular formula is also functionally determined by the values of the atomic formulas.

However, instead of talking about truth-values, it would be better to talk about *semantic values*, since, as we have argued earlier, the values 0 and 1 attributed to formulas need not be *always* interpreted as false and true. The point is that we do not want to commit ourselves from the start to just the values true and false attributed to formulas.

Now let us put things in a more neutral and precise way. Let us say that a semantics for a logic  $L$  is called *matrix functional* (instead of truth functional) if the semantic value of a formula of  $L$  is functionally determined by means of a finite matrix. This is the case for classical logic, but not for intuitionistic logic, as Gödel has proven in [15.8], nor is the case for *mbC*. In fact, not only *mbC* but all logics of the da Costa hierarchy  $C_n$ , and most LFIs, are not characterizable by finite matrices [15.7, theorem 121].

A non-matrix-functional semantics for paraconsistent logics was proposed by *da Costa* and *Alves* in [15.9]. There, a bivalued semantics for the logic  $C_1$  is found. The semantic clause for the paraconsistent negation has only *half* of the clause for classical negation: if  $v(\neg A) = 0$ , then  $v(A) = 1$ . The idea is that it cannot be the case that  $A$  and  $\neg A$  simultaneously receive the value 0. But the possibility is open for both to receive the value 1. This kind of semantics is described by the so-called *quasi-matrices*. The quasi-matrix for negation is as follows

$A$	$\neg A$
1	0
	1
0	1

It is clear that the semantic value of  $\neg A$  is not functionally determined by the semantic value of  $A$ : when  $v(A) = 1$ ,  $v(\neg A)$  may be 1 or 0. For this reason, this semantics is clearly nonfunctional, and in this case we say also that the semantics is *nondeterministic*. A bivalued sound and complete semantics for *mbC*, is as follows:

### Definition 15.1

An *mbC-valuation* is a function that assigns the values 0 and 1 to formulas of the language  $L_1$ , satisfying the following clauses:

- (i)  $v(A \wedge B) = 1$  iff  $v(A) = 1$  and  $v(B) = 1$
- (ii)  $v(A \vee B) = 1$  iff  $v(A) = 1$  or  $v(B) = 1$
- (iii)  $v(A \rightarrow B) = 1$  iff  $v(A) = 0$  or  $v(B) = 1$
- (iv)  $v(\neg A) = 0$  implies  $v(A) = 1$
- (v)  $v(\circ A) = 1$  implies  $v(A) = 0$  or  $v(\neg A) = 0$ .

A valuation  $v$  is a model of a set  $\Gamma$  if and only if every proposition of  $\Gamma$  receives the value 1 in  $v$ , and  $v$  is a model of a proposition  $A$  if and only if  $A$  receives the value 1 in  $v$ .

The notion of logical consequence is defined as usual: A formula  $A$  is an *mbC-consequence* of a set  $\Gamma$  ( $\Gamma \models_{\text{mbC}} A$ ), if and only if for every valuation  $v$ , if  $v$  is a model of  $\Gamma$ , then  $v$  is a model of  $A$  (when there is no risk of ambiguity, we simply write  $\models$  and  $\vdash$  without subscripts).

Soundness and completeness proofs for the semantics above are to be found in [15.7, pp. 38ff]. Notice that the clauses for  $\wedge$ ,  $\vee$  and  $\rightarrow$  are exactly the same as in classical logic. By clause (iv), the system is paraconsistent but not paracomplete, since the excluded middle (for paraconsistent negation) holds. We suggest that the values 0 and 1 are not to be understood as false and true respectively, but rather as absence and presence of evidence. Accordingly:  $v(A) = 1$  means *there is evidence that  $A$  is true*;  $v(A) = 0$  means *there is no evidence that  $A$  is true*;  $v(\neg A) = 1$  means *there is evidence that  $A$  is false*; and  $v(\neg A) = 0$  means *there is no evidence that  $A$  is false*. The same counter example invalidates explosion and disjunctive syllogism:  $v(A) = 1$ ,  $v(B) = 0$  and  $v(\neg A) = 1$ . Noncontradiction is also invalid:  $v(A) = v(\neg A) = 1$ , hence  $v(A \wedge \neg A) = 1$ , but  $v(\neg(A \wedge \neg A))$  may be 0. Due to clause (v), it may be the case that  $v(A) = 1$ ,  $v(\neg A) = 0$  (or vice versa) but  $v(\circ A) = 0$ . In this valuation,  $v(\neg(A \wedge \neg A)) = 1$ , hence the nonequivalence between  $\circ A$  and  $\neg(A \wedge \neg A)$ . Also, due to clause (v), it is clear that *mbC* does not admit a trivial model, i. e., a model such that  $v(A) = 1$  for every formula  $A$ .

We would like to make some comments with respect to the validity of the excluded middle in *mbC*, given the intended interpretation in terms of evidence. Indeed, in some circumstances there may be a situation such that there is no evidence at all, neither for the truth nor for the falsity of a proposition  $A$ , but this scenario cannot be represented in *mbC*. In fact, in the next section, it will be shown that *mbC* may be easily modified in order to be able to represent such a situation. On the other hand, the validity of the excluded middle may be justified when one, by default, attributes evi-

dence for  $A$  or for  $\neg A$  when there is no evidence at all. This happens, for instance, in a criminal investigation in which one begins by considering everyone (in some group of people) not guilty until proof to the contrary (see Example 15.5 in Sect. 15.3.1). In fact, any context of reasoning such that a final decision must be made in a finite amount of time demands that  $A$  or  $\neg A$ , or even both, have to be in some sense accepted. Whether the excluded middle should be valid from the start, or be recovered once some information has been added may be seen as a methodological decision that depends on the reasoning scenario we want to represent.

### 15.2.2 A Logic of Evidence and Truth

The logic  $mbC$  may be slightly modified to be able to express a scenario such that no evidence is available. The duality between the principles of explosion and excluded middle that corresponds to a duality between paraconsistency and paracompleteness has been mentioned in Sect. 15.2. Now, in a way analogous to that by which we recover the explosion with respect to a formula  $A$ , in a paracomplete logic, the validity of the excluded middle with respect to a formula  $A$  may be recovered by means of the following axiom

$$\text{Ax. bd1. } \circ A \rightarrow (A \vee \neg A) .$$

A semantic clause for the axioms  $bc1$  and  $bd1$  is defined as follows:

(vi) if  $v(\circ A) = 1$ , then  $[v(A) = 1 \text{ iff } v(\neg A) = 0]$ .

If the excluded middle holds for  $A$ , we say that  $A$  is determined.  $bd$  stands for *basic property of determinedness* (bd). A system in which both  $bd1$  and  $bc1$  holds is thus paracomplete and paraconsistent. It is better to call  $\circ$ , in this context, not a consistent operator but rather a *classicality operator*, since  $\circ A$  recovers classical truth conditions with respect to  $A$ . But  $\circ A$  still may be informally understood as meaning that the truth value of  $A$  has been conclusively established. In fact, the basic idea of restricting the validity of the principle of explosion may be generalized. The validity of some inference rule (or axiom) may be restricted in such a way that some *logical property* does not hold unless some information is added to the system. In particular, the excluded middle may be restricted in a way analogous to the restriction imposed to the explosion.

Now, with  $bd1$  and  $bc1$ , we have the resources to express the following situations, besides nonconclusive evidence; no evidence at all:  $v(A) = v(\neg A) = 0$  and conflicting evidence:  $v(A) = v(\neg A) = 1$ . The system obtained by adding the axioms  $bd1$  and  $bc1$  to CPL+ is called  $mbCD$ , a minimal logic of inconsistency and

undeterminedness.  $mbCD$  is correct and complete with respect to a bivalued semantics defined by clauses (i) to (iii) of Def. 15.1 plus the clause (vi) above.

Although  $mbCD$  is able to express also the absence of evidence, the negation still may be improved to better represent the deductive behavior of the notion of *preservation of evidence*. The logic  $LET_K$  (the logic of evidence and truth based on CPL+) is thus obtained by adding to  $mbCD$  the axioms 11 to 14 below

$$\text{Ax. 11. } A \leftrightarrow \neg \neg A ,$$

$$\text{Ax. 12. } \neg(A \wedge B) \leftrightarrow (\neg A \vee \neg B) ,$$

$$\text{Ax. 13. } \neg(A \vee B) \leftrightarrow (\neg A \wedge \neg B) ,$$

$$\text{Ax. 14. } \neg(A \rightarrow B) \leftrightarrow (A \wedge \neg B) .$$

The axioms above fit the intuitive meaning of the simultaneous attribution of the value 0 or the value 1 to a pair of propositions  $A$  and  $\neg A$  as absence and presence of evidence respectively. Let us consider axiom 12. It is reasonable to conclude that if there is some evidence that a conjunction is false, that same evidence must be evidence that one of the conjuncts is false. On the other hand, if there is some evidence that  $A$  is false, that same evidence must be evidence that  $A \wedge B$  is false, for any  $B$ . Analogous reasoning applies for disjunction and implication.

A bivalued complete and correct semantics for  $LET_K$  is obtained by adding to the semantics of  $mbCD$  the following clauses

(vii)  $v(A) = 1$  iff  $v(\neg \neg A) = 1$ ,

(viii)  $v(\neg(A \wedge B)) = 1$  iff  $v(\neg A) = 1$  or  $v(\neg B) = 1$ ,

(ix)  $v(\neg(A \vee B)) = 1$  iff  $v(\neg A) = 1$  and  $v(\neg B) = 1$ ,

(x)  $v(\neg(A \rightarrow B)) = 1$  iff  $v(A) = 1$  and  $v(\neg B) = 1$ .

The logic  $LET_K$  can be proven without much trouble to be sound and complete with respect to the semantics above. The proof needs only to drop the clauses related to Ax. 10 and to extend the soundness and completeness proofs for  $mbC$  [15.7, pp. 38ff] to the new axioms and semantic clauses, which can be done without difficulties. In  $LET_K$ , a DAT holds as in  $mbC$  and a classical negation is definable in the same way as in  $mbC$ , thus  $LET_K$  may be also seen as an extension of propositional classical logic. It is worth noting that according to the intuitive interpretation proposed,  $LET_K$  like  $mbC$  does not tolerate true contradictions: indeed, a true contradiction yields triviality, as in classical logic. If  $A$  is simultaneously true and false, this is expressed by  $(\circ A \wedge A) \wedge (\circ A \wedge \neg A)$ , that, in its turn, is equivalent to  $\circ A \wedge A \wedge \neg A$ , but the latter is nothing but a *bottom particle*  $\perp$ , i. e., a formula that alone implies triviality: for any  $B$ ,  $\perp \vdash B$ .

## 15.3 Abduction

The problem of abduction has been formulated by *Peirce* as the process of forming hypotheses with explanatory purposes. It is, thus, a kind of a reversed explanation [15.10, CP 7.202]:

“Accepting the conclusion that an explanation is needed when facts contrary to what we should expect emerge, it follows that the explanation must be such a proposition as would lead to the prediction of the observed facts, either as necessary consequences or at least as very probable under the circumstances. A hypothesis then, has to be adopted, which is likely in itself, and renders the facts likely. This step of adopting a hypothesis as being suggested by the facts, is what I call *abduction*.”

The basic idea may be expressed as follows: When some fact is discovered that is not explained by the available theory (i. e., is not a consequence of the available theory), a set of new premises is added as a hypothetical solution to the problem. However, the act of adding something before using it as an explanation poses a second problem: how is it possible to generate an abductive *explanans*? First, we have to recognize that characterizing the concept of explanation is one of the greatest challenges in the philosophy of science. This problem is even harder in logic and mathematics, where explanations are sometimes confused with proofs [15.11].

Although we are not suggesting that *explaining* can be reduced to *deducing*, it is certainly acceptable that the idea of explanation in deductive sciences includes the query for missing hypotheses; it is in this context that the general abductive process can be formulated as the process of generating new hypotheses within arbitrary deductive systems, and afterwards using them in deductive terms. The former task (generating new hypotheses) is referred to here as *creative abduction*, while the latter (using such new hypotheses) as *explicative abduction*. The term *explicative* is here understood under the following proviso: A missing link in a deduction certainly does not exhaust the need for an explanation, but does constitute the first necessary step towards explaining a surprising (i. e., not yet deducible) fact.

Two natural assumptions about explanations that can be posed are the following: First, there can be various explanations for the same surprising fact, and second, there can be explanations of various *degrees* for the same surprising fact. For example, searching for the ultimate scientific explanation as to why the grass of your garden is wet in the morning and discovering that the sprinkler was left on all night may be two different things. Both explain the fact, but responding to different

needs. The question is analogous to the one in automatic theorem proving: finding any proof is one thing, while finding a *philosophically interesting* proof is another. In the same manner as automatic theorem proving is satisfied with the first level of proofs, so automatic abduction will be satisfied with a first-level explanation. We are mainly interested here in abductions with no *obvious* explanations, particularly those in which contradictions may be involved.

Let  $\vdash$  be a deductive relation; if  $\Gamma \not\vdash A$ , the creative abductive step consists of finding an appropriate  $\Delta$  so that  $\Gamma \cup \Delta \vdash A$ . In this case, the discovered  $\Delta$  performs the explicative abductive step. Obviously, there must be some constraints, otherwise  $\Delta = \{\perp\}$  would be a trivial solution for the abduction problem in most deduction relations. Usually, if the underlying logic is explosive (e.g., classical or intuitionistic), another constraint is that  $\Gamma \not\vdash \neg A$ , for this would imply that any explicative  $\Delta$  would be a trivial explanation. This restriction, however, will not be necessary in our case, as the following developments and the examples will make clear.

From the point of view of general argumentation (and not only deduction), abduction concerns the search for hypotheses or the search for explanatory instances that support reasoning. In this sense, it can be seen as a complement to argumentation, in the same manner that in the philosophy of science, the context of discovery is a complement to the context of justification. And moreover, further pursuing the analogy, the question of the logical possibility of creative abduction lies on the same side of the famous question of the logical possibility of scientific discovery.

A renewed interest in abduction acquired impetus due to the factual treatment of data and the question of virtual causality in the information age. The enormous amount of data stored on the World Wide Web and in complex systems, as well as the virtual relationship among such data, continuously demands new tools for automatic reasoning. These tools should incorporate general logical methods which are at the same time machine-understandable, and sufficiently close to human semantics as to perform sensible automated reasoning.

An example wherein abductive inference is highly relevant is the model-based diagnosis in engineering and AI (artificial intelligence). Suppose that a complex system, such as an aircraft, is being tested before a transatlantic flight. The electronic circuitry permits the testers to predict certain outputs based on specific input tests. If the instruments show something distinct from the expected, it is a task of model-based diagnosis to discover an explanation for the anomaly and use it to

separate the components responsible for the problem, instead of disassembling the whole aircraft.

Another example occurs in the process of updating in the so-called datalog databases. Suppose we have a logic program (see introductory chapter for a brief overview of abduction in logic programming) composed of the following clauses, where  $desc(x, y)$  means  $x$  is a descendent of  $y$ ,  $parent(y, x)$  means  $y$  is a parent of  $x$  and  $\beta \leftarrow \alpha$  means that the database contents plus  $\alpha$  produce (or answer)  $\beta$

$$\begin{aligned} desc(x, y) &\leftarrow parent(y, x), \\ desc(x, y) &\leftarrow parent(z, x), desc(z, y). \end{aligned}$$

There is a subtle difference between inserting information in the database in an explicit versus implicit manner: information of the form  $y$  is a parent of  $x$  is a basic fact, and can be inserted explicitly, while information of the form  $x$  is a descendent of  $y$  is either factual knowledge or is a consequence of the machine reasoning (as simple as it can be). If one wishes to insert a piece of implicit information, it is necessary to modify the set of facts stored in the database in such a way that this information can be deduced: this is an example of creative abduction and of explicative abduction at the same time. For instance, if we have stored  $desc(Zeus, Uranus)$  and  $parent(Uranus, Cronus)$ , for implicitly inserting  $desc(Aphrodite, Uranus)$  there are two different ways: we may either insert  $parent(Zeus, Aphrodite)$  or alternatively insert  $desc(Aphrodite, Cronus)$ . These two alternative additions are examples of abductive explanations for  $desc(Aphrodite, Uranus)$ . In fact, logic programming uses this abductive mechanism for answering queries, in the form: *Is the fact  $desc(Aphrodite, Uranus)$  compatible with the program clauses and data?* or *Is there any  $x$  such that  $desc(x, Uranus)$ ?* The whole procedure is creative as much as it can be automatized. Therefore it is evident that a useful abductive mechanism for databases should be based on first-order logic, and not merely on propositional logic.

Moreover, abductive approaches are also used to integrate different ontologies and database schemes, or for integrating distinct data sources under the same ontology, as for example [15.12], where an abductive-based application for database integration is developed. Suppose that, while a query is being processed by a user, another data source had inserted  $desc(Uranus, Aphrodite)$  in our database, plus a constraint of the form: *For no  $x$  and  $y$ , simultaneously  $desc(x, y)$  and  $desc(y, x)$  can be maintained in the database.* If  $parent(Zeus, Aphrodite)$  had been inserted for one data source, the insertion of  $desc(Uranus, Aphrodite)$  by the second data source would cause a collapse in view of the constraint. What

can be done? The alternative of deleting all data seems inconceivable, and the one of having all queries be answered positively (since a database established on classical logic grounds would deduce anything from a contradiction) is of course intolerable. Thus a legitimate logic to ground this process would have to be a first-order logic that sanctions useful reasoning in the presence of contradictions.

Proposals from this perspective have been investigated in [15.13]. A first-order LFI,  $QmbC$ , that is an extension of the logic  $mbC$ , has been presented and investigated in [15.14]. We will argue here that simple yet powerful techniques for automatic abduction can be usefully implemented by means of tableau proof-procedures for the logic  $mbC$ , which may be extended to  $QmbC$ .

Although a wide-scoped study of tableaux and abduction was offered by [15.15] in 1997, the quite natural idea of using the backward mechanism of tableaux for gaining automatic explanations occurred earlier: [15.16] in 1992 already proposed a fully detailed treatment for the question of *completing* a database in a way as to deduce (in classical propositional logic) a previously undeducible formula. In the same year a tableau proof system for da Costa's logic  $C_1$  was proposed in [15.17], and several examples of using such tableau systems for automatic solving dilemmatic situations were extensively discussed. Even though neither of these references explicitly mentions the concept of abduction, these papers undoubtedly proposed ways to solve the abductive problem, for classical propositional logic and for the propositional paraconsistent logic  $C_1$  respectively. In [15.18], tableau systems for LFIs were proposed, and this logical formalism was used as a method to devise database repairs in [15.13]. Such methods are based upon many-valued semantics, or upon bivalued semantics.

The question of abduction thus involves two independent, but complementary problems:

1. Finding a method to automatically perform abduction (and, if possible, to automatically generate abductive data), and
2. Doing this within a robust reasoning environment, in a such a way as to keep running and providing reasonable output even in the presence of the possible contradictions that this search would engender.

Any contradictions found in the process of producing lucid outputs are a condemnation of the whole process if the underlying logic is classical, so, the abductive experience can sometimes appear to be lethal. We argue here, however, that very simple and natural logical models can be designed for dealing with abduc-

tion, by means of defining them in terms of refutation procedures based on LFIs.

It is well accepted that abduction does not go in the forward direction of deduction. It is not difficult to accept, either, that abduction cannot coincide with any backward form of classical deduction, but it does not follow that another form of backward deduction would not work. In this sense, some LFIs are good candidates. Let us take as example *mbC*: it does not prove anything that classical logic would not prove; it tolerates contradictions, but, nonetheless, it can encode the whole of classical reasoning. Backwards proof procedures for LFIs indeed constitute a suitable approach for abduction, and we intend to show how this approach can be programmed and treated on a natural basis departing from a very simple formalism.

We have already seen here that it is typical that cognitive situations can enter into a situation of (presumably temporary) contradictory state. Of course, in a situation where we have serious theories competing around a contradiction, there is little sense in rejecting one of them a priori just to save the principle of explosion. It seems to be out of question that it is more convenient to tame the logic, rather than to sacrifice a precious (and possibly correct) theory.

This is not only the case for scientific theories. A single digit in a database can of course be extremely valuable than to be just thrown away, and it is already widely recognized that no automated reasoning is possible without means of controlling logical explosion. What is yet not clear is whether the act of guessing is involved in the discovery context of abduction, and furthermore under such conditions, can be the subject of logic. Although it seems that Peirce maintained the negative, we argue that in many interesting cases the process of guessing can be solved semi-automatically by means of careful manipulation of the concept of consistency, viewed as a primitive notion independent from the concept of contradiction, as shown in Sect. 15.2.1 above. In this way we can obtain a reasonably efficient and conceptually simple method for discovering new logical hypotheses that will serve as explanans for a given explanandum.

### 15.3.1 mbC-Tableaux

The beginnings of automatic heuristics by means of paraconsistent tableaux can be traced back to [15.17], although the logic used there is da Costa's  $C_1$ . As argued in the preceding sections, an important LFI is the logic *mbC*. A relevant feature of *mbC* is that it can be defined by means of refutative tableau-type proof procedures. Such backward proof procedures are very convenient for formalizing abductive routines. The ba-

sic idea is that the open branches may be seen as a heuristic device that helps indicate the formulas that would close the tableau, and these formulas are then taken as the explicative hypotheses.

We present below a definition of the notion of an abductive explanation:

#### Definition 15.2

Let  $\Gamma$ ,  $\Delta$  be finite sets of sentences and  $A$  be a sentence in the language of a given logic  $L$ .  $\Gamma$  and  $A$  form an abductive problem and  $\Delta$  is an abductive explanation for the abductive problem if:

1. (Abductive problem): The context  $\Gamma$  is not sufficient to entail  $A$ , that is,  $\Gamma \not\vdash A$
2. (Abductive solution): The enriched context  $\Gamma$  plus  $\Delta$  is sufficient to entail  $A$ , that is,  $\Gamma, \Delta \vdash A$
3. (Nontriviality of solution): The enriched context  $\Gamma$  plus  $\Delta$  is nontrivial, that is, there exists  $B$  such that  $\Gamma, \Delta \not\vdash B$
4. (Vocabulary restriction of solution):  $Var(\Delta) \subseteq Var(\Gamma) \cup Var(A)$
5. (Minimality of solution): by lack of any other criteria, a mathematically minimal  $\Delta$  is a good explanation (in the sense, for example, that it is composed of a set with minimal cardinality and with formulas of minimal length).

While conditions (1) and (2) just define what is an abductive problem and what is a solution, conditions (3), (4) and (5) impose restrictions for a solution to be considered relevant: condition (3) avoids, for instance, that  $\Delta$  be taken as the collection of all formulas, or as a single bottom particle (which would entail any other formulas). Since the compactness theorem holds for *mbC*,  $\Gamma$  and  $\Delta$  can always be taken as finite sets.

Below, we present a tableau system for *mbC*, based on the bivaluation semantics presented in Definition 15.1 [15.7, p. 48]. We use 0 and 1 as syntactic labels to represent the semantic values 0 and 1.

$$\begin{array}{l}
 R1 \frac{\mathbf{0}(\neg X)}{\mathbf{1}(X)} \\
 R2 \frac{\mathbf{1}(\circ X)}{\mathbf{0}(X) \mid \mathbf{0}(\neg X)} \\
 R3 \frac{}{\mathbf{1}(X) \mid \mathbf{0}(X)} \\
 R4 \frac{\mathbf{1}(X_1 \wedge X_2)}{\mathbf{1}(X_1), \mathbf{1}(X_2)} \\
 R5 \frac{\mathbf{0}(X_1 \wedge X_2)}{\mathbf{0}(X_1) \mid \mathbf{0}(X_2)}
 \end{array}$$



$$\begin{array}{l}
 R6 \quad \frac{\mathbf{1}(X_1 \vee X_2)}{\mathbf{1}(X_1) \mid \mathbf{1}(X_2)} \\
 R7 \quad \frac{\mathbf{0}(X_1 \vee X_2)}{\mathbf{0}(X_1), \mathbf{0}(X_2)} \\
 R8 \quad \frac{\mathbf{1}(X_1 \rightarrow X_2)}{\mathbf{0}(X_1) \mid \mathbf{1}(X_2)} \\
 R9 \quad \frac{\mathbf{0}(X_1 \rightarrow X_2)}{\mathbf{1}(X_1), \mathbf{0}(X_2)}
 \end{array}$$

A branch is closed when it contains a couple of labeled formulas  $\mathbf{1}(X)$  and  $\mathbf{0}(X)$ , or when it contains a triad of labeled formulas  $\mathbf{1}(X)$ ,  $\mathbf{1}(\neg X)$  and  $\mathbf{1}(\circ X)$ . The rule  $R3$  could in principle be eliminated, but it is convenient to keep it in the system. The results proven in [15.19] guarantee that the tableau system is sound and complete for  $mbC$ .

Now, making use of the tableau system above, we can illustrate how an abduction mechanism based on the logic  $mbC$  works. Let us consider, first, an example from the folklore; a theory  $\Gamma$  containing the following sentences:  $\Gamma = \{A \rightarrow C, B \rightarrow C\}$ , where  $A$  means *It rained last night*,  $B$  means *the sprinkler was left on*, and  $C$  means *the grass is wet*. If we observe that the grass is wet, and we want to explain why this is so, *It rained last night* is an explanation, but *the sprinkler was left on* is another competitive (though not incompatible) explanation. Tableaux allow for automatically computing nontrivial explanations. After we have such explanations, we may employ some criteria to discard some explanations or choose the best one among competitors. For instance, the hypothetical explanation that the sprinkler was left on may be true, but canceled by the fact that the main water register was known to be off. In any case, some choices may be necessary in order to implement a preference policy for ranking multiple explanations – facts may have precedence over hypothetical explanations, and likelihood may be used to classify explanations. Although this is an important part of the whole question that will affect the usefulness of the automatic explanations produced, it is not part of the abduction problem as originally posed. It is worth noting that the position we are holding here does not require a nonmonotonic logic. The logic used here to produce the abductive output,  $mbC$ , is monotonic. Nonmonotonic reasoning, if necessary, may be used in further steps.

#### Example 15.1

A case where  $mbC$ -tableaux and classical tableaux give the same result:

Let  $\Gamma = \{A \rightarrow C, B \rightarrow C\}$ ; of course  $\Gamma \not\vdash C$ . Running an  $mbC$ -tableau for  $\mathbf{1}(\Gamma) \cup \{\mathbf{0}(C)\}$  produces an open branch containing  $\mathbf{0}(A)$  and  $\mathbf{0}(B)$ . Clearly, this branch

would be closed by  $\mathbf{1}(A)$  or  $\mathbf{1}(B)$ , which indicates that there are three possible abductive solutions:  $\Delta_1 = \{A\}$ ,  $\Delta_2 = \{B\}$  and  $\Delta_3 = \{A, B\}$ .

In the example above, by principle, we consider that a  $\Delta$  minimal provides a better explanation. However, in real-life contexts of reasoning, the choice between these solutions is a problem that may depend on data and criteria to be established by the user of the system.

#### Example 15.2

Let  $\Gamma = \{A \rightarrow B, B \rightarrow C\}$ ; clearly,  $\Gamma \not\vdash C$ . Running an  $mbC$ -tableau for  $\mathbf{1}(\Gamma) \cup \{\mathbf{0}(C)\}$  produces an open branch containing  $\mathbf{0}(A)$  and  $\mathbf{0}(B)$ . Again, it indicates the same three possible abductive solutions:  $\Delta_1 = \{A\}$ ,  $\Delta_2 = \{B\}$  and  $\Delta_3 = \{A, B\}$ .

#### Example 15.3 Impossible Explanations Explained

Suppose that we know that, if it rained last night, then the grass is wet; we know that the grass is wet, but we also know (or have been informed, or have independent evidence for it) that it did not rain. How can we explain that the grass is wet? Let the situation be represented as  $\Gamma = \{A \rightarrow B, \neg A\}$ ; here,  $\Gamma \not\vdash B$ , but no classical tableau is able to find an explanation, since the only possible candidate,  $A$ , has to be ruled out by clause (3) of Definition 15.2 as it entails triviality. However,  $mbC$ -tableaux will be able to provide a solution. In situations like this, common sense suggests that rain may be accepted as an explanation, if the information suggesting that it did not rain is uncertain or dubious. Running an  $mbC$ -tableau yields an open branch containing  $\mathbf{1}(A)$ . Clearly,  $\Delta = \{A\}$  would close the branch, but classically it would not be a solution, since the set of premises contain a formula  $\neg A$ . But  $A$  is indeed a solution, that also indicates the proposition  $\neg A$  is not well established as true – that is, is not consistent. In fact, in  $mbC$ ,  $A, \neg A \vdash \neg \circ A$ . For this reason, this explanation does not violate Definition 15.2. Notice that this scenario cannot be represented by a classical tableau.

#### Example 15.4 Explanations that Avoid Hasty Conclusions

We know that taking certain drugs has beneficial consequences for health, but also the same drugs, under certain conditions, will produce undesirable effects on our health. Represent this situation as  $A \rightarrow B$  and  $A \rightarrow \neg B$ . Under classical reasoning (using classical tableaux, or any other classical inference mechanism) an immediate conclusion would be  $\neg A$ , that is, we should not take these drugs. However, the negative effects could be explained by inappropriate doses, or by different

health conditions in different people, and so on. Using *mbC*-tableaux, however, this case turns out to be an interesting abduction problem, since in *mbC*  $A \rightarrow B, A \rightarrow \neg B \not\vdash \neg A$ , as can be checked by the reader, by consulting the semantics given in Sect. 15.2.1, A Semantics for *mbC*. We are thus invited to look for an abductive explanation: this explanation, automatically produced by the *mbC*-tableau, is that the drug is to be banned only if the contradictory effects are undeniable, that is, if  $\circ B$ . Indeed, in *mbC*  $A \rightarrow B, A \rightarrow \neg B, \circ B \vdash \neg A$ . Hence,  $\Delta = \{\circ B\}$  is an explanation: the resulting *mbC*-tableau is closed.

### Example 15.5 Whodunit?

A diamond was stolen in a hotel room and only two people had entered the room on two different days, Bob and Alice. Since there is only nonconclusive evidence against them and the standard of a proof in a criminal trial must be so strong that there should be *no shadow of doubt*, the police initially consider that they are not guilty, but certainly one of them is guilty, that is, the evidence basis contains  $\Gamma = \{\neg A, \neg B, A \vee B\}$  where *A* and *B* stand, respectively, for *Alice is guilty* and *Bob is guilty*. At this point,  $\Gamma \not\vdash A$  and  $\Gamma \not\vdash B$ , so we have two abductive problems. Now, by running the respective *mbC*-tableaux, we easily see that either  $\circ A$  (meaning that the initial supposition about Alice's innocence was indeed consistent) or  $\circ B$  (meaning, alternatively, that the initial supposition about Bob's innocence was indeed consistent) would decide the question. Indeed, in *mbC*,

$$A \vee B, \neg A, \neg B, \circ A \vdash B \quad \text{and}$$

$$A \vee B, \neg A, \neg B, \circ B \vdash A$$

The presumed innocence of exactly one of them must be revised. Defending the innocence of one of them amounts to the culpability of the other. Notice that they cannot be both innocent – if this were the case, we get triviality.

These examples illustrate the fact that employing logics of formal inconsistency in the general problem of abduction has interesting consequences, automatically producing meaningful explanations that would be imperceptible within the classical environment. *mbC* is not the only choice, and other LFIs would play a similar role. It is worth noting that *mbC* is decidable, and the complexity of its satisfiability problem is no worse than that of the classical satisfiability problem.

## 15.3.2 Quantification

The extension of the ideas of obtaining abductive explanations by means of tableaux to the first-order case

is not only quite natural, but expected in real applications. Although there are some technical complications, from the tableau-proof-theoretical standpoint, all the grounding constructions are already at our disposal: the logic *QmbC*, first-order extension of the propositional *mbC*, has been studied in detail in [15.14]. We recall the main ideas about *QmbC* and show how the underlying tableau procedure can be used in abductive problems. Let  $\Sigma$  be the language of *mbC* enriched with  $\forall$  and  $\exists$ , and *Var* be a set of variables. The formulas of *QmbC* are defined as usual in first-order logics; all the familiar syntactic notions of free and bound variables, closed formulas (sentences), substitution etc., are defined as usual. From the semantic side, sentences of *QmbC* are interpreted by adding the following to the semantics of *mbC*:

- i.  $v(\exists xA) = 1$  iff  $v(A[x/t]) = 1$  for some term  $t$  in  $L$
- ii.  $v(\forall xA) = 1$  iff  $v(A[x/t]) = 1$  for all  $t$  in  $L$
- iii. If  $A$  is a variant of  $B$ , then  $v(A) = v(B)$

We say that  $A$  is a variant of  $B$  (and vice versa) if  $A$  can be obtained from  $B$  by means of addition or deletion of void quantifiers, or by renaming bound variables. It is a theorem of classical first-order logic that if  $A$  and  $B$  are variants of each other, then  $A$  and  $B$  are logically equivalent. However, in *QmbC* the clause (3) above has to be postulated to solve some technical problems that will not be considered in detail here, but the reader can find in [15.14]. From a syntactical perspective, what interests us here for the sake of abduction, a tableau system for *QmbC* is obtained by adding to the tableau rules of *mbC* the following rules for the quantifiers

$$R10 \frac{\mathbf{1}(\forall xA)}{\mathbf{1}(A(x/t))}$$

$$R11 \frac{\mathbf{1}(\exists xA)}{\mathbf{1}(A(x/s))}$$

$$R12 \frac{\mathbf{0}(\forall xA)}{\mathbf{0}(A(x/s))}$$

$$R13 \frac{\mathbf{0}(\exists xA)}{\mathbf{0}(A(x/t))}$$

R14 If  $B$  is variant of  $A$  :

$$\frac{\mathbf{1}(A)}{\mathbf{1}(B)} \quad \text{and} \quad \frac{\mathbf{0}(A)}{\mathbf{0}(B)}$$

The rules above are subjected to the following restriction:  $t$  is an arbitrary term and  $s$  is a new term with respect to  $\forall xA$ , i. e., it does not appear in any branch containing  $\forall xA$  (respectively for  $\exists xA$ ).

The method introduced here for obtaining automatic explanations can thus be extended to first-order theories. Of course, this involves some additional com-

plications, because *QmbC*-tableaux, as much as their classical counterparts, are not a decision procedure for *QmbC*-validity – indeed, *QmbC* is undecidable. Let us see an example below.

#### Example 15.6

Consider the following set of propositions:  $\Gamma = \{\forall x(Cx \rightarrow Bx), \forall x(Gx \rightarrow Bx), \neg Ca\}$ . Here,  $\Gamma \not\vdash Ba$ . Running an *mbC*-tableau for  $\mathbf{1}(\Gamma) \cup \{\mathbf{0}(Ba)\}$  produces an open branch containing  $\mathbf{0}(Ca)$ ,  $\mathbf{1}(\neg Ca)$  and  $\mathbf{0}(Ga)$ . Classically, the only candidate to be an abductive explanation is *Ga*. But from the point of view of *QmbC*, we obtained two possible explanations, since  $\mathbf{1}(Ca)$  also closes that branch. In the latter case, a further conclusion is that *Ca* is not *consistent*, i. e., is not well established as a true proposition. Thus, this explanation does not violate Definition 15.2.

## 15.4 Modality

Modal logics and paraconsistent logics are cousins. In 1948, while attempting to answer a question posed by J. Lukasiewicz, S. Jaśkowski presented the first formal system for a paraconsistent logic with his *discussive logic*. Interestingly enough, his logic was framed in terms of modalities, and later on it was proven to be a particular case of the family of LFIs [15.7]. However, it was only in 1986 that the first modal paraconsistent system was proposed in [15.20], with the aim of dealing with deontic paradoxes. That system was a modal extension of da Costa's paraconsistent logic  $C_1$ . This approach has been extended by means of deontic modalities combined with LFIs, as developed in [15.21, 22].

Paraconsistent negation can be regarded as a kind of modal operator, considering the fact that the classical negation for possibility (and, a fortiori, for necessity) has a paraconsistent behavior. Namely, the operator  $\neg$ , defined as

$$\neg A \stackrel{\text{def}}{=} \diamond \sim A,$$

is a paraconsistent negation where, as usual,  $\sim$  denotes the classical negation. This relationship has been studied in [15.23], both with respect to the standard modal logic *S5* and to four-valued modal logics [15.24]. It is worth noting that the fact that  $\diamond \sim A$  defines a paraconsistent negation was already observed in 1987 in [15.25], when a Kripke-style semantics was proposed for Sette's logic *PI* based on Kripke frames for modal logic *T*.

One of the interests in paraconsistent modal logics is the potential of dealing with, or even avoiding,

To the extent that LFIs permit fine control of reasoning in the presence of inconclusively established hypotheses, particularly under contradictions, the mechanism presented here is thus capable of proposing solutions for an extensive class of abductive problems. As we have seen, the *mbC*-tableaux increase the range of options provided by classical reasoning. The issues discussed here have much in common with belief revision, default reasoning, the closed-world assumption, and negation as failure of logic programming, as well as databases with evolutionary constraints, thus making our proposal valuable for several applications. Abduction, however, can also be regarded, from a much more abstract standpoint, as a companion for argumentation (see chapter by Barés and Fontaine, this section, for a proposal in this direction). From this perspective, any attempt to make abduction somewhat closer to deduction is welcome.

some modal paradoxes. Moral dilemmas are a typical situation in which paraconsistent modal logics provide a tool to handling contradictions without triviality. Let us take as an example the well-known dilemma, posed by [15.26], of the man in occupied France who, on the one hand, wants to fight the Nazis but, on the other, must take care of his mother. He believes that each alternative is a moral obligation, but doing one implies not doing the other. Let *A* and *B* be, respectively, *fight the Nazis* and *take care of his mother*. Let the symbols *O* and *P* mean, respectively *it is obligatory that* and *it is permitted that* (as usual  $\square$  means necessity and  $\diamond$  means possibility). From the premises

$$OA, OB, \sim \diamond(A \wedge B),$$

plus the following principles of deontic logic

$$\square(A \rightarrow B) \rightarrow (OA \rightarrow OB),$$

$$OA \rightarrow \sim O \sim A,$$

and given that in classical modal logic  $\sim \diamond(A \wedge B)$  is equivalent to  $\square(A \rightarrow \sim B)$ , a contradiction may be obtained in a few steps. On the other hand, a paraconsistent modal logic may handle the contradiction without triviality.

Another example is Urmson's paradox. In this case, the modal paradox is just avoided. Consider the following proposition

- (X) It is optional that you attend my talk or not, but your choice is not indifferent.

It is clear that the notions *optional* ( $Opt$ ) and *indifferent* ( $Ind$ ) must be distinct in (X). Again, let  $P$  and  $O$  mean *permitted* and *obligatory*. It is natural in modal logic to formalize  $Opt$  and  $Ind$  as

$$Opt(A) \stackrel{\text{def}}{=} PA \wedge P\sim A.$$

$$Ind(A) \stackrel{\text{def}}{=} \sim OA \wedge \sim O\sim A.$$

In classical modal logic a contradiction occurs because  $\sim$  is a classical negation and  $OA$  is equivalent to  $\sim P\sim A$ . Hence, it is easy to see that  $Opt$  and  $Ind$  are equivalent. So, attending the talk is simultaneously optional and not optional. On the other hand, if a nonexplosive negation  $\neg$  is available, it can be used to express the notions  $Opt$  and  $Ind$ . In this way, no paradox occurs because  $OA$  and  $\neg P\neg A$  are no longer equivalent.

Paraconsistent deontic logics have also been studied in the literature for quite a some time [15.20], and deontic counterparts of LFIs, the *logics of deontic (in)consistency* (LDIs), have been introduced in [15.21]. These logics are shown to be able to handle deontic paradoxes, as the well-known Chisholm's paradox. Since contradictory obligations do not trivialize such LDIs, several paradoxes involving conflicting obligations are dissolved [15.22].

It is important to note, however, that the potential of combining paraconsistency and modalities extends far beyond deontic issues. Not only can some problems, described in [15.27], be thought in paraconsistent terms, but also certain problems and paradoxes in epistemic and doxastic logics gain new insight when regarded from paraconsistent perspective.

A detailed investigation of the relationship between LFIs and their modal versions is carried out in [15.28], where the so-called *anodic systems* (purely positive modal systems) introduced in [15.29] are extended by adding certain paraconsistent axioms based on LFIs, defining a class of modal systems called *cathodic systems* (modal systems involving degrees of negations). For an explanation of the terms *anodic* and *cathodic* see [15.29]. A semantic interpretation of cathodic systems is given in [15.28], where it is shown that the cathodic systems can be semantically characterized in two different ways: by means of *Kripke-style semantics* and by means of *modal possible-translations semantics*.

In the following sections we start by presenting a positive (i. e., *anodic*) modal system, that can be enhanced with degrees of negation, as shown in [15.7], so as to obtain a family of *cathodic* systems. We start with the anodic modal system  $K^\diamond$ , a negationless fragment of the well-known modal system  $K$ .

The first paraconsistent modal system we shall consider is  $mbC^\square$ , which will be obtained as an extension

of the anodic modal system  $K^\diamond$ . Then, we show how paraconsistent modal logics that are fragments of the familiar systems  $T$ ,  $S4$  and  $S5$  may be obtained, as extensions of  $mbC^\square$ . Correct and complete semantics are presented for all of these systems.

### 15.4.1 The Anodic System $K^\diamond$

In this section, a purely positive bimodal system  $K^\diamond$  will be defined in a negationless language that is both an extension of CPL+ and a fragment of  $K$ . The only modal axioms are positive versions of the distribution axiom ( $K$ ) (namely, ( $K$ ), ( $K1$ ), ( $K2$ ) and ( $K3$ )) together with the usual necessitation rule ( $Nec$ ).

The language  $L_2$  of  $K^\diamond$  has the following set of connectives:  $\{\vee, \wedge, \rightarrow, \square, \diamond\}$ . Notice that both modal operators are needed as primitive because one cannot be defined in terms of the other, given that no negation is available. The set of *formulas* of  $K^\diamond$  is obtained as typically done in modal logic. The formulas of  $K^\diamond$  are represented by Roman capital letters, and sets of formulas are represented by uppercase Greek letters  $\Gamma, \Delta$  etc. The definition of a derivation of  $A$  from a set of premises  $\Gamma$  ( $\Gamma \vdash_{K^\diamond} A$ ) is the usual one: a finite sequence of formulas  $B_1 \dots B_n$  such that  $A$  is  $B_n$  and each  $B_i$ ,  $1 \leq i \leq n$  is an axiom, a formula that belongs to  $\Gamma$ , or a result of an inference rule. A theorem is a formula derived from the empty set of premises. When there is no risk of ambiguity, we write just  $\vdash$  instead of  $\vdash_{K^\diamond}$ .

#### Definition 15.3

The anodic modal system  $K^\diamond$  is defined by adding to CPL+ the following modal axiom-schemas and modal rule:

- |       |  |
|-------|--|
| (K)   | $\square(A \rightarrow B) \rightarrow (\square A \rightarrow \square B)$   |
| (K1)  | $\square(A \rightarrow B) \rightarrow (\diamond A \rightarrow \diamond B)$ |
| (K2)  | $\diamond(A \vee B) \rightarrow \diamond A \vee \diamond B$                |
| (K3)  | $(\diamond A \rightarrow \square B) \rightarrow \square(A \rightarrow B)$  |
| (Nec) | $\vdash A$ implies $\vdash \square A$                                      |

A modal system is called *normal* if it contains the distribution axiom ( $K$ ) and the necessitation rule ( $Nec$ ) among its axioms and rules, and *minimal* if it has only ( $K$ ) as a modal axiom and ( $Nec$ ) as a modal rule.  $K^\diamond$  is minimal and normal. In addition, it is not difficult to see that  $K^\diamond$  is a fragment of the system  $K$ , since the axioms ( $K1$ )-( $K3$ ) can be derived in the system  $K$ , as the reader can verify as an exercise (remember that  $K$  is obtained by adding to CPL+ the axiom  $K$  and the necessitation rule). As well as  $mbC$ ,  $K^\diamond$  satisfies the properties of reflexivity, monotonicity, cut and compactness. Besides, the deduction theorem is also valid.

### A Semantics for $K^\diamond$

Now a Kripke-style semantics for  $K^\diamond$  will be presented, or in other words, a possible world semantics. Such a semantics has two basic and primitive notions: (i) the notion of possible world, usually understood as a way things in the world might have been, and (ii) the notion of accessibility, or relative possibility, between worlds. The basic idea of (ii) is that it may be the case that some alternative worlds are not possible with respect to a given world. These intuitions are expressed as follows:

#### Definition 15.4

An *anodic frame* is a birelational structure  $\mathfrak{F} = \langle W, R_\square, R_\diamond \rangle$ .  $W$  is a nonempty set (a set of possible worlds), and  $R_\square$  and  $R_\diamond$  are any binary relations on  $W$  (accessibility relations between worlds).

Notice that the definition above needs two binary relations because  $\square$  and  $\diamond$  are not interdefinable. Frames play a central role in modal logic because the conditions imposed on the accessibility relation are placed on frames. In  $K^\diamond$ , as well in  $K$ , there is no condition on the relations  $R_\square$  and  $R_\diamond$ .  $\mathfrak{F}$  is said to be a *frame* for a modal system  $S$  if every theorem of  $S$  is valid on  $\mathfrak{F}$ . It will become clear throughout this section.

A *model* for a modal system  $S$  is defined by specifying what formulas receive the semantic value 1 in which worlds. The following definition specifies what a model for the anodic system  $K^\diamond$  is:

#### Definition 15.5

A model for  $K^\diamond$  is a pair  $\mathfrak{M} = \langle \mathfrak{F}, v \rangle$  where  $\mathfrak{F}$  is a frame for  $K^\diamond$  and  $v : \text{Var} \times W \rightarrow \{0, 1\}$  is a function satisfying:

- (i)  $v(p, w) = 1$  or  $v(p, w) = 0$
- (ii)  $v(A \rightarrow B, w) = 1$  iff  $v(A, w) = 0$  or  $v(B, w) = 1$
- (iii)  $v(A \wedge B, w) = 1$  iff  $v(A, w) = 1$  and  $v(B, w) = 1$
- (iv)  $v(A \vee B, w) = 1$  iff  $v(A, w) = 1$  or  $v(B, w) = 1$
- (v)  $v(\square A, w) = 1$  iff  $v(A, w') = 1$ , for all  $w' \in W$  such that  $wR_\square w'$
- (vi)  $v(\diamond A, w) = 1$  iff  $v(A, w') = 1$ , for some  $w' \in W$  such that  $wR_\diamond w'$ .

A sentence  $A$  is said to be *satisfied in a model*  $\mathfrak{M}$ , if there is a  $w \in W$  such that  $v(A, w) = 1$  (notation:  $\mathfrak{M}, w \models A$ ). A sentence  $A$  is said to be *valid in a model*  $\mathfrak{M}$ , if  $v(A, w) = 1$  for all  $w \in W$  (notation:  $\mathfrak{M} \models A$ ). A sentence  $A$  is said to be *valid in a frame*  $\mathfrak{F}$ , if  $A$  is valid in all models  $\mathfrak{M}$  based on  $\mathfrak{F}$  (notation:  $\mathfrak{F} \models A$ ).

A special class of frames  $\mathcal{F}$  is the collection of frames that satisfy some condition imposed on the relation  $R$ . Examples are the special class of reflexive

frames, where  $R$  is reflexive, and the class of general frames, where there is no condition imposed on  $R$ . Remember that there is no condition on the accessibility relations in  $K^\diamond$ .

A sentence  $A$  is a *semantic consequence* of  $\Gamma$  with respect to the class  $\mathcal{F}$  of frames if  $\mathfrak{F} \models \Gamma$  then  $\mathfrak{F} \models A$ , for each  $\mathfrak{F} \in \mathcal{F}$ , where  $\mathfrak{F} \models \Gamma$  means that  $\mathfrak{F} \models B$  for all  $B \in \Gamma$ .

Notice that the semantics in the clauses above, (i) to (iv) are classical in the sense that, in each world  $w$ , the propositional connectives behave classically. This is expected, since  $K^\diamond$  is an extension of CPL+. The system  $K^\diamond$  can be proven sound and complete with respect to the semantics given above. The proof is a bit complicated, and requires some technical details [15.30].

### 15.4.2 The Logic $mbC^\square$

Catholic systems can be obtained by extending  $K^\diamond$ , adding to its language a paraconsistent negation  $\neg$  and the consistency operator  $\circ$  plus specific paraconsistent axioms. The modal paraconsistent logic  $mbC^\square$  is defined by adding to  $K^\diamond$  the following axioms

$$\text{Ax. bc1. } \circ A \rightarrow (A \rightarrow (\neg A \rightarrow B)),$$

$$\text{Ax. 10. } A \vee \neg A.$$

It is clear that  $mbC^\square$  is both an extension of  $mbC$  and of  $K^\diamond$ . Notice that the properties of reflexivity, monotonicity, cut and compactness hold for  $mbC^\square$ , as well as the deduction theorem. Given that it is possible to define a classical negation  $\sim$  in  $mbC$ , then the possibility operator  $\diamond$  can be defined from the necessity operator  $\square$  as usual in modal logic.

$$\diamond A \stackrel{\text{Def}}{=} \sim \square \sim A \tag{15.5}$$

Hence, the axioms (K1)–(K3) are innocuous in  $mbC^\square$ , since they can be easily proven as theorems, as the reader can verify.

#### A Semantics for $mbC^\square$

##### Definition 15.6

A *frame* is a relational structure  $\mathfrak{F} = \langle W, R \rangle$ , where  $W \neq \emptyset$  is a universe and  $R$  is a binary relation on  $W$  (notice that now we need only one relation that covers both  $\diamond$  and  $\square$ ).

A bivalued relational model for the catholic system  $mbC^\square$  is defined as follows:

##### Definition 15.7

A *bivalued relational model*  $\mathfrak{M}$  for  $mbC^\square$  is a pair

$\langle \mathfrak{F}, v \rangle$  where  $\mathfrak{F}$  is a frame and  $v : Var \times W \longrightarrow \{0, 1\}$  is a bivalued modal assignment satisfying the conditions:

- (i)  $v(p, w) = 1$  or  $v(p, w) = 0$ ;
- (ii)  $v(A \rightarrow B, w) = 1$  iff  $v(A, w) = 0$   
or  $v(B, w) = 1$
- (iii)  $v(A \wedge B, w) = 1$  iff  $v(A, w) = 1$   
and  $v(B, w) = 1$
- (iv)  $v(A \vee B, w) = 1$  iff  $v(A, w) = 1$  or  $v(B, w) = 1$
- (v)  $v(A, w) = 0$  implies  $v(\neg A, w) = 1$
- (vi)  $v(\Box A, w) = 1$  iff  $\forall w'(wRw'), v(A, w') = 1$
- (vii)  $v(\Diamond A, w) = 1$  iff  $\exists w'(wRw'), v(A, w') = 1$
- (viii)  $v(\circ A, w) = 1$  implies  $v(A, w) = 0$   
or  $v(\neg A, w) = 0$ .

The notion of validity in a frame is defined as usual.

### 15.4.3 Extensions of $mbC^\Box$

From  $mbC^\Box$ , stronger systems may be defined. The well-known modal systems  $D$ ,  $T$ ,  $S4$ ,  $B$  and  $S5$  are obtained by adding one or more of the following axioms to system  $K$

- (D)  $\Box A \rightarrow \Diamond A$  ;
- (T)  $\Box A \rightarrow A$  ;
- (4)  $\Box A \rightarrow \Box \Box A$  ;
- (B)  $A \rightarrow \Box \Diamond A$  ;
- (5)  $\Diamond A \rightarrow \Box \Diamond A$  .

Now, axiomatic systems for the modal logics listed below are obtained as follows

$$\begin{aligned} D &\stackrel{\text{def}}{=} K + (D) \\ T &\stackrel{\text{def}}{=} K + (T) \\ S4 &\stackrel{\text{def}}{=} K + (T) + (4) \\ B &\stackrel{\text{def}}{=} K + (T) + (B) \\ S5 &\stackrel{\text{def}}{=} K + (T) + (5) \end{aligned}$$

Sound and complete semantics can be obtained for the systems above by imposing appropriate conditions on frames. A *frame*  $\mathfrak{F}$  is

Reflexive iff for every  $w \in W$ ,  $wRw$  ;

Symmetric iff for every  $w, w' \in W$ ,

$wRw'$  implies  $w'Rw$  ;

Transitive iff for every  $w, w', w'' \in W$ ,

$wRw'$  and  $w'Rw''$  implies  $wRw''$  ;

Serial iff for every  $w \in W$  there is some  $w' \in W$  .  
such that  $wRw'$

The modal logics  $K$ ,  $K^\Diamond$  and  $mbC^\Box$  have no condition on frames – that is, they have no condition on the relation  $R$  of accessibility between worlds. Starting from  $K$ , the systems  $D$ ,  $T$ ,  $S4$ ,  $B$  and  $S5$  are obtained imposing the following condition on frames

$D$  : serial;

$T$  : reflexive;

$B$  : reflexive and symmetric;

$S4$  : reflexive, transitive;

$S5$  : reflexive, symmetric, transitive.

Now,  $mbC^\Box$  may be extended, obtaining paraconsistent modal systems that are both extensions and fragments of each of the systems above.

$$\begin{aligned} mbC^D &\stackrel{\text{def}}{=} mbC^\Box + (D) , \\ mbC^T &\stackrel{\text{def}}{=} mbC^\Box + (T) , \\ mbC^B &\stackrel{\text{def}}{=} mbC^\Box + (T) + (B) , \\ mbC^{S4} &\stackrel{\text{def}}{=} mbC^\Box + (T) + (4) , \\ mbC^{S5} &\stackrel{\text{def}}{=} mbC^\Box + (T) + (5) . \end{aligned}$$

Kripke-style semantics for the above paraconsistent modal systems may be obtained by just adding clauses corresponding to the respective restrictions on frames to the semantics for  $mbC^\Box$ .

The modal logic  $S5$ , where the relation is reflexive, symmetric and transitive, or in other words every possible world may access every possible world, is usually accepted as the system that better expresses the intuitive idea of Leibniz according to which possibility amounts to truth in some possible world and necessity amounts to truth in every possible world. Kripke has refined the basic intuition of Leibniz, ingeniously providing complete and sound semantics for the axiomatic systems that were already available.  $S5$  is also taken as the strongest propositional modal system among those presented here. Besides, it is claimed, although not without dispute, that  $S5$  is the system that better expresses the notion of logical necessity. It is worth noting that the system  $mbC^{S5}$  may be seen an extension of  $S5$ , since a classical negation is definable in it. Hence,  $mbC^{S5}$  presents itself as a very powerful paraconsistent modal system, capable of expressing everything that is expressed by  $S5$  plus the resources of the paraconsistent logic  $mbC$ . Applications of  $mbC^{S5}$

in philosophical issues in modal logic, as well as the possibility of enhancing it in order to fit some contexts of reasoning or philosophical problems related to modalities in general are topics that deserve further attention.

Although a large class of anodic and cathodic multimodal logics can be shown to be complete with respect to Kripke frames, an interesting point about anodic and cathodic modal logics is that some incomplete systems can be found in both families. *Bueno-Soler* in [15.31] shows that some classes of cathodic multimodal paraconsistent logics (that is, logics endowed with weak forms of negation) are incomplete with

respect to Kripke semantics. The meaning of this kind of incompleteness is also discussed in [15.31], but, surprisingly, the phenomenon of modal incompleteness is also found among purely positive (multi)modal logics: *Bueno-Soler* in [15.30] obtains some examples of Kripke-incompletable purely positive modal logics, demonstrating that modal incompleteness is a result of the interaction of modalities, independent of negation. The incompleteness in the case of cathodic modal logics, however, does not obtain with respect to possible-translations semantics, marking a distinction between this kind of semantics and Kripke semantics for modalities.

## 15.5 On Alternative Semantics for mbC

The semantic characterization of nonclassical logics in general is not an easy task. As it was proven in [15.7], *mbC*, as well as the majority of LFIs, cannot be characterized by a finite logical matrix. Moreover, *mbC* cannot be algebraizable, even in the wide sense of *Blok* and *Pigozzi* [15.32]. These restricting results also hold for several LFIs extending *mbC*, and for the logics of the  $C_n$  hierarchy. Being so, these systems lie outside standard semantic analysis such as categorical or algebraic semantics. Because of this, the development of alternative semantic techniques for these kinds of LFIs is an important and challenging task.

This section briefly describes some alternative semantic approaches to *mbC* and many other LFIs. All these semantics have an intrinsic nondeterministic character, suggesting that nondeterminism is the correct perspective for this kind of logics.

Let us begin by recalling the valuation semantics for *mbC* introduced above (Definition 15.1). The clauses for the binary connectives  $\rightarrow$ ,  $\vee$  and  $\wedge$  are the same as for classical logic. However, as we have seen, the clauses for the paraconsistent negation  $\neg$  and the consistency operator  $\circ$  give the following (nondeterministic) quasi-matrices

$A$	$\neg A$	$\circ A$	
1	1	0	$v_1$
	0	1	$v_2$
		0	$v_3$
0	1	1	$v_4$
		0	$v_5$

Accordingly, there are five possible valuations for *mbC* concerning propositions  $A$ ,  $\neg A$  and  $\circ A$ , namely  $v_1$  to  $v_5$ , such that  $v_1(A) = v_1(\neg A) = 1$  and  $v_1(\circ A) = 0$ ;  $v_2(A) = v_2(\circ A) = 1$  and  $v_2(\neg A) = 0$ , and so on. Ob-

serve that, if  $v(A) = v(\neg A) = 1$ , then  $v(\circ A)$  is forced to be 0, by the gentle explosion law (*bc1*). Otherwise, when  $v(A) \neq v(\neg A)$  the value of  $v(\circ A)$  is arbitrary. As *Avron* observed in [15.33], simultaneously taking into account the semantic values of  $A$ ,  $\neg A$  and  $\circ A$  instead of taking these values separately, five semantic values derived from the five valuations  $v_1$ – $v_5$  can be obtained, namely  $B_5 = \{t, T, t_0, F, f_0\}$ , where  $t = (1, 1, 0)$ ,  $T = (1, 0, 1)$ ,  $t_0 = (1, 0, 0)$ ,  $F = (0, 1, 1)$  and  $f_0 = (0, 1, 0)$ . This allows a semantic characterization of *mbC* in terms of nondeterministic matrices (see below).

This approach, however, can be traced back to *Fidel's* ideas, already presented in his definition of the notion of twist structures in [15.34] (independently of *Vakarelov* in [15.35]). Indeed, in his PhD thesis [15.36], *Fidel* claims that, by considering  $n$ -tuples of a given class of algebras (for some  $n \geq 2$ ), it is possible to analyze the structure of the algebraic models of other nonclassical logics. Besides twist structures for Nelson's logic (where  $n = 2$ ), he introduces in [15.36, Chap. 4] a new semantics for the logic of Ockham algebras  $P_{3,1}$  by considering triples of elements of distributive lattices. Any triple  $(a, b, c)$  is such that  $a$  represents the value of a formula  $A$ , while  $b$  and  $c$  represent the values of  $\neg A$  and  $\neg\neg A$ , respectively.

These ideas from *Fidel* inspired [15.37], wherein the notions of *snapshots* and *swap structures* for *mbC* (and some LFIs extending it) were introduced.

### Definition 15.8

Let  $\mathcal{A} = \langle A, \wedge, \vee, \rightarrow, 0, 1 \rangle$  be a Boolean algebra, and let  $\mathbb{B}_{\mathcal{A}} = \{(a, b, c) \in A \times A \times A : a \vee b = 1 \text{ and } a \wedge b \wedge c = 0\}$ . A *swap structure* for *mbC* over  $\mathcal{A}$  is any multialgebra  $\mathcal{B} = \langle B, \wedge, \vee, \rightarrow, \neg, \circ \rangle$  such that  $B \subseteq \mathbb{B}_{\mathcal{A}}$  and where the multivalued operations satisfy the following

conditions, for every  $(a, b, c)$  and  $(a', b', c')$  in  $B$ , and for each  $\# \in \{\wedge, \vee, \rightarrow\}$ :

- (i)  $(a, b, c)\#(a', b', c')$   
 $\stackrel{\text{def}}{=} \{(a'', b'', c'') \in B : a'' = a\#a'\};$
- (ii)  $\neg(a, b, c) \stackrel{\text{def}}{=} \{(a'', b'', c'') \in B : a'' = b\};$
- (iii)  $\circ(a, b, c) \stackrel{\text{def}}{=} \{(a'', b'', c'') \in B : a'' = c\}.$

The elements of a given swap structure are called *snapshots*. Intuitively, a snapshot  $x = (a, b, c)$  simultaneously keeps track of the value  $a$  of a given formula  $A$ , the value  $b$  of  $\neg A$ , and the value  $c$  of  $\circ A$ . Because of this,  $a \vee b = 1$  by the principle of excluded middle, and  $a \wedge b \wedge c = 0$  by the gentle explosion law, which are both valid in  $mbC$ .

The binary operations (cf. clause (i)) kept fix the first coordinate of the given snapshots: the second and third coordinates of the output do not depend on the given data. It reflects the fact that the binary connectives are classical, but the truth value of  $\neg(A\#B)$  and  $\circ(A\#B)$  are unrelated to the truth values of  $A, B, \neg A, \neg B, \circ A$  and  $\circ B$ . With respect to the unary connectives (cf. clauses (ii) and (iii)), the negation  $\neg x$  of a snapshot  $x$  is the set of snapshots with the second coordinate of  $x$  on the first place. Accordingly, the consistency  $\circ x$  of  $x$  has the third coordinate of  $x$  in the first place. This reflects the intuitive meaning of the components of a snapshot, as mentioned above. As in the case of (i), the second and third coordinates of  $\neg x$  are independent from  $x$ , since  $\neg\neg A$  and  $\circ\neg A$  are independent from  $A, \neg A$  and  $\circ A$ . The same observation holds for  $\circ x$ .

Swap structures are multialgebras defined over suitable subsets of  $A^3$ , for any Boolean algebra  $A$ . Hence they naturally determine a family of nondeterministic matrices in the sense of Avron and Lev (see below) which semantically characterize  $mbC$  and other LFIs.

Any swap structure  $\mathcal{B}$  for  $mbC$  determines a nondeterministic matrix  $\mathcal{M}(\mathcal{B}) \stackrel{\text{def}}{=} \langle \mathcal{B}, D_{\mathcal{B}} \rangle$  such that  $D_{\mathcal{B}} = \{x \in \mathcal{B} : x_1 = 1\}$  is the set of designated truth values (here,  $x_1$  denotes the first coordinate of the snapshot  $x$ ). Let  $\mathbb{K}$  be the class of such nondeterministic matrices, and define the semantic consequence with respect to swap structures for  $mbC$  as follows:  $\Gamma \models_{\mathbb{K}} A$  iff  $\Gamma \models_{\mathcal{M}(\mathcal{B})} A$ , for every swap structure  $\mathcal{B}$  for  $mbC$  (the consequence relation on each nondeterministic matrix is defined by means of valuations in the sense of Avron and Lev, see below). Then, the following result can be proven [15.37]:

#### **Theorem 15.1 Soundness and Completeness**

Let  $\Gamma \cup \{A\}$  be a set of formulas in  $mbC$ . Then,  $\Gamma \vdash_{mbC} A$  if and only if  $\Gamma \models_{\mathbb{K}} A$ .

The last result can be strongly improved: Let  $\mathcal{A}_2$  be the two-elements Boolean algebra, and let  $\mathbb{K}_2$  be the nondeterministic matrix  $\mathcal{M}(\mathcal{B})$  induced by the unique swap structure  $\mathcal{B}$  for  $mbC$  over  $\mathcal{A}_2$  with domain  $\mathbb{B}_{\mathcal{A}_2}$ . Observe that  $\mathbb{B}_{\mathcal{A}_2}$  coincides with the set  $B_5 = \{t, T, t_0, F, f_0\}$  mentioned at the beginning of this section (with the notation introduced therein). Then, the following result can be proven [15.37]:

#### **Theorem 15.2 Soundness and Completeness with Respect to $\mathcal{A}_2$**

Let  $\Gamma \cup \{A\}$  be a set of formulas in  $mbC$ . Then,  $\Gamma \vdash_{mbC} A$  if and only if  $\Gamma \models_{\mathbb{K}_2} A$ .

Theorem 15.2 is nothing more than the semantic characterization of  $mbC$  by means of Nmatrices obtained in [15.33]. The notion of *nondeterministic matrices* (or *Nmatrices*) was proposed by Avron and Lev in [15.38], and afterwards studied by Avron and his collaborators. Basically, an Nmatrix is a logical matrix  $\mathcal{M} = \langle \mathcal{A}, D \rangle$  such that each operation in the algebra  $\mathcal{A}$  is multivalued, that is, if  $c^{\mathcal{M}}$  is an  $n$ -ary operator of  $\mathcal{A}$  (interpreting a connective  $c$ ) and  $(a_1, \dots, a_n) \in A^n$  (where  $A$  is the domain of the multialgebra) then  $c^{\mathcal{M}}(a_1, \dots, a_n)$  is a finite, nonempty subset of  $A$ . The valuations are mappings  $v$  assigning to each formula of the logic being interpreted by  $\mathcal{M}$  an element of  $A$  in a coherent way, as follows:  $v(c(A_1, \dots, A_n)) \in c^{\mathcal{M}}(v(A_1), \dots, v(A_n))$ . The semantic consequence is defined as in the case of standard logical matrices, namely:  $A$  follows from a set  $\Gamma$  of formulas if, for every valuation  $v$ ,  $v(A) \in D$  (the set of designated truth values) whenever  $v(\gamma) \in D$  for every  $\gamma$  in  $\Gamma$ .

As observed above, the Nmatrices for  $mbC$  considered by Avron were defined by means of a notion equivalent to snapshots in  $B_5$ , with the same motivations described here. Swap structures propose a generalization of this approach to arbitrary Boolean algebras (instead of taking the two-valued Boolean algebra). From this, some interesting perspectives to study LFIs from the point of view of multialgebras arise, by adapting appropriate techniques from algebraic logic.

Finally, the *possible-translations semantics* paradigm will be briefly surveyed here. The *possible-translations semantics* (PTS) were introduced by Carnielli in 1990 [15.39] as an attempt to offer a more palatable interpretation, from the philosophical point of view, for some nonclassical logics, and especially for paraconsistent logics. PTSs are based on the notion of translations between logics (that is, mappings that preserve consequence relations). Using an analogy with natural languages, translations can be thought of as different *world views*, and the concept of PTSs is a way



of interpreting a given logic  $L$  as the combination of all possible *world views*, represented by an appropriate set of translations of the formulas of  $L$  into a class of logics with known consequence relation. By choosing an adequate collection of such translations, the object logic  $L$  acquires a semantic meaning throughout the logics into which it is translated (the so-called *traducts*). When the translations and traducts are decidable, PTSs offer a decision procedure.

In formal terms, a possible-translations semantics for a logic  $L$  is a pair of families  $\{\{L_i\}_{i \in I}, \{f_i\}_{i \in I}\}$  such that each  $L_i$  is a logic and each  $f_i$  is a mapping from  $L$  to  $L_i$  such that  $\Gamma \vdash_L A$  if and only if  $f_i[\Gamma] \vdash_{L_i} f_i(A)$ , for every  $i \in I$ .

In [15.40] it was shown that possible-translations semantics are able to express Nmatrices, which implies that the latter is a particular case of the former. On the

other hand, from the viewpoint of the consequence relation, and leaving aside algebraic aspects, swap-structure semantics is nothing more than a semantics defined by a family of nondeterministic matrices. Therefore, it can be seen as a particular case of possible-translations semantics, as pointed out above.

From the previous considerations, it can be seen that possible-translations semantics is a semantic tool having a wide scope of applications and a high degree of generality. On the other hand, Nmatrices and swap structures offer semantic interpretations with a more algebraic-oriented perspective, thus being more intuitive. In particular, swap structures offer not only a promising way to study *mbC* and other LFIs from the point of view of algebraic logic, but also a new angle for understanding why such logics matter for the analysis of reasoning.

## 15.6 Conclusions

Classical logic is a very powerful tool for modeling both informal and scientific reasoning. However, the fact that it is not able to handle contradictions is an important constraint. Although paraconsistent logics have been gaining an increasingly relevant place in contemporary philosophical debate, there is still some resistance against recognizing their philosophical significance, and it is likely that this reluctance is primarily related to the awkwardness of the claim that there are meaningful contradictory propositions about reality which are true. An epistemological interpretation of the acceptance of contradictions by paraconsistent logics has been presented and defended here. It has also been argued that the logics of formal inconsistency are capable of expressing contradictions as *conflicting evidence*, a notion weaker than truth that occurs in several contexts of informal reasoning and scientific research. The unary operator  $\circ$  initially had the purpose of representing the metatheoretical notion of consistency within the object language. But the idea has been further developed in such a way that it may receive alternative meanings. Actually, any *logical property* may have its validity restricted to a group of propositions, depending on the context of reasoning one wants to represent. This has been done in the logics *mbCD* and its extension *LET<sub>k</sub>*, which restrict not only explosion but also excluded middle.

In Sect. 15.3 we have argued that the logic *mbC*, and its first-order extension *QmbC*, are naturally connected to the general question of finding solutions for (respectively, sentential and quantified) problems of abduction. Although the use of tableaux in abductive contexts is not a novelty, paraconsistent tableaux do represent an ad-

vance, due to the problem of contradictory character of certain abductive solutions. The subject of abduction is currently investigated, and employed, in several fields such as artificial intelligence research, formal systems of law and norms, diagnostic expert systems and databases. In such cases, theory and practice rely more and more on the paraconsistency paradigm. In this way, logic-based abduction, regarded from a paraconsistent perspective, acquires special interest. Taking into account, for instance, that probabilistic abductive reasoning as a form of taking decisions is already extensively used in areas of higher degree of uncertainty such as medical diagnoses and pharmaceutical tests, further investigation on paraconsistent probabilistic methods combined with tableau would open a new research window, adding interest to the type of approach here expounded.

With respect to the paraconsistent modal logics surveyed in the Sect. 15.4, it is worth noting that the epistemological interpretation of contradictions in reasoning scenarios can be naturally combined with modal logics. The informal interpretation here suggested, according to which a pair of contradictory propositions  $A$  and  $\neg A$  means that there is conflicting evidence about  $A$ , and that  $\circ A$  means that the truth value of  $A$  has been conclusively established, can be combined with modalities with interesting new aspects. For example,  $\circ \Box A$  may be understood as meaning that the question about whether or not  $A$  is necessary has been conclusively established, and  $\Box \circ A$  may be understood as meaning that necessarily the truth value of  $A$  will be conclusively established, assigning thus a realistic determinism on the proposition  $A$ . The mere possibility of performing this

separation by means of cathodic and anodic modalities already seems to offer new logical perspectives to some modal dilemmas, although this is not the appropriate place to develop such analysis. In all cases, the ideas here surveyed certainly offer new tools for philosophers and logicians.

Finally, in Sect. 15.5, some alternative semantics were discussed for *mbC* (and for nonclassical logics in general). The fact that *mbC* lies outside the scope of the traditional algebraic methods has furthered the development of new kinds of semantics. Three paradigms were briefly surveyed, and their interrelations were dis-

cussed: swap structures, nondeterministic matrices and possible-translations semantics. These paradigms are bounded nondeterministic in nature, in the sense that the result of an input by the semantic procedure can produce more than one single output, but within a previously determined set of values. This strongly suggests that bounded nondeterminism is a suitable approach when studying *mbC* as well as other logics of the same kind. Besides characterizing *mbC*, the formal properties of such semantics deserve future studies. In particular, the algebraic properties of swap structures constitute an instigating topic of research.

## References

- 15.1 N.C.A. da Costa, S. French: *Science and Partial Truth: A Unitary Approach to Models and Scientific Reasoning* (Oxford Univ. Press, Oxford 2003)
- 15.2 T. Nickles: From Copernicus to Ptolemy: Inconsistency and method. In: *Inconsistency in Science*, ed. by J. Meheus (Springer, Dordrecht 2002)
- 15.3 N.C.A. da Costa: *Sistemas formais inconsistentes (Inconsistent formal systems)* (Editora UFPR, Curitiba 1993), in Portuguese
- 15.4 N.C.A. da Costa: The philosophical import of paraconsistent logic, *J. Non-Class. Log.* **1**, 1–19 (1982)
- 15.5 A. Einstein: *Relativity: The Special and General Theory* (Emporium Books, Belrose 2013)
- 15.6 R.P. Feynman, R.B. Leighton, M. Sands: *The Feynman Lectures on Physics*, Vol. 1 (Basic Books, New York 2010)
- 15.7 W.A. Carnielli, M.E. Coniglio, J. Marcos: Logics of formal inconsistency. In: *Handbook of Philosophical Logic*, Vol. 14, ed. by D.M. Gabbay, F. Guenther (Springer, Dordrecht 2007)
- 15.8 K. Gödel: Zum intuitionistischen Aussagenkalkül, *Anz. Akad. Wiss. Wien* **69**, 65–66 (1932)
- 15.9 N.C.A. da Costa, E.H. Alves: A semantical analysis of the calculi *Cn*, *Notre Dame J. Formal Log.* **18**, 621–630 (1977)
- 15.10 C.S. Peirce: *Collected Papers* (Harvard Univ. Press, Cambridge 1931)
- 15.11 P. Mancosu: Mathematical explanation: Problems and prospects, *Topoi* **20**(1), 97–117 (2001)
- 15.12 O. Arieli, M. Denecker, B. van Nuffelen, M. Bruynooghe: Coherent integration of databases by abductive logic programming, *J. Artif. Intell. Res.* **21**, 245–286 (2004)
- 15.13 W.A. Carnielli, S. De Amo, J. Marcos: A logical framework for integrating inconsistent information in multiple databases, *Lect. Notes Comput. Sci.* **2284**, 67–84 (2002)
- 15.14 W.A. Carnielli, M.E. Coniglio, R. Podiacki, T. Rodrigues: On the way to a wider model theory: Completeness theorems for first-order logics of formal inconsistency, *Rev. Symb. Log.* **7**(63), 548–578 (2014)
- 15.15 A. Aliseda: Seeking Explanations: Abduction in Logic, Philosophy of Science and Artificial Intelligence, Ph.D. Thesis (Stanford University, Stanford 1997)
- 15.16 M.E. Coniglio: Obtención de respuestas en bases de conocimiento a partir de completaciones, Query procedures in knowledge bases by means of completions, Proc. 21th JAIIO Argent. Symp. Inf. Operat. Res. (SADIO) (1992), in Spanish
- 15.17 W.A. Carnielli, M. Lima-Marques: Reasoning under inconsistent knowledge, *J. Appl. Non-Class. Log.* **2**(1), 49–79 (1992)
- 15.18 W.A. Carnielli, J. Marcos: Tableau systems for logics of formal inconsistency, Proc. Int. Conf. Artif. Intell. (IC-AI 01), Vol. II (2001) pp. 848–852
- 15.19 C. Caleiro, W. A. Carnielli, M. E. Coniglio, J. Marcos: Dyadic semantics for many-valued logics, Draft <http://sqjg.math.ist.utl.pt/pub/caleiro/03-CCM-dyadic2.pdf>
- 15.20 N.C.A. da Costa, W.A. Carnielli: On paraconsistent deontic logic, *Philosophia* **16**(3/4), 293–305 (1986)
- 15.21 M.E. Coniglio: Logics of deontic inconsistency, *Revista Brasileira de Filosofia* **233**, 162–186 (2009)
- 15.22 M.E. Coniglio, N.M. Peron: A paraconsistentist approach to Chisholm's paradox, *Principia* **13**(3), 299–326 (2009)
- 15.23 J.-Y. Béziau: S5 is a paraconsistent logic and so is first-order classical logic, *Log. Stud.* **9**, 301–309 (2002)
- 15.24 J.-Y. Béziau: Paraconsistent logic from a modal viewpoint, *J. Appl. Log.* **3**, 7–14 (2005)
- 15.25 A.L. de Araújo, E.H. Alves, J.A.D. Guerzoni: Some relations between modal and paraconsistent logic, *J. Non-Class. Log.* **4**(2), 33–44 (1987)
- 15.26 J.P. Sartre: *Existentialism is a Humanism* (Yale Univ. Press, New Haven 2007)
- 15.27 J. Hansen, G. Pigozzi, L. van der Torre: Ten philosophical problems in deontic logic, Dagstuhl Seminar Proceeding 07122 (2007), <http://drops.dagstuhl.de/opus/volltexte/2007/941/>
- 15.28 J. Bueno-Soler: Two semantical approaches to paraconsistent modalities, *Log. Univ.* **4**(1), 137–160 (2010)

- 15.29 J. Bueno-Soler: Multimodalidades anódicas e catódicas: A negação controlada em lógicas multimodais e seu poder expressivo (Anodic and Cathodic Multimodalities: Controlled Negation in Multimodal Logics and Their Expressive Power, Ph.D. Thesis (IFCH–Unicamp, Campinas 2009), in Portuguese
- 15.30 J. Bueno-Soler: Completeness and incompleteness for anodic modal logics, *J. Appl. Non-Class. Log.* **19**, 291–310 (2009)
- 15.31 J. Bueno-Soler: Multimodal incompleteness under weak negations, *Log. Univ.* **7**, 21–31 (2013)
- 15.32 W.J. Blok, D. Pigozzi: *Algebraizable Logics*, *Memoirs of the American Mathematical Society Ser.*, Vol. 77 (396) (American Mathematical Society, Providence 1989)
- 15.33 A. Avron: Non-deterministic matrices and modular semantics of rules. In: *Logica Universalis*, ed. by J.-Y. Béziau (Birkhäuser Verlag, Basel 2005) pp. 149–167
- 15.34 M. Fidel: An algebraic study of a propositional system of Nelson. In: *Mathematical Logic. Proceedings of the First Brazilian Conference on Mathematical Logic, Campinas 1977*, *Lecture Notes in Pure and Applied Mathematics Ser.*, Vol. 39, ed. by A.I. Arruda, N.C.A. da Costa, R. Chuaqui (Marcel Dekker, New York 1978) pp. 99–117
- 15.35 D. Vakarelov: Notes on N-lattices and constructive logic with strong negation, *Stud. Log.* **36**(1/2), 109–125 (1977)
- 15.36 M.M. Fidel: *Nuevos enfoques en Lógica Algebraica (New Approaches to Algebraic Logic)*, Ph.D. Thesis (Universidad Nacional del Sur, Bahía Blanca 2003), in Spanish
- 15.37 W.A. Carnielli, M.E. Coniglio: Swap structures for LFIs, *CLE e-Prints* **14**(1), 1–39 (2014)
- 15.38 A. Avron, I. Lev: Canonical propositional Gentzen-type systems, *Lect. Notes Artif. Intell.* **2083**, 529–544 (2001)
- 15.39 W.A. Carnielli: Many-valued logics and plausible reasoning, *Proc. 20th Int. Symp. Multiple-Valued Log.* (The IEEE Computer Society Press, Los Alamitos 1990) pp. 328–335
- 15.40 W.A. Carnielli, M.E. Coniglio: Splitting logics. In: *We Will Show Them! Essays in Honour of Dov Gabbay*, Vol. 1, ed. by S. Artemov, H. Barringer, A.A. Garcez (College Publications, London 2005)

# Model-Based Reasoning

## Part D

### Part D Model-Based Reasoning in Science and the History of Science

Ed. by Nora Alejandrina Schwartz

- 16 Metaphor and Model-Based Reasoning in Mathematical Physics**  
Ryan D. Tweney, Bowling Green, USA
- 17 Nancy Nersessian's Cognitive-Historical Approach**  
Nora Alejandrina Schwartz, Buenos Aires, Argentina
- 18 Physically Similar Systems – A History of the Concept**  
Susan G. Sterrett, Wichita, USA
- 19 Hypothetical Models in Social Science**  
Alessandra Basso, Helsinki, Finland  
Chiara Lisciandra, Groningen, The Netherlands  
Caterina Marchionni, Helsinki, Finland
- 20 Model-Based Diagnosis**  
Antoni Ligęza, Krakow, Poland  
Bartłomiej Górny, Krakow, Poland
- 21 Thought Experiments in Model-Based Reasoning**  
Margherita Arcangeli, Berlin, Germany

The chapters contained in Part D, *Model-Based Reasoning in Science and the History of Science*, provide conceptual tools that allow us to understand model-based-reasoning in current science and the history of science and akin notions such as similar system-based inferences. On the one hand, they give analytic frames – cognitive, historical, and methodological ones – (Chaps. 16–19) to help us understand that kind of scientific reasoning in any domain or across the social sciences. On the other hand, they address specific forms of model-based reasoning: a paradigm of model-based diagnostic reasoning, supported by a formal theory of diagnostic reasoning (Chap. 20) and a review of the debate on thought experiment (Chap. 21).

Chapter 16 proposes an interpretative frame based on cognitive science to understand the effects mathematical representations may have on scientists' model-based reasoning, specifically on that of physicists. This frame is constituted by the concept of *model-based reasoning*, the concept of *metaphorical processes* founded on embodied cognition and on more basic conceptual spaces, and the concept of *long-term working memory*. Ryan Tweney defends this proposal on the basis of three claims: (i) that mathematical representations used in physics exemplify model-based-reasoning, (b) that the working of such models depends on acquired metaphors and conceptual blend, and (c) that the acquisition of these metaphorical grounds can be explained by developing long-term working memory. He illustrates the argument that the above-mentioned cognitive schemes can be understood as the basis of mathematical representations in physics, developing the analysis of one part of J.C. Maxwell's field theory of electromagnetism.

Chapter 17 seeks to clarify and underline the merits of the cognitive-historical approach elaborated by Nancy Nersessian, an environmental perspective within cognitive studies of science, with which a central aspect of model-based reasoning in science is treated: a process to solve representational problems generating historically creative ideas. In order to achieve this goal, Nora A. Schwartz introduces the main problems and solutions provided using the cognitive-historical method. Accordingly, the chapter consists of three parts: questions about the creation of the scientific concepts, epistemic virtues of the cognitive-historical analysis, and a hypothesis about the creation of scientific concepts. The first one is focused on the nature of cognitive processes implied in the creation of ideas and the search of an account for their effectiveness in achieving successful results. The second one exhibits the epistemic virtues of the historical and cognitive dimensions of the method. The last one introduces the dynamic hypothesis of cog-

nitive processes implicated in scientific change and also develops the argument that model-based reasoning is effective to create new candidate representations because it facilitates the changes of *constraints*.

Chapter 18 helps to improve the understanding and appreciation of the notion of physically similar systems in the philosophy of science. Susan Sterrett characterizes this concept as it is understood currently, based on the article by Edgar Buckingham *On Physically Similar Systems: Illustrations of the Use of Dimensional Equations*. Then she draws a path from the earliest precursors of the concept in the Renaissance to its plain articulation in the twentieth century, bringing out, on the one hand, the key ideas of function, which was developed in the eighteenth century, and, on the other hand, the idea of a coherent system of units, which was developed in the late nineteenth century. Also, Sterrett discusses the role that the notion has in reasoning and drawing inferences: The concept of similar systems has been useful in developing methods to draw inferences about values of specific quantities in a system, based on observations in other systems. Sterrett emphasizes that the success of this approach in physical chemistry promoted the extension of a similar system approach to electromagnetic theory and gas kinetic theory.

Chapter 19 focuses on the distinctive features of the method of hypothetical modeling in social sciences by treating it as one style of reasoning: abstract or theoretical model-based reasoning. With this goal, Caterina Marchionni, Alessandra Basso, and Chiara Lisciandra compare this method with other styles of reasoning employed in social science: experiments and computer simulations. Differences between hypothetical modeling and experiments are found, and the consequences they have for making inferences about the world are explored. Considered closely, computer simulations are also viewed as a different style of reasoning from that of analytical models, in that they are particularly apt for dealing with complex systems. Also, the legitimacy of hypothetical modeling as a way of learning about social scientific phenomena is examined. From recognizing the little philosophical agreement on the issue, the discussion is rebuilt by organizing the different perspectives around the function of models that is taken as primary.

Chapter 20 presents *model-based-diagnostic reasoning*, understood as a paradigm of diagnostic inference aimed to give rational explanations of some faulty behavior of the system under discussion. The main idea of this paradigm is the comparison between the behavior of the observed system and the one which can be predicted using knowledge about the system model.

---

The model-based diagnostic approach is placed within knowledge engineering methods from the artificial intelligence domain and is based on the formal theory of diagnostic reasoning by *R. Reiter*. *Antoni Ligeza* and *Bartłomiej Górny* illustrate the method in detail with applications; particularly they mention the dynamic system of three tanks.

**Chapter 21** is structured in five parts. In the first one, *Margherita Archangeli* introduces in a historical context a sample of examples of thought experimentation with the purpose of giving a precise idea of the issues under discussion. In the second part, she refers to the more relevant steps in the history of thought experiments: the beginning, when the term was coined, and the two phases into which the debate can be divided, the *classic* one and the *contemporary* one. In the last three

sections, *Archangeli* tackles the following questions: what is a thought experiment? What is the function of thought experiments? How do they achieve their function? She reviews what has been said about thought experiment definitions placed within an experimental domain and within a theoretical domain and, finally, she refers to the main features that should help us to identify thought experiments. Further, she deals with the central epistemological questions treated in the literature on thought experiments: what kind of knowledge do thought experiments produce? To what extent are thought experiments a reliable source of information? What role do thought experiments play in the processes of rational choice? Finally, she reviews what has been said about the cognitive underpinnings of thought experimentation and focuses on the role of imagination in thought experiments.

# 16. Metaphor and Model-Based Reasoning in Mathematical Physics

Ryan D. Tweney

The role of model-based reasoning in experimental and theoretical scientific thinking has been extensively studied. However, little work has been done on the role of mathematical representations in such thinking. This chapter will describe how the nature of mathematical expressions in physics can be analyzed using an extension of the metaphoric analysis of mathematics. In *Where Mathematics Comes From*, Lakoff and Núñez argued that embodied metaphors underlie basic mathematical ideas (e.g., the concept of *number* is based on the embodied operations of *collecting objects*), with more complex expressions developed via conceptual blends from simpler expressions (e.g., *addition as combining collections*). In physics, however, the need to represent physical processes and observed entities (including measurements) places different demands on the blending processes. In model-based reasoning, conceptual blends must often be based on immediately available embodiments as well as highly developed mathematical expressions that draw upon expert use of long term working memory. Thus, Faraday's representations of magnetic fields as *lines of force* were modeled by Maxwell as vectors. In this chapter, we compare Faraday's experimental investigation of the magnetic field within a magnet to Maxwell's mathematical treatment of the same problem. Both can be understood by unpacking the metaphoric underpinnings as physical representations. The implications for analogical and model-based reasoning accounts of scientific thinking are discussed.

16.1	<b>Cognitive Tools for Interpretive Understanding</b> .....	343
16.1.1	Model-Based Reasoning .....	343
16.1.2	Metaphoric Processes .....	344
16.1.3	Long Term Working Memory.....	344
16.2	<b>Maxwell's Use of Mathematical Representation</b> .....	345
16.2.1	From Faraday to Maxwell .....	345
16.2.2	Faraday: Magnetic Lines Within a Magnet.....	346
16.2.3	Maxwell: Magnetic Lines Within a Magnet.....	347
16.3	<b>Unpacking the Model-Based Reasoning</b> .....	348
16.4	<b>Cognition and Metaphor in Mathematical Physics</b> .....	350
16.5	<b>Conclusions</b> .....	351
	<b>References</b> .....	352

Mathematics is central in science; it is frequently used as the basis for calculation, as a means of derivation of new expressions, and – the focus of this chapter – as a means of *representation*. Oddly, however, there are few attempts to deal with the power of mathematics as a representational medium in science, in spite of

extensive work on the psychological and cognitive underpinnings of scientific thought in general [16.1].

To clarify what is meant by representation, consider the following. *Isaac Newton*, in his *Principia Mathematica* [16.2] formulated a law of universal gravitation which is usually today expressed with the following

equation

$$F = G \frac{m_1 m_2}{r^2}. \quad (16.1)$$

The equation gives the force  $F$  between two bodies of mass  $m_1$  and  $m_2$ , separated by the distance  $r$ ;  $G$  is the universal gravitational constant. About a century after Newton, *Lagrange* [16.3] showed that there was an alternate way to represent the dynamics among a system of bodies

$$L = T - U. \quad (16.2)$$

Here  $L$ , now known as the Lagrangian, is given as the difference between the kinetic energy  $T$  and the potential energy  $U$ , quantities which can be defined for every point in the space between two (or more) objects. Lagrange showed that his formulation could solve all the same problems as Newton's equations and in many cases was easier to use; it possessed calculation advantages and led to new derivational possibilities.

In fact, however, the two expressions, Newton's and Lagrange's, are fundamentally different in the way they represent the same physical reality. Newton's equation is based on an *action at a distance* view; it tells the relation between the masses in terms of the distance between them, but says nothing about the intervening space; the gravitational attraction just *happens* across empty space. Lagrange's equation, however, is defined at every location *between* the masses. Thus, where Newton's equation is nonlocal in character, Lagrange's equation is local. In this sense, it is more compatible with a field-like conception of the gravitational forces. As representations, therefore, the two expressions convey something entirely different about the dynamics of gravitational attractions. Furthermore, given this difference, it is appropriate to ask what effects the differing representations might have on the way in which physicists reason about gravitation. In particular, how might such representational differences affect the model-based reasoning of a physicist?

It is a commonplace to say that different kinds of mathematics are needed to deal with different kinds of physics. Thus, an Aristotelian world view, which is focused upon a world of objects, is associated with Euclidean geometry, an extremely powerful way of dealing with object shape and size. During the seventeenth century, and the emergence of analytic geometry, it became easier to talk about the relations among objects. For example, one could readily determine the intersection between two curved lines or surfaces. The physics that emerged as the result of the Galilean/Newtonian

world view, in turn, drove the development of calculus as a means of determining and describing the motion of objects, and, more generally, of changing quantities in general. The eighteenth century saw extensive development of the tools of calculus (Lagrange's work being just one example), a development that continued in the nineteenth century [16.4].

But the nineteenth century brought a new kind of physics on the scene, one based on field theory, as exemplified by the work of Michael Faraday and James Clerk Maxwell. New ways of thinking required new kinds of mathematics, the mathematics of space filling, vectorial, expressions. Field theories require attention to the entire space surrounding objects (and even, as we shall see, inside the objects) and gave new meaning to Lagrange's approach. For field theories, Euclidean formulations are inadequate, and even analytic geometric methods can be tedious and unilluminating. Developments in the calculus during the eighteenth century overcame these limitations to a larger extent; in particular, as partial differential equations became available, it became easier to represent phenomena that were not tied to the object-centered world of objects and motion [16.5].

Note that in speaking of the representational power of mathematical expressions, the use of mathematics in science is really being talked about, rather than of mathematics as such. Both Newton and Maxwell, for example, were powerful mathematical thinkers, but they were also finely tuned to the representational use of mathematical expressions. For *Newton*, this centered on a geometric mathematics; for modern readers, his *Principia Mathematica* [16.2] is difficult to read (in part) because we have lost the feel for how his physics can be represented in this way. Translated into the (today) more familiar Leibnizean notation for the calculus; however, the underlying representations become more transparent. For Maxwell, the notation is more familiar (to those who have had a physics course in electricity and magnetism). While some translation is still needed [16.6], Maxwell's field-like use of integral and differential vector expressions, as exemplified in his *Treatise on Electricity and Magnetism* of 1873 [16.7] is still important.

Newton's mechanics, especially as it was understood after Newton, assumed that the fundamental principle of motion depended upon forces that acted at a distance. Two masses attract each other because the gravitational force centered on each produces the motion. Throughout the eighteenth and most of the nineteenth centuries, similar action at a distance forces were presumed to be responsible for electric and magnetic actions. Just as gravitational force obeys an *inverse square law* (as in (16.1)), so also did the attractive or



repulsive force between two magnets or two electric charges. The action at a distance account was challenged by Michael Faraday, who instead argued that electric and magnetic forces depended upon *lines of force*; the first true field theory in physics. By the end of his life, Faraday believed he had demonstrated the physical reality of the lines as immaterial but real centers of *power* [16.8].

Faraday was mainly well-known for his many experimental researches and discoveries, but his theoretical account had almost no adherents – except the young Maxwell. For Maxwell, Faraday’s account was a seminal one, and he set about to translate it into mathematical expressions. Eventually, he was able to show that the prevalent action at a distance theories of electromagnetic effects were less tenable than a true field theory (although this account also was slow to gain acceptance, as *Hunt* has shown [16.9]).

In this chapter, using a part of Maxwell’s account, it will be shown how cognitive science can provide an analytic framework for an understanding of the role of mathematics in physics. Maxwell’s reformulation of classical physical ideas can thus be understood in cognitive terms, using recent formulations of model-based reasoning in science, and recent analyses of the under-

lying metaphoric bases of mathematics. The argument is based on three claims:

1. That mathematical representations can serve in model-based reasoning, and
2. That an understanding of how they are used requires attention to the embodied metaphoric understandings of the expressions. The metaphoric bases are in turn
3. Dependent upon automated cognitive processes related to the employment of long term working memory (LTWM).

In this way, the external representation in the form of a mathematical expression is coordinated with an internal representation.

One terminological point is needed. In distinguishing between metaphors and analogies, an unconventional division between two terms often seen as interchangeable is used. In the present usage (following [16.10]), *Metaphor* is used to signal a taken-for-granted, tacit, comparison. I use *analogy* to signal a comparison between a source and a target that must be explicitly argued. In the particular case of Maxwell’s physics, there have been many studies of his use of analogy in this sense, but little about his use of metaphor.

## 16.1 Cognitive Tools for Interpretive Understanding

Each of the three cognitive claims has a somewhat different epistemic grounding. Here the first claim has been taken to be given. That is, abundant research and scholarship, some reflected in the other chapters of this volume, have shown that model-based reasoning is ubiquitous in science – this will not be argued as such in this chapter. On the other hand, the embodied metaphoric claim is an extension of a current approach, one which is not without controversy. While we will not review the pros and cons, nor claim sides, we do hope to convince the reader that the use of a metaphoric analysis of the tacit, taken-for-granted, aspects of mathematical physics can illuminate the representational power of mathematics. The embodiment of metaphor will be assumed here, and is important to the notion of model-based reasoning as a species of abduction [16.10–14]. Note also that *Simpson* [16.15, 16], while emphasizing the rhetoric of Maxwell’s *Treatise*, is advancing a similar argument. Finally, we use recent research on expertise and long term working memory as an explanatory tool, a way of justifying the metaphoric analysis and of suggesting ways in which such model-based reasoning can be learned and acquired as a working tool.

### 16.1.1 Model-Based Reasoning

Model-based reasoning rests on the claim that scientific thinking is largely a matter of the development of mental models of a richly varied sort; models that are involved in constructive and manipulative activities and that draw upon information in various formats, including linguistic and imagistic simulations, as well as external representations [16.14]. The traditional cognitive views of mental models [16.17], which centered on linguistic and propositional reasoning, have been extended in their application to scientific thinking. Thus, *Nersessian* [16.14], drawing partly on Maxwell’s use of analogy, described a model-based reasoning process which included the mental simulation of complex physical systems (see also [16.18]). *Clement* [16.19] emphasized the recursive character of model-based reasoning, arguing for a *Generate-Evaluate-Modify* cycle. As with *Nersessian*’s approach, *Clement* emphasized the way in which scientific models are successively modified and tested. By studying both scientists and advanced college students in real time, *Clement* was able to track these processes from their initial formulations to the final, tested and justified, model.

### 16.1.2 Metaphoric Processes

In recent years, linguists and cognitive scientists have explored the metaphoric underpinnings of language. The claim is that common expressions like *falling in love*, or *building an argument* are actually based on the specific metaphors of physical falling or of building construction. This has been argued as a way to connect the abstractness of language with sensori-motor cognition, and of *embodied* cognition in general [16.20].

Lakoff and Núñez [16.21] argued that even the most abstract of mathematical formulations are also grounded in basic cognitive embodiments via the use of metaphor. For example, the arithmetical operation of addition is related to the elementary cognitive operations of collecting objects. Thus, *Object collection* as source is mapped onto *Arithmetic* as target. *Collections of objects of the same size* are mapped onto *Numbers*, *Putting collections together* onto *Addition*, and *Taking a smaller collection from a larger one* onto *Subtraction* [16.21, p. 55]. Arithmetic itself can then become the source for further extensions to new target domains. *Grounding metaphors* according to Lakoff and Núñez are linked directly to sensori-motor experience (as in the examples), and these are then the source for further conceptual metaphors.

Turner [16.22] has argued that *conceptual integration*, the *blending* of disparate conceptual spaces is a basic cognitive operation that underlies the emergence of new meaning. Thus, in the metaphor, *The surgeon is a butcher*, the spaces corresponding to the source and target of the metaphor each contribute some meanings to the blend, but the emergent meaning of the whole is something not characteristic of either of the parent spaces. Turner has shown how non-Euclidean geometry can be interpreted as a conceptual blending from Euclidean geometry [16.22, Appendix C, pp. 163–168]. In this fashion, as Lakoff and Núñez also argue, the seemingly abstract spaces of mathematics can be unwrapped by showing their origins in successively more basic conceptual spaces. The approach is general; for example, Núñez [16.23] has used it to interpret the historical case of the development of transfinite cardinal numbers by Georg Cantor.

The conceptual theory of metaphor and its role in science has been the subject of some controversy (see, e.g., the critiques by Murphy [16.24, 25] and Weiskopf [16.26] and the reply by Gibbs Jr. [16.27]). Still, for present purposes, in which the approach is used to structure an interpretive framework, the outcome of the controversy is not directly relevant. For the present analysis, what counts is the ability of the approach to provide a tool for the untangling of what is usually implicit in mathematical physics. Note also

that my approach differs from accounts that regard metaphor as a somewhat loose use of similarity, while analogy has been regarded as founded on more severe constraints; thus Gentner and Jeziorski [16.28] adopt such a view. By contrast, we are using the two terms in unconventional fashion, with metaphor referring to implicit comparisons and analogy to those drawn explicitly.

### 16.1.3 Long Term Working Memory

Cognitive scientists have long distinguished between (1) short term memory (STM), which holds a limited amount of new information for a brief time, (2) long term memory (LTM), a larger, more permanent, store, and (3) WM which holds material recently retrieved from LTM as needed in a specific task. Ericsson and Kintsch [16.29] extended the concept of WM by noting that, among experts, specific kinds of processes seemed to be taking place when domain-specific material was retrieved. Referring to this as (4) LTWM, Ericsson and Kintsch suggested that many of the results of expertise can be explained by the emergence of LTWM. In particular, rather than relying upon specific retrieval cues, experts have acquired *structured* retrieval mechanisms to bring domain-relevant material and skills into WM. In the case of mathematical reasoning in science, a differential equation, say, can be thought of as entraining a series of other components of the knowledge of calculus into LTWM.

Ericsson and Kintsch showed that expert readers (but not inexperienced readers) can keep the thread of a book's argument in mind long after the contents of ordinary (short term) WM have been replaced by new information. In effect, LTM remains immediately accessible. Experts thus have a specific set of retrieval structures that make this possible. The relevant skill is more than simply possession of a set of retrieval cues. The expert retrieval structures also imply an anticipatory element that flags what might be relevant in the near or far distant future. Such structures are domain specific and develop only after extensive deliberate practice. In the case of expert reading, the larger gist of text remains available across long stretches of text. The same is true for differential equations in physics [16.30]; in addition to a specific cue, the skills required to use and interpret such equations become automatic.

For Ericsson and Kintsch, LTWM is acquired as an aspect of the acquisition of expertise and comes about via the extended deliberate practice characteristic of the highest levels of expertise. Thus, physics professors perform differently than physics graduate students on problems where both have the same specific con-

tent knowledge, as *Chi et al.* [16.31] have shown (see also [16.32]). The professor subjects have developed such LTWM retrieval structures centered around the ba-

sic principles of physics, while the graduate students are still acquiring them and are more dependent upon surface-level cues.

## 16.2 Maxwell's Use of Mathematical Representation

In the previous section, three cognitive concepts have been outlined; these will serve as the interpretive framework for the following discussion. We will argue

1. That the mathematical representations used in physics exemplify model-based reasoning
2. That the functioning of such models depends upon acquired metaphors and conceptual blends, and
3. That the acquisition of such metaphoric foundations can be explained by the development of LTWM.

To illustrate the argument, we will develop an analysis of one part of Maxwell's field theory of electromagnetism, a *mini* case study. To provide context, we will give a brief account of experimental work by Faraday which is directly relevant.

### 16.2.1 From Faraday to Maxwell

In a conventional view that finds its way into many textbooks, Michael Faraday (1791–1867) was one of the greatest experimental scientists of the nineteenth century, responsible for a long string of discoveries, most famously in electricity and magnetism. Still, his theoretical ideas were couched in a nonmathematical language that did not, by and large, appeal to his contemporaries. By contrast, as the conventional view has it, James Clerk Maxwell (1831–1879), was one of the greatest of the mathematical physicists of the century. His *translation* of Faraday's theory into mathematical expressions and his subsequent extension of those theories was the ultimate triumph of classical physics. A good brief introduction to Faraday's work is [16.33]. For Maxwell, a good beginning is [16.34].

The conventional view, while broadly correct, misses the nuances of the relation between Faraday's and Maxwell's theory. In particular, Maxwell saw in Faraday an intuitive mathematician of the highest order [16.7, Vol. 1, p. ix]:

“As I proceeded with the study of Faraday, I perceived that his method of conceiving the phenomena was also a mathematical one, though not exhibited in the conventional form of mathematical symbols.”

In the case described below, this will become more clear.

Across thousands of experiments, Faraday developed by 1850 a coherent theory of electric and magnetic fields and the relation between the two [16.8, 35, 36]. Centering on the notion of *lines of force*, which he conceptualized as space-filling immaterial entities possessing dynamic properties, he argued that these were physically real and that his experiments had proved their existence and determined many of their properties. Faraday acknowledged the incompleteness of the theory, in part because it was not possible to determine the velocity with which such fields moved. And, while he had shown by experiment [16.37] a possible relation between electromagnetic fields and light (in the form of a rotation of the direction of polarization of light when traversing a dense transparent glass subjected to a strong magnetic field), he could only speculate on the physical nature of the relationship.

*Thomson* (later Lord Kelvin) [16.38, 39] was the first to attempt a mathematical treatment of Faraday's lines of force. Thomson showed that there was an analogy between the equations describing the distribution of electric and magnetic force and the equations describing the distribution of heat within a solid. In developing the analogy, *Thomson* took no position on the reality of the lines of force, although he later claimed that the equations constituted “a full theory of the characteristics of the lines of force” [16.38, p. 1, footnote added in 1854].

*Maxwell* began his account of Faraday's theory in a series of three papers in 1855–1856, 1861–1862, and 1864, and summarized the final state of his theory in the 1873 *Treatise on Electricity and Magnetism* [16.7]. The development across the three early papers has been extensively analyzed (see especially [16.40, 41]). In the course of the three papers, Maxwell did in fact *translate* Faraday's theory into mathematical form (as the conventional view has it), but there were significant changes along the way. Beginning, like Thomson, with an analogy, Maxwell considered the lines of force in Faraday's theory as if they were tubes carrying an incompressible fluid, then developed a mechanical model (based on vortex wheels in a mechanical ether, again, as an analogy), and finally re-expressed Faraday's notion of *force* into a new form, one based on a dynamical theory with *energy* as the focus; this last view was then fully developed in the *Treatise*. Maxwell's famous

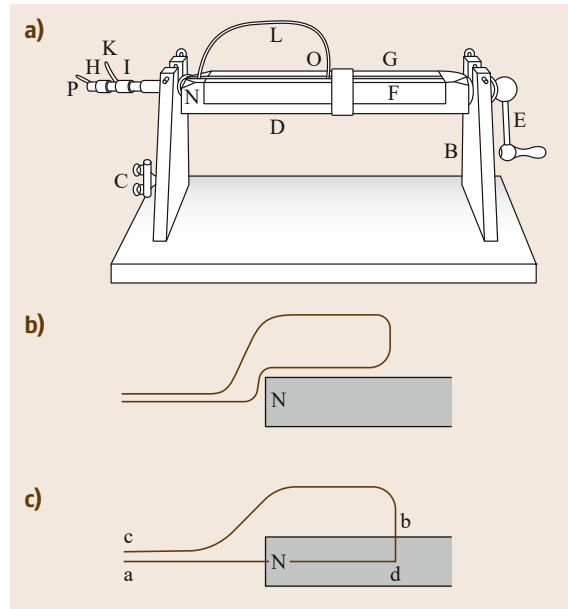
derivation, suggesting that light was an electromagnetic manifestation, appeared initially in the second paper, was re-expressed in the third paper, and finalized at the end of the *Treatise*.

Maxwell's *Treatise* is a complex work with multiple goals. Conceived as a textbook, it includes much material on the fundamental empirical facts of electricity and magnetism, accounts of experiments and measuring devices, and a *dialectical* development of the final theory [16.6, 15, 16]. In its modern form, Maxwell's final account is summarized as *Maxwell's equations*, four vector equations that represent the electric and magnetic fields and the relation between the two. I have outlined how the four equations can be tied metaphorically to primitive embodied notions of stress and strain [16.10]. Here, I compare one aspect of Maxwell's treatment of magnetism to a parallel case examined experimentally by Faraday. This, in turn, will allow an account of how the three cognitive schemas outlined in the previous section can be understood as the bases of mathematical representations in physics.

### 16.2.2 Faraday: Magnetic Lines Within a Magnet

The year 1846 was a crucial one for the development of Faraday's theory of magnetism. In that year, he published three papers on the nature of magnetic interactions, first with light [16.37], then with matter (a brief account is in [16.36] and a more thorough account in [16.8]). Confirming his belief that magnetic lines of force extended through all of space, even penetrating into material bodies, he argued that lines of force were *conducted* within the substance of material bodies, thus establishing that diamagnetic substances (such as bismuth or glass), as well as paramagnetic substances (such as iron) were subject to magnetic influence. Further, by showing the rotation of a polarized light beam in a magnetic field, he was able to argue that magnetic lines of force were perhaps implicated in the nature of light.

Still, he needed to show that the lines of magnetic force could be observed even within the substance of a magnet and that they were closed curves [16.42]. To do this, he conducted an interesting series of experiments in which two long bar magnets of equal strength were placed side by side, with a small gap between (thus acting as a single, thicker, magnet). These were mounted on a shaft within an apparatus that allowed their rotation (Fig. 16.1). With commutators on the shaft of the rotating apparatus, he could then run a wire alongside or within the slot between the magnets, and the wire could be rotated together with or independent



**Fig. 16.1a-c** Faraday's experiments on the lines of force within a magnet. (a) The apparatus used; F and G are two identical magnets mounted on a shaft with a small gap between. Commutators are shown at H, I, and O. (b) The wire separated from and entirely outside the magnet. (c) The wire run through the inside of the magnet. Segments a-d and b-d can be rotated independently, or together with b-c. No current is produced by the entire loop, or by a-d alone, but current is produced by b-d alone or by c-b alone (in the opposite direction) (after [16.42, pp. 333, 338])

of the magnets (which were equivalent to a single magnet with a slot down the middle). Connecting the ends of the wire to a galvanometer, he was able to detect any induced currents in the wire.

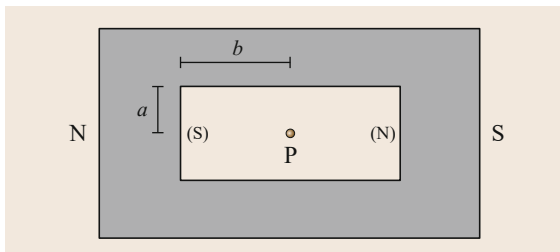
Faraday tried a variety of configurations (I simplify his many arrangements in this description), first rotating the wire and magnet together (no current was produced), then the wire alone or the magnet alone (as in Fig. 16.1b, again no current was produced). He then separated the wire at the point b (Fig. 16.1c), which permitted the segments to be rotated separately while maintaining electrical contact. He found that rotating the magnet with the segment b-d and without the segment b-c *did* produce a current, and rotating the wire segment b-c without the segment b-d *also* produced a current, but in the opposite direction. He argued that the segment of the wire from b to d was cutting all the lines of force when revolved, as did segment b to c. Further, the size of the current produced was the same in both cases. Because the currents were in opposite directions, when the whole wire (a-d-b-c) was revolved and the magnet kept stationary, no current was

observed: Each wire was cutting *all* the lines of force but the generated currents were in opposite directions, thus cancelling each other. This was the result he was after: the lines of magnetic force ran through the magnet, out at one end, curved around through space, and re-entered at the other end of the magnet. Magnetic lines of force are closed curves.

### 16.2.3 Maxwell: Magnetic Lines Within a Magnet

Maxwell's *Treatise* is divided into four parts, with the fourth part developing the final form of his theory of electromagnetism and the third presenting his account of magnetism. The first chapter of the third part considered the magnetic potential at any point outside of a nearby magnet, showing that the force on a *unit magnetic pole* is equal to the gradient of the potential ( $\nabla V$ , where  $V$  is a scalar function), that is, to the rate of change of the potential in the direction of greatest change. In Chapter II, Maxwell considered the forces *within* a magnet. In contrast to Faraday, however, he did not here conduct experiments, nor replicate Faraday's (although they are cited). Instead, he conducted a series of thought experiments.

He began by imagining a cylindrical hollow cavity within a bar magnet (Fig. 16.2). Taking its length as  $2b$  and its radius as  $a$ , he then imagined a unit magnetic pole centered within the cavity. Such a pole is an imaginary object, since magnetic monopoles do not exist (that is, if you break a magnet into two pieces, each piece will have a North and South pole, breaking them again, each piece will have two poles, and so on). Still, *were such a thing to exist*, it is possible to represent the forces it would experience. There are two sources; first the forces due to magnetic induction from the ends of the cavity. Since the field lines are parallel to the walls of the cylinder, the walls play no role, only the circular ends are involved. Second, there are forces due to



**Fig. 16.2** Maxwell's thought experiment: a bar magnet with a cavity inside. The cavity is cylindrical, of length  $2b$  and with faces of radius  $a$ . Note that the polarity of the faces is the reverse of the polarity of the nearest end of the magnet

the potential field within the cavity. That is, there is an overall field because the cavity is within a magnet and a specific field due to the surface distribution of magnetism on the ends of the cylindrical cavity. Note that the forces due to the circular surfaces are of opposite polarity to the ends of the magnet.

Maxwell first considered the field due to the surface distribution on the cylinder ends, claiming that the forces on the monopole are equal and in the same direction (because the monopole will be attracted by one surface and repelled by the other). This force will be

$$R = 4\pi I \left( 1 - \frac{b}{\sqrt{a^2 + b^2}} \right), \quad (16.3)$$

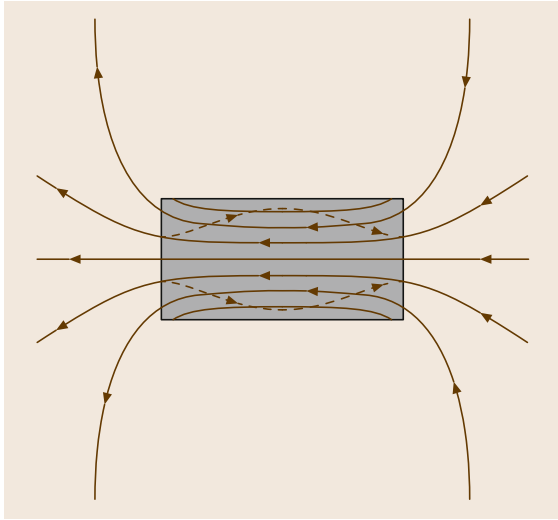
where  $R$  is the force and  $I$  is the intensity of magnetization. Because the dimensions of the cavity are involved, the force is dependent upon the shape of the cavity. Interestingly, Maxwell does not show how this equation is obtained, taking for granted that the reader will know how to do this (while not lengthy, I will not carry this out – see [16.6, the comment on 396.2] and the discussion that follows).

With this in hand, Maxwell now asked the reader to consider two cases. In the first, imagine that  $a$  is very small, that is, shrink the diameter of the cylinder cavity. From (16.3), note that  $R$  will approach 0 as  $a$  approaches 0. In the second case, let the cylinder shrink in length. As  $b$  approaches 0, then  $R$  approaches  $4\pi I$ . This means that, in the first case, a long and thin cylinder, the force will simply be that due to the overall field; it will be the gradient of the potential. Maxwell calls this *magnetic force within the magnet* and symbolized it as a vector,  $\mathbf{H}$  (here using bold-face, to indicate a vector). In the second case, which becomes a flat disk as the cylinder length shrinks, the force is dependent on  $R$  and is compounded of  $4\pi I$  and  $\mathbf{H}$ . He symbolizes this new quantity as  $\mathbf{B}$  and calls it the *magnetic induction*. The two terms are related by a simple equation, via the overall intensity of magnetization,  $I$ , which, written as a vector, is

$$\mathbf{B} = \mathbf{H} + 4\pi \mathbf{I}. \quad (16.4)$$

Note from (16.4) that the distinction between  $\mathbf{B}$  and  $\mathbf{H}$  will hold only within a magnet; in the absence of a surrounding magnet, that is, when  $\mathbf{I} = 0$ , the two are identical (Fig. 16.3).

Maxwell used the relation between  $\mathbf{B}$  and  $\mathbf{H}$  to clarify a paradox in Faraday's notion of lines of force. The paradox arose because the directions of  $\mathbf{B}$  and of  $\mathbf{H}$  differ. That is, the magnetic force due to  $\mathbf{H}$  always goes from the North Pole of the magnet to the South Pole – both inside and outside the magnet! As a result, they



**Fig. 16.3** Showing the lines of force within and without a bar magnet; the North Pole is to the left. Solid lines are lines of induction,  $\mathbf{B}$ ; dashed lines are lines of  $\mathbf{H}$ . On the outside of the magnet, the two fields are identical

meedst head on, as it were, at the South Pole, which then constitutes a termination. But Faraday’s lines of force are continuous closed curves and do not meet; they run from the North Pole to the South Pole when outside the magnet and continue from the South Pole to the North Pole inside the magnet, thus constituting closed curves. The magnetic lines due to  $\mathbf{B}$  have the needed property – they run from north to south outside and from south to north inside. For this reason [16.7, Vol. 2, p. 28]:

“All that Faraday says about lines of force [...] is mathematically true, if understood of the lines [...] of magnetic induction.”

## 16.3 Unpacking the Model-Based Reasoning

How do we use the cognitive framework sketched in the first part of this chapter to understand the case study? Maxwell, like Faraday, used model-based reasoning in the example, as should be clear. Like Faraday, Maxwell described for the reader a series of actively used constructions to make the argument for forces within a magnet. Faraday described actual experiments, inviting the reader to construct a mental model of the apparatus, procedures, and results. Maxwell used a thought experiment in the same way; that is, his reader was asked to construct a mental model of an experiment that could be done only in the mind’s eye and

Thomson [16.43] had considered a problem similar to Maxwell’s, in determining the force on a unit pole placed within a small cavity in a magnet. However, he did not resolve the directional paradox between the directions of what were later called  $\mathbf{B}$  and  $\mathbf{H}$  by Maxwell. Smith and Wise [16.44] describe Thomson’s approach and indicate that he did not fully publish his results.

Maxwell’s clarification of the difference between  $\mathbf{B}$  and  $\mathbf{H}$  was highly consequential. In particular, it allowed him to distinguish between magnetic effects which were mechanical in nature and those which were involved in the production of currents in a nearby conductor. In the case of  $\mathbf{H}$ , one is speaking of the *magnetic force*, and this is purely mechanical and can be manifested by the effect on a compass needle or an iron filing. In the case of  $\mathbf{B}$ , the *magnetic induction*, the force can be manifested as an *electromotive force*, that is, as one producing a current in a conductor. For Maxwell (mathematically), as for Faraday (experimentally), these corresponded to two different ways to detect the presence of magnetic field.

Faraday initially mapped magnetic fields by using a small magnetized needle suspended from a thread; such a needle will orient itself along lines of force. He later used a small loop of wire attached to a sensitive galvanometer. When moved in a field, a current would be generated in the loop, a current detectable by a sensitive galvanometer at a distance. This *moving wire* became his favored method, mapping, in Maxwell’s terms, the lines of induction. In subsequent chapters of the *Treatise*, Maxwell developed the mathematical representation of such mappings in great detail, arguing that the magnetic induction and not the magnetic force is the physically significant quantity.

not in reality. That is not to say that actual experiments were entirely absent in Maxwell’s account, rather, they were presumed to be present in the reader’s knowledge, based, in part, on the previous chapters of the book and in the references to Faraday’s experiments.

Still, it is not the case that we can make a one-to-one mapping between the kinds of knowledge that underlies Faraday’s reasoning and that of Maxwell. This is because Maxwell also had to rely upon a kind of knowledge base not used by Faraday. In particular, Maxwell relied upon the *metaphoric* understandings embedded within the mathematical expressions used. For exam-

ple, consider (16.3) from the previous section

$$R = 4\pi I \left( 1 - \frac{b}{\sqrt{a^2 + b^2}} \right),$$

As noted earlier, Maxwell did not provide a derivation of this result, instead assuming that his readers would be able to recognize it. To show its metaphoric nature, first consider the term  $a^2 + b^2$ . From Fig. 16.2, it is apparent that this is related, via the Pythagorean theorem, to the length of the hypotenuse of the triangle with sides  $a$  and  $b$ . If we take the square root and call this  $r$ , then we can simplify (16.3)

$$R = 4\pi I \left( 1 - \frac{b}{r} \right), \quad (16.5a)$$

This, in turn, becomes

$$R = 4\pi I - 4\pi I \frac{b}{r}. \quad (16.5b)$$

Now suppose that  $a$  shrinks (Maxwell's first case). Then  $b/r$  goes to 1 and  $R$  goes to 0. And if  $b$  shrinks (Maxwell's second case), then  $R$  goes to  $4\pi I$ .

The attentive reader can now see how the metaphoric underpinnings worked in the discussion of this equation. For, in fact, what has been asked of *you* to do is what Maxwell (with, to be sure, more extensive metaphors assumed) asked of his readers! That is, we drew upon your knowledge of the Pythagorean theorem and upon your metaphoric sense of what happens when geometric terms like  $a$  and  $b$  change. Further, how the sense of algebraic equations can be modified, as in going from (16.3) to (16.5a) and (16.5b), was also involved. These did not need to be specifically argued because, as Lakoff and Núñez [16.21] argued, these have been acquired on the basis of long practice – they are conceptual blends with metaphoric groundings. On my account, they are not analogies, because the links between source and target are implicit and assumed to be known among his readers. This is why Maxwell does not explicate (16.3).

However, (16.3) is not yet fully explicated for our purposes. Where does the  $4\pi I$  come from? In the previous section, Maxwell had considered the force on a small magnet due to the distribution of a surface of magnetic *matter* (like the imagined cavity and the magnetic monopole, this is another convenient fiction). That discussion, in turn, relied upon results achieved in the first volume of the *Treatise*, in which he showed that the surface distribution of an electric charge on a conductor exerted a force near to the conductor equal to  $4\pi\sigma$ , where  $\sigma$  is the surface distribution of charge. In the present case,  $I$  is equivalent to the charge in the

earlier case. In particular, both charge and magnetic entities exert force according to an inverse square law, that is, inversely as the square of the distance. Thus,  $4\pi I$ , unlike the other part of (16.3) is an analogy, albeit itself grounded in the mathematics of earlier parts of the book [16.7, Vol. 2, p. 5]:

“Since the expression of the law of force between given quantities of *Magnetism* has exactly the same mathematical form as the law of force between quantities of *Electricity* of equal numerical value, much of the mathematical treatment of magnetism must be similar to that of electricity.”

Maxwell is able to carry over the expression for the magnetic surface density from the equivalent expression for electric surface density: he does not need to repeat the derivation (which is also built on metaphoric grounds and hence can be taken as given), he only needs to have shown the analogy.

We can again obtain an informal understanding by asking where the multiplier  $4\pi$  comes from. Note first that the monopole at point P is subjected to an attractive force from one face of the cylindrical cavity and a repelling force from the other face. Both forces are in the same direction, so any one face is contributing  $2\pi$  to the result. But  $2\pi$  is the circumference of a circle of radius 1. Here, it appears as if Maxwell is relying upon a previous result from the first volume of the *Treatise*, namely Stokes's theorem, which states that the surface integral of a function describing a surface is equal to the line integral of the curve bounding that surface. Explaining this would go beyond the scope of this chapter, but it implies in the case of the circular face of the cavity that the force due to the face can be construed as either based on the density of magnetization of the surface or, equivalently, as based on a circulation around the closed curve (the circle) that bounds it. Thus,  $2\pi$  emerges!

Note that for Maxwell's readers Stokes's theorem would have been assumed knowledge (it is explained in a *preliminary* chapter [16.7, Vol. 1, p. 29]). For the present purpose, however, it is enough to catch some glimpse of how the factor emerges; in the following chapter, Maxwell uses Stokes's theorem to make a more explicitly physical representation. There, he shows that a magnetic *shell* (a surface bounded by a closed curve) can be represented equivalently by an electric current in a conductor that follows the same closed curve.

One final point: Maxwell's *Treatise* is notable in part for its use of vectors as representational entities. In the selection here, these appear as  $\mathbf{H}$ ,  $\mathbf{B}$ , and  $\mathbf{I}$ . We have previously discussed the metaphoric basis of vector representations [16.10]. For the present case, it needs only to be noted that vectors are quantities that represent both magnitude *and* direction. They can be grounded on el-

elementary notions of muscular force and direction, and can then be conceptually blended with other mathematical concepts. Throughout the *Treatise*, Maxwell uses them (and the vector calculus) as part of his overall rep-

resentation of fields (as in Fig. 16.3). The introduction of such vector analysis was an important milestone in mathematical physics generally, one that continues to be used today [16.5, 45].

## 16.4 Cognition and Metaphor in Mathematical Physics

This chapter has presented a sketch of a mode of analysis that has important implications for understanding how mathematical representations have gained such great importance in science. There have been many analyses of the role of analogy in model-based reasoning, even extending to accounts of Maxwell's physics. However, the metaphoric aspect of mathematical representations holds the key to understanding how the tacit knowledge embedded within mathematical expressions can become an active part of model-based reasoning.

Three points were discussed in the introduction of this chapter: That mathematical physics does involve model-based reasoning, that metaphor underlies the representational use of mathematics, and that such metaphoric grounding is tacit and acquired (via LTWM) through the acquisition of expertise. We will discuss each in the reverse order.

Since the role of LTWM in Maxwell's case [16.46] has already been discussed, only brief comment is needed here. Maxwell wrote the *Treatise* partly intending it as a text for the new Tripos exam in Natural Philosophy at Cambridge University (he had been appointed to the newly established professorship of natural philosophy in 1871 [16.47, 48]). Maxwell's own education in science and mathematics (primarily at Edinburgh and Cambridge, but beginning even in his childhood) provided him with an extensive knowledge of the mathematics and physics that he took for granted in the book, and it is likely that he expected his students would have similar knowledge. He was writing the *Treatise* for those with the kind of retrieval structures that are fundamental to expertise. In recasting the case study for my readers, we have also made some assumptions; for example, that the reader would know the formula for the circumference of a circle, have algebraic skills, and know at least something about electricity and magnetism. Access to all of these relies upon a similar LTWM capacity; the cognitive underpinnings for Maxwell's students and my readers were not different in principle.

We have also previously spelled out the role of metaphors in the understanding of mathematical physics, using the modern form of Maxwell's equations [16.10]. For the present case study, we have in-

stead relied more closely on the actual text written by Maxwell. Although closer to Maxwell's argument, much has been left out. Further, the analysis is informal and adapted to my readership, not Maxwell's. In this sense, what we have provided is not an analysis of the actual historical materials, but rather a reconstruction of a *possible world*. It is interesting to note the similarity of this maneuver to that used by Lakatos in *Proofs and refutations*, which used a similar ploy to discuss the nature of discovery in mathematical proof [16.49]. Even so, it should be possible to see the way in which metaphors and conceptual blends play a role in the arguments made by Maxwell. The analytic task here is to work backward from the argument as presented by Maxwell to the underlying structure of the mathematical representations.

While a great deal has been written about Maxwell's use of analogy (especially [16.14, 41]), we believe our analyses are the first attempts to use metaphor and conceptual blends to describe the *tacit* knowledge which Maxwell brought to bear on his arguments (see, for a similar attempt, not rooted in metaphor in the same fashion, [16.50]). It has been argued that, from a cognitive point of view, there is no inherent difference between analogy and metaphor. Indeed, the terms, analogy and metaphor, have had a flexible boundary in much of the writing about their use in science. Thus, for example, much of what Bradie [16.51] has written about metaphor applies equally to analogy. The best-known such argument is due to Gentner [16.52]. Her *structure mapping theory* of analogy and metaphor is based on the processes involved in mapping relations from a source to a target, and there is much evidence to suggest that this correctly captures many of the phenomena. Nersessian has extended this, describing (using Maxwell, in part) how analogies can participate in the creation of new conceptual content in science. In turn, this chapter supplements and extends all of these accounts.

Model-based reasoning is an integral part of all naturalistic accounts of science, and this chapter is no exception. That Maxwell used it in presenting his analysis of the magnetic field within a magnet should be clear, even from the brief segment considered. The



thought experiment he presents is fundamentally anchored in the reader's ability to follow the claims made via the construction of a model and via the implementation of the mathematical representations involved. Note that they lead up to the expression of an *identity*, not an *equation* in the usual sense. That is, (16.4),  $\mathbf{B} = \mathbf{H} + 4\pi\mathbf{I}$ , is presented, not because it has calculational uses but because it shows the reader the relationships among key terms and because, by using vector notation for the first time in the section, it reiterates the directional character of the lines of induction, of force, and of magnetic intensity. It is important because of its representational character.

As noted earlier Faraday represented magnetic *lines of force* experimentally, by constructing apparatus that

enabled the detection of the lines within a magnet. Maxwell achieved the same thing using a thought experiment, a move that allowed him to distinguish between  $\mathbf{H}$  and  $\mathbf{B}$ , thus identifying  $\mathbf{B}$  as the physically significant quantity. The two approaches complement each other in an interesting fashion. Thus, Faraday's science is replete with *hand-eye-mind* representations [16.13]; for him, the lines of force were physically real to the extent that he could observe their effects and manipulate their character. For Maxwell, the observation and manipulation were based, not on experiment directly, but on the expression of a mental model and its extension via the metaphoric underpinnings of the mathematical representations. It, too, had a *hand-eye-mind* character.

## 16.5 Conclusions

Ultimately, then, this is the true fashion in which Faraday and Maxwell can be seen as similar: both were doing science in a style dependent upon a fundamental *embodiment* of the conceptual representations they created. For both, this was, in fact, a conscious goal. When Faraday is seeking the physical reality of his lines of force, he is doing just what Maxwell was doing in identifying the vector  $\mathbf{B}$  as the *physically significant* quantity. That they followed different pathways, that Faraday's was experimental and Maxwell's mathematical, is not, in the end, the most important aspect for an understanding of their creative achievements.

Beyond these two cases, however, there is a more general point to be made. That model-based reasoning is ubiquitous in science should be clear to the reader of this volume. What case studies of the type offered here can provide is a method of discovery of the finer points with which such reasoning is carried out. While not every scientist will resemble either Faraday or Maxwell in the way in which they employ such reasoning, still, the nuances may be quite general across cases. In particular, the importance of distinguishing between those aspects of model-based reasoning that are tacit (and hence unargued – what we have referred to as metaphorical in nature) versus those that are explicit (i. e., analogical in nature) is central to any understanding of scientific thinking. That is why determining the role of expertise and of long-term WM is so helpful in understanding the particularities of a case – any case.

The case studies also reflect a challenge to the common view that science deals with increasingly

*abstract* entities, particularly in situations in which mathematical representation is involved. In fact, however, the presence of arcane symbols and equations do not mean that, *for the scientist*, these are necessarily abstract, however they appear to the uninitiated. In the present chapter, we have tried to show how both Faraday and Maxwell were anchored in quite concrete representations of their respective models of the electromagnetic field within a magnet. Those representations depended for their utility on the highly skilled and easily accessible expertise that each investigator possessed. The presence of such expertise is the necessary cognitive grounding of creative achievement in science.

**Acknowledgments.** Thanks are due especially to Howard Fisher, who has saved me from many errors and is not responsible for remaining ones! I have benefitted greatly from discussions of Maxwell with John Clement, Howard Fisher, Frank James, Nancy Nersessian, and Thomas Simpson. The chapter's ultimate origin stems from discussions with the late David Gooding and with Elke Kurz-Milcke. The proximate origin is a paper given at MBR012 in Sestri Levante, Italy, in June, 2012; I am grateful for the questions and comments of the other participants and to Lorenzo Magnani for his support. Matt Lira and Frank James provided helpful comments on an early draft, for which I am grateful. An earlier version was published in L. Magnani (Ed.): *Model Based Reasoning in Science and Technology* (Springer, Berlin 2014) pp. 395–414.

## References

- 16.1 M.E. Gorman, R.D. Tweney, D.C. Gooding, A.P. Kin-cannon (Eds.): *Scientific and Technological Think-ing* (Lawrence Erlbaum, Mahwah 2005)
- 16.2 I. Newton: *The Principia: Mathematical Principles of Natural Philosophy* (Univ. California Press, Berkeley 1999), transl. by I.B. Cohen, A. Whitman, originally published 1687
- 16.3 J.L. Lagrange: *Analytical Mechanics* (Kluwer, Boston 1997), transl. and ed. by A.C. Boissonnade, V.N. Vagliente, originally published 1788
- 16.4 I. Grattan-Guinness: *The Fontana History of the Mathematical Sciences: The Rainbow of Mathe-matics* (London, Fontana Press 1997)
- 16.5 E. Garber: *The Language of Physics: The Calculus and the Development of Theoretical Physics in Eu-rope, 1750–1914* (Birkhäuser, Boston 1999)
- 16.6 H. Fisher: *Maxwell's Treatise on Electricity and Magnetism: The Central Argument* (Green Lion, Santa Fe 2014)
- 16.7 J.C. Maxwell: *A Treatise on Electricity and Mag-netism*, 3rd edn. (Clarendon Press, Oxford 1891), 2 Volumes, revised by J.J. Thompson, originally pub-lished 1873
- 16.8 D.C. Gooding: Final steps to the field theory: Far-aday's study of magnetic phenomena, *Hist. Stud. Phys. Sci.* **11**, 231–275 (1981)
- 16.9 B.R. Hunt: *The Maxwellians* (Cornell Univ. Press, Ithaca 2005)
- 16.10 R.D. Tweney: On the unreasonable reasonableness of mathematical physics: A cognitive view. In: *Psy-chology of Science: Implicit and Explicit Processes*, ed. by R.W. Proctor, E.J. Capaldi (Oxford Univ. Press, Oxford 2012) pp. 406–435
- 16.11 L. Magnani: *Abduction, Reason, and Sci-ence: Processes of Discovery and Explanation* (Kluwer/Plenum, New York 2001)
- 16.12 J. Cat: Into the 'regions of physical and meta-physical chaos': Maxwell's scientific metaphysics and natural philosophy of action (agency, deter-minacy and necessity from theology, moral phi-losophy and history to mathematics, theory and experiment, *Stud. Hist. Philos. Sci. Part A* **43**, 91–104 (2011)
- 16.13 D. Gooding: *Experiment and the Making of Mean-ing: Human Agency in Scientific Observation and Experiment* (Kluwer, Dordrecht 1990)
- 16.14 N.J. Nersessian: *Creating Scientific Concepts* (MIT Press, Cambridge, MA 2008)
- 16.15 T.K. Simpson: *Figures of Thought: A Literary Ap-preciation of Maxwell's Treatise on Electricity and Magnetism* (Green Lion Press, Santa Fe 2005)
- 16.16 T.K. Simpson: *Maxwell's Mathematical Rhetoric: Rethinking the Treatise on Electricity and Mag-netism* (Green Lion Press, Santa Fe 2010)
- 16.17 P.N. Johnson-Laird: Mental models in cognitive science, *Cogn. Sci.* **4**, 71–115 (1980)
- 16.18 K. Forbus: Reasoning about space and motion. In: *Mental Models*, ed. by D. Gentner, A. Stevens (Lawrence Erlbaum, Hillsdale 1983) pp. 53–74
- 16.19 J. Clement: *Creative Model Construction in Scien-tists and Students: Imagery, Analogy, and Mental Simulation* (Springer, Dordrecht 2008)
- 16.20 G. Lakoff, M. Johnson: *Philosophy in the Flesh: The Embodied Mind and its Challenge to Modern Thought* (Basic Books, New York 1999)
- 16.21 G. Lakoff, R.E. Núñez: *Where Mathematics Comes From: How the Embodied Mind Brings Mathematics into Being* (Basic Books, New York 2000)
- 16.22 M. Turner: *Cognitive Dimensions of Social Science* (Oxford Univ. Press, Oxford 2001)
- 16.23 R.E. Núñez: Creating mathematical infinities: Metaphor, blending, and the beauty of transfinite cardinals, *J. Pragmat.* **37**, 1717–1741 (2005)
- 16.24 G.L. Murphy: On metaphoric representation, *Cog-nit.* **60**, 173–204 (1996)
- 16.25 G.L. Murphy: Reasons to doubt the present evi-dence for metaphoric representation, *Cognit.* **62**, 99–108 (1997)
- 16.26 D.A. Weiskopf: Embodied cognition and linguistic comprehension, *Stud. Hist. Philos. Sci.* **41**, 294–304 (2010)
- 16.27 R.W. Gibbs Jr.: Why many concepts are metaphori-cal, *Cognit.* **61**, 309–319 (1996)
- 16.28 D. Gentner, M. Jeziorski: The shift from metaphor to analogy in Western science. In: *Metaphor and Thought*, ed. by A. Ortony (Cambridge Univ. Press, Cambridge 1993) pp. 447–480
- 16.29 K.A. Ericsson, W. Kintsch: Long-term working memory, *Psychol. Rev.* **102**, 211–245 (1995)
- 16.30 E. Kurz-Milcke: The authority of representations. In: *Experts in Science and Society*, ed. by E. Kurz-Milcke, G. Gigerenzer (Kluwer/Plenum, New York 2004) pp. 281–302
- 16.31 M.T.H. Chi, P.J. Feltovich, R. Glaser: Categorization and representation of physics problems by experts and novices, *Cogn. Sci.* **5**, 121–152 (1981)
- 16.32 J.H. Larkin, J. McDermott, D.P. Simon, H.A. Simon: Models of competence in solving physics problems, *Cogn. Sci.* **4**, 317–345 (1980)
- 16.33 F.A.J.L. James: *Michael Faraday: A Very Short Intro-duction* (Oxford Univ. Press, Oxford 2010)
- 16.34 C.W.F. Everitt: *James Clerk Maxwell: Physicist and Natural Philosopher* (Charles Scribner's Sons, New York 1975)
- 16.35 M. Faraday: On the Physical Character of the Lines of Magnetic Force. In: *Experimental Researches in Electricity*, Vol. 3, ed. by M. Faraday (Taylor Francis, London 1855) pp. 407–437, first published 1852
- 16.36 R.D. Tweney: Inventing the field: Michael Faraday and the creative 'engineering' of electromagnetic field theory. In: *Inventive minds: Creativity in Tech-nology*, ed. by R.J. Weber, D.N. Perkins (Oxford Univ. Press, Oxford 1992) pp. 31–47
- 16.37 M. Faraday: Experimental researches in electricity, Nineteenth series. On the magnetization of light and the illumination of magnetic lines of force. In: *Experimental Researches in Electricity*, Vol. 3, ed. by M. Faraday (Taylor Francis, London 1855) pp. 1–26, (originally published 1846)

- 16.38 W. Thomson (Lord Kelvin): On the uniform motion of heat in homogeneous solid bodies, and its connexion with the mathematical theory of electricity. In: *Reprint of Papers on Electrostatics and Magnetism*, ed. by Sir W. Thomson (Macmillan Co., London 1872), pp. 1–14 (originally published 1842)
- 16.39 W. Thomson (Lord Kelvin): On the mathematical theory of electricity in equilibrium. I. On the elementary laws of statical electricity. In: *Reprint of Papers on Electrostatics and Magnetism*, ed. by Sir W. Thomson (Macmillan Co., London 1872) pp. 15–37 (originally published 1845)
- 16.40 N. Nersessian: *Faraday to Einstein: Constructing Meaning in Scientific Theories* (Nijhoff, Dordrecht 1984)
- 16.41 D.M. Siegel: *Innovation in Maxwell's Electromagnetic Theory: Molecular Vortices, Displacement Current, and Light* (Cambridge Univ. Press, Cambridge 1991)
- 16.42 M. Faraday: Experimental researches in electricity, Twenty-eighth series. On lines of magnetic force: Their definite character; and their distribution within a magnet and through space. In: *Experimental Researches in Electricity*, Vol. 3, ed. by M. Faraday (Taylor Francis, London 1855) pp. 328–370, originally published 1851)
- 16.43 W. Thomson (Lord Kelvin): A mathematical theory of magnetism. In: *Reprint of Papers on Electrostatics and Magnetism*, ed. by Sir W. Thomson (Macmillan Co., London 1872) pp. 340–425 (originally published 1849)
- 16.44 C. Smith, M.N. Wise: *Energy and Empire: A Biographical Study of Lord Kelvin* (Cambridge Univ. Press, Cambridge 1989)
- 16.45 M.J. Crowe: *A History of Vector Analysis* (Univ. Notre Dame Press, South Bend 1967)
- 16.46 R.D. Tweney: Representing the electromagnetic field: How Maxwell's mathematics empowered Faraday's field theory, *Sci. Educ.* **20**(7/8), 687–700 (2011)
- 16.47 P.M. Harman: *The Natural Philosophy of James Clerk Maxwell* (Cambridge Univ. Press, Cambridge 1998)
- 16.48 A. Warwick: *Masters of Theory: Cambridge and the Rise of Mathematical Physics* (Univ. Chicago Press, Chicago 2003)
- 16.49 I. Lakatos: *Proofs and Refutations* (Cambridge Univ. Press, Cambridge 1976), (originally published 1963–1964)
- 16.50 J. Cat: *On understanding: Maxwell on the methods of illustration and scientific metaphor*. *Stud. Hist. Philos. Modern Phys* (32, 395–441 2001)
- 16.51 M. Bradie: Models and metaphors in science: The metaphorical turn, *Protosociol.* **12**, 305–318 (1998)
- 16.52 D. Gentner, B. Bowdle: Metaphor as structure–mapping. In: *The Cambridge Handbook of Metaphor and Thought*, ed. by R.W. Gibbs Jr. (Cambridge Univ. Press, Cambridge 2008) pp. 109–128

## 17. Nancy Nersessian's Cognitive–Historical Approach

Nora Alejandrina Schwartz

Nancy Nersessian raises questions about the creation of scientific concepts and proposes answers to them based on the cognitive–historical approach. These problems are mainly about the nature of the cognitive processes involved in the generation of ideas fundamentally new in human history and the efficacy of those mechanisms in achieving successful results. In this chapter, I intend to show the epistemic virtues that make this method a useful tool for establishing the *dynamic hypothesis* about the creation of knowledge in science. I also point out that, compared to other methods of cognitive studies on the creation of scientific knowledge – ethnography, in vivo observation, and laboratory experiments – the cognitive–historical approach turns out to be primary. I analyze Nersessian's idea that scientists often employ model-based reasoning, in an iterative way, in order to solve representational problems in the target domain. Additionally, I examine her claim that model-based reasoning facilitates the conceptual change. This hypothesis involves a representation of concepts illustrated by the *dynamic frames* theory about concepts.

17.1	<b>Questions About the Creation of Scientific Concepts</b> .....	356
17.1.1	The Problem of Conceptual Change .....	356
17.1.2	The Naturalistic Approach to Science: Revision of the Problem .....	357
17.1.3	The Naturalistic Recasting.....	357
17.2	<b>The Epistemic Virtues of Cognitive Historical Analysis</b> .....	359
17.2.1	The Cognitive–Historical Approach .....	359
17.2.2	Epistemic Virtues and Dimensions of this Approach.....	360
17.2.3	Cognitive Methods to Investigate Conceptual Innovation.....	362
17.3	<b>Hypothesis About the Creation of Scientific Concepts</b> .....	363
17.3.1	Dynamic Hypothesis .....	364
17.3.2	The Power of Model-Based Reasoning.	367
17.4	<b>Conclusions</b> .....	373
	<b>References</b> .....	373

Nancy Nersessian has studied the creation of scientific concepts from a naturalized perspective in the philosophy of science: the cognitive historical approach. Why did she do this? What properties does this method possess that justify such employment? The main purpose of this chapter is to clarify the way in which she understands this method and to underline its merits. In order to achieve this goal, I will introduce the main problems about the issue she deals with by this method, and, on the other hand, I will mention the solutions to these questions that she has been able to provide using the cognitive-historical method.

This chapter consists of three parts. The first, Sect. 17.1 *Questions About the Creation of the Scientific Concepts*, highlights that, within Nersessian's recasting of the problem of conceptual change, there lies a fundamental question related to the creation of

scientific concepts, and that yet another question arises from this one. The basic question is about the nature of cognitive processes implied in the generation of ideas fundamentally new in human history, which leads to the assessment of their effectiveness in achieving successful results. First, I will introduce Nersessian's review of the way in which logical positivism and the historicist philosophy of science have framed the problem of conceptual change, as well as her critical evaluation of this matter.

The second part, Sect. 17.2 *Epistemic Virtues of the Cognitive-Historical Analysis*, deals with Nersessian's conception of the cognitive-historical method, emphasizing those qualities that make it a useful tool for answering the open questions about the creation of scientific concepts. In addition, it is pointed out here that, compared to other approaches to the creation of scien-

tific knowledge, the cognitive-historical method can be considered primary for it is the one that establishes the generative mechanisms of creative concepts in a historical sense.

The last part, Sect. 17.3 *Hypothesis About the Creation of Scientific Concepts*, has two sections. The first one introduces the dynamic hypothesis proposed by Nersessian to give a solution to the problem of the nature of practices that generate new scientific concepts. Mainly, it will be observed that, through her cogni-

tive-historical investigations, Nersessian confirms that scientists often employ model-based reasoning in an iterative way until they solve representational problems in the target domain. The second section presents the general conception of the meaning or the representation of concepts that Nersessian proposed in the first place so as to understand the change of conceptual structures, and the one that she has more recently suggested, in order to explain the effectiveness of model-based reasoning for creating new concepts.

## 17.1 Questions About the Creation of Scientific Concepts

This section will elucidate the questions about the creation of scientific concepts that have motivated Nersessian's investigation. For this I will show that, to a large extent, they arise from the need to find a suitable answer to the *problem of conceptual change*, and, in turn, from the assumption that such an answer depends on a recasting of the problem itself. Nersessian suggests posing this question from the naturalistic point of view in the philosophy of science, that is, the historical-cognitive method. We will see that her questions regarding concept formation are a core part of this new way of understanding the problem of conceptual change, or else, that they are further questions derived from the given answers to this problem.

In Sect. 17.1.1, I will deal with the way in which the problem of conceptual change has been presented by logical positivism and by the *historicist* philosophy of science, and I will also treat the evaluation of this formulation by some naturalistic scholars of science. In Sect. 17.1.2, I will examine the proposal of tackling the problem of conceptual change from a naturalistic point of view. In Sect. 17.1.3, I will present the new naturalistic casting of this problem, particularly its historical-cognitive version. I mention here that it is a core part of this casting to try to determine the nature of the cognitive processes that generate new concepts.

### 17.1.1 The Problem of Conceptual Change

The problem of conceptual change in science has been interpreted and faced in several ways. Although a crucial issue can be identified in it – “How, if in any manner at all, are successive scientific conceptualizations of a domain related to one another?” [17.1] – Nersessian revises the specific manner in which it has been posed and faced by logical positivism and by the *historicist* philosophy of science, and she refers to how some naturalistic researchers, including herself, have viewed this. Regarding the first, she considers that they understood

the problem by focusing on the terms of the change, that is, on the concepts, and that they analyzed those terms as linguistic structures. The neo-positivists faced the problem with the idea that the new conceptual structures are logical extensions of the previous ones; that is why they thought of conceptual change as continuous and cumulative; Kuhn, on the other hand, tackled it by introducing the idea that conceptual change is abrupt and discontinuous, which he developed in his thesis on incommensurability.

Nersessian, like other scholars of science, considers that both positions – the logical positivism's and Kuhn's – are unsatisfactory. Particularly, the thesis of the incommensurability seems to them anti-intuitive and contrary to historical evidence. In fact, they pointed out that, according to the results of individual studies on scientific creativity made by historians of science, conceptual change is continuous and noncumulative [17.2]. Nersessian highlighted that the answers of these two great philosophical movements of the twentieth century are based on an unsuitable treatment of the question, and that this was due to two main reasons:

1. That most of the philosophers of logical positivism assumed that the analysis of science is done in two contexts, the context of justification and the context of discovery; and that they were convinced that the context concerning philosophy of science is the justification one, and not the discovery one.
2. That the group of science philosophers and historians of which Kuhn was part, together with Hanson and Feyerabend, in spite of having taken the initiative to study the context of discovery, did not have the analytical tools necessary to investigate in depth the scientific activities that it seeks to comprehend.

I will give a further explanation of these two points. Regarding the first reason mentioned, Nersessian recounts that the neo-positivists worked within what they

called *the justification context*. This means that they tended to make *rational reconstructions* of science, in particular artificial maps of the logical relationships between concepts [17.3, p. x]. They analyzed scientific concepts as linguistic structures and considered that logical and conceptual studies were enough to understand the meaning of the scientific theories, not studying the real scientific activity [17.4, p. 4]. With respect to the second reason why she thinks that the treatment of scientific change has been unsatisfactory, *Nersessian* highlights the fact that, though *Kuhn* and *Feyerabend* resorted to scientific knowledge in order to understand conceptual change, the Gestalt psychology available then did not give them the suitable tools for that [17.1, p. 6]. She points out that the perceptive metaphor of *change of Gestalt*, which *Kuhn* took from the Gestalt psychology, had an adverse effect [17.1, p. 6].

“[...] By emphasizing the endpoints of a conceptual change (e.g., Newtonian mechanics and relativistic mechanics) [...], the change of Gestalt was made to appear artificially abrupt and discontinuous.”

Moreover, *Nersessian* wrote that, as a philosopher of science, *Kuhn* neglected the processual aspect of conceptual change [17.1, p. 7].

“Significantly, although *Kuhn* does talk about discovery as an *extended process* [17.5, pp. 45ff] and, in his role as historian of science, has provided detailed examinations of such processes, in his role as philosopher of science he identifies conceptual change with *the last act when the pieces fall together* [17.6].”

Regarding the discussion of the problem of conceptual change in the second half of the twentieth century, *Arabatzis* and *Kindi* [17.4], *Andersen* and *Nersessian* [17.7], and *Thagard* [17.8] can be consulted.

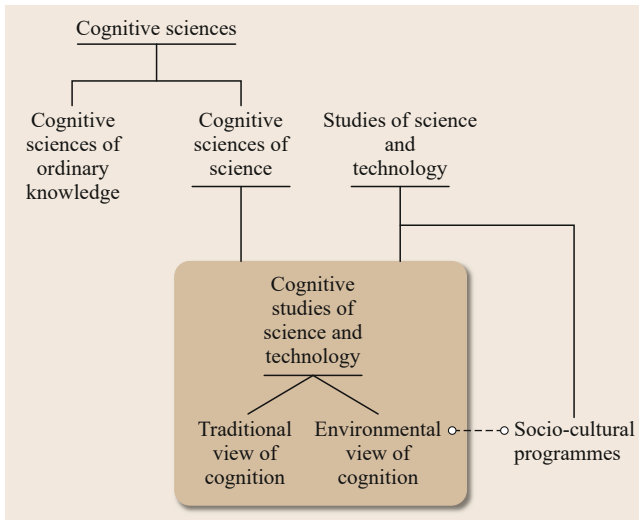
### 17.1.2 The Naturalistic Approach to Science: Revision of the Problem

During the 1980s and 1990s, *Nersessian*, *Paul Thagard* and *Hanne Andersen*, and *Peter Barker* and *Xiang Chen* concurred in revising the *Kuhnian* idea of a radical and sudden conceptual change in science, and they realized that something like that would be, at the most, exceptional. Thus, *Thagard* proposed that to think of changes as gestaltic ones “[...] makes it hard to see how conceptual change can take place” [17.8, p. 49]; see also [17.9]. The conviction that a naturalist scientific approach would enable facing the problem in a suitable way began to impose itself. That is why several scholars

of science further developed such an approach which had been started by the *historicist* philosophers. From this developed a methodological form of naturalism, which defends the need to appeal to science in order to understand science [17.10, Introduction]. Methodological naturalism is one of the various kinds of naturalistic philosophy that have been proposed. *Nersessian's* approach, like *Ronald Giere's*, belong to this position. Usually this one is attributed to *Kuhn* in *The Structure of Scientific Revolutions* [17.5], but was pioneered by *Ludwick Fleck* in *Genesis and Development of a Scientific Fact* [17.11]. Within the domain of philosophy of science, *Fleck* insisted on studying the practice of science instead of propounding rational reconstructions of the logics of the investigation. Therefore, the real subject of the science is an issue of methodological naturalism concern [17.10, p. 3]. *Ronald Giere* characterized the stance as “[...] the vision that all human activities should be understood entirely as natural phenomena, as are activities of chemicals or animals” [17.12, p. 8]. As for the naturalistic studies of science, he describes them as the perspective of “[...] using science in the attempt to understand science itself” [17.13, p. 145]. *Nersessian* bases her own research of science on a naturalistic approach that, to understand scientific knowledge, philosophical theories need to have the best scientific information available on the human subject and about the practices for constructing knowledge used by scientists; she also holds the view that empirical methods are admissible in developing and testing philosophical hypothesis [17.14, pp. 4–5].

### 17.1.3 The Naturalistic Recasting

With the adoption of the naturalistic approach to science, the presentation of the problem of conceptual change is modified. The contrived logical reconstructions of science are replaced by the study of effective scientific practices aiming at explaining the continuous and noncumulative character of conceptual change. *Kuhn* conceived normal scientific cognition in terms of *practices* of solving puzzles guided by solutions to *exemplar* problems [17.15, Chap. 6]. The post-*Kuhnian* group of cognitive orientation holds that interest in scientific activity, focusing, specifically, on the practices of formation and changing of scientific concepts. In general, scientific practices can be understood as procedures carried out by scientific agents. For example, *David Gooding* characterizes the notion of *procedure* implied in such practices as “[...] a sequence of acts or operations whose inferential structure is undecided” [17.16, p. 8]. An overview of the philosophical discussions related to scientific practices can be found in the work of *Joseph Rouse* [17.17]. Although



**Fig. 17.1** Participation of the environmental cognitive studies of science within the cognitive sciences of science and within the studies of science and technology

a scientific procedure is a singular process, two complementary aspects can be distinguished within it:

1. The experimental practice, that is, the manipulation of objects, tools, and experience
2. The intellectual or theoretical practice, that is, the manipulation of concepts, models, propositions, and formalisms.

*Nersessian* presents the problem of conceptual change in the following way: “[...] *how it is that scientists build on existing structures while creating genuine novelty*” [17.1, p. 9]. Unlike the logical positivists, the naturalistic researchers of science understand the question in a way focused on the practices developed by scientists. Thus, they incorporate the *context of discovery* in epistemology. *Nersessian* even came to appreciate the need to demarcate a new domain of investigation, different than both the justification context and the discovery one, which she called *context of development* [17.1, p. 6].

“The context of development is the domain for inquiry into the processes through which a vague speculation gets articulated into a new scientific theory, gets communicated to other scientists, and comes to replace existing representations of a domain.”

I interpret that her goal in doing this was to emphasize that her investigative interest did not focus on

sudden acts of innovation and conceptual change, but on long-lasting processes.

In particular, *Nersessian* develops a naturalistic approach with which she recasts the problem and tries to explain it: the cognitive-historical approach. She says that [17.1, p. 8]:

“In cognitive-historical analysis the problem of conceptual change appears as follows. It is the problem of understanding how scientists combine their human cognitive abilities with the conceptual resources available to them as members of scientific communities and wider social contexts to create and communicate new scientific representations of a domain.”

So her presentation aims at understanding how the cognitive abilities that scientists have as human beings within an environment – and which they share with people who have ordinary knowledge – enable them to construct new concepts. These underlying abilities behind scientific practices are integrated within the cultural, material, and social context, and this environment provides them with conceptual resources [17.14, p. 5]. Although I will analyze the cognitive-historical method in the next section, here I will place it within the cognitive studies of science and technology, one of the naturalistic perspectives of science (Fig. 17.1).

Cognitive science – a confederation of disciplines that includes philosophy, cognitive psychology, artificial intelligence, neurology, cognitive anthropology, and many other areas of study – does research on the cognitive processes and structures implied in ordinary knowledge. It also examines the cognitive mechanisms and the representations involved in scientific knowledge. The *contemporary cognitive science of science* is a naturalist perspective that began to be elaborated in the 1980s and which *Nersessian*, together with Ronald Giere, Lindley Darden, and other philosophers, promoted within the field of philosophy [17.18, p. 2]. One of their main objectives is to explain how scientists construct science naturally. Giere describes this purpose as the aim to reveal how scientists manage to interact with the world when they make science. He points out that scientists, as human beings, have several cognitive abilities, biologically founded, to make that possible [17.12, p. 5].

Within the cognitive science of science, there is a variety of approaches. Some of them, denominated *cognitive studies of science* or *cognitive studies*, consider it relevant to take into account the cultural and social dimensions of scientific practice. These approaches, besides being part of the cognitive science of

science, also participate in the studies of science and technology (STS), together with the social–cultural programs. Some of the methods of the cognitive studies of science use the traditional view of cognitive science. This means that they assume that cognition is a processing of symbols that occurs within the individual minds of humans. Therefore, the cultural and social dimensions of scientific practice are not an integral part of their analysis. On the other hand, other approaches of cognitive studies, the environmental ones, recognize that material, cultural, and social environments in which science is practiced, are crucial to understand scientific cognition. Nersessian argues that, in order to give accounts which capture the fusion of the cultural-cognitive-social dimensions in the practices producing scientific and engineering knowledge, it is convenient to use the view of environmental perspectives in cognitive science. *Nersessian* [17.19] provides an overview of the environmental analysis lines of research that have been delineated between the 1980s and the year 2000. The environmental paradigm emphasizes that sociocultural and body factors have a substantial role in cognitive processes [17.19, 20]. Although it should be pointed out that researchers in cognitive science that study science based on the environmental perspective resist the view of those who think that

all the aspects relevant to science can be explained in terms of sociocultural factors. Nersessian considers that this view is a form of reductionism which is manifested, for example, in the declaration of a 10 year moratorium on cognitive science studies, which was initiated by *Bruno Latour* and *Stephen Woolgar* in 1986 [17.21].

In short, the cognitive-historical method is an environmental approach inscribed within the cognitive studies of science with which Nersessian interprets the problem of conceptual change. The new version of the problem of conceptual change contains a basic question related to the creation of scientific concepts: Which situated cognitive processes, that is, integrated within their environment, do scientists develop in order to come to articulate new concepts from vague notions? As will be treated in Sect. 17.2, this issue has led to investigations that indicate the sought-after processes are *model-based reasoning*. This conclusion, in turn, impelled the posing of new questions, and one among them that stands out for its relevance and potential to determine why this kind of reasoning affords an appropriate medium to generate new scientific concepts is: “*What features of model-based-reasoning make it a particularly effective means of conceptual innovation?*” [17.14, p. 186].

## 17.2 The Epistemic Virtues of Cognitive Historical Analysis

Nersessian studies scientific practices from the perspective of the cognitive-historical approach, aiming at providing suitable answers to the questions related to the creation of scientific concepts mentioned in Sect. 17.1. Here, I will analyze the way in which she conceives of this method, highlighting the advantages it offers to carry out such investigations about science.

In Sect. 17.2.1, I will characterize the cognitive-historical approach in a general way, and will present a brief overview of its expressions and of the conception Nersessian has about it. In Sect. 17.2.2, the dimensions that constitute the cognitive-historical method highlighting the virtues that enable it to provide knowledge about the creation of scientific concepts will be analyzed. In Sect 17.2.3, the cognitive-historical analysis with other approaches used to study the creation of scientific knowledge, for example, ethnography, in vivo observation, and laboratory experiments will be compared. These contrasts will highlight a distinctive property that deserves to be considered a *primary* approach to the study of the creation of scientific concepts with a historical impact.

### 17.2.1 The Cognitive–Historical Approach

Nersessian mentions several approaches to cognitive studies of science provided by the various disciplinary fields within them – in particular philosophy of science, history, cognitive psychology, and cognitive anthropology – from which perspectives creative scientific practices are investigated. She states that ([17.22, p. 127]; [17.18, Chap. 1])

“In contemporary cognitive studies of science, the methodologies employed in investigating the practices scientists use in creating knowledge are ethnography, in vivo observation, laboratory experiments, and cognitive-historical analysis.”

Among these methodologies, she evaluates one of them as particularly advantageous in order to undertake the investigation of those creative practices: cognitive historical analysis. What characteristics does it have? A broad way of describing it is that it is the result of the combination of cognitive and historical methods [17.23, p. 42]. It was *Nersessian* who coined the



term *cognitive history*: “Recently a research frontier I call *cognitive history* has emerged within the history of science and is finding its place in this confederation (Cognitive science)” [17.24, p. 194]. However, this kind of mixed approach was used by other authors, at least since the 1970s. *Tweney* [17.25] presented an overview of its versions that includes the contributions published principally since the 1980s. He restricts his attention to

“those studies that are truly *cognitive-historical*, that is, to those studies that have used cognitive frameworks for the understanding of historical episodes *in such a way* that the methodological niceties of both history and cognitive science are respected.”

They comprise:

1. The application of the solving-problem approach to scientific discoveries by Herbert Simon and his colleagues, understanding the solving of problems as the search within a space of problems which follows heuristic principles, for example, *Kulkarni* and *Simon* [17.26, 27]
2. Kuhn’s appeal to cognitive psychology in order to understand the change in paradigms in *The Structure of Scientific Revolutions* and the work on Kuhnian psychology of science by *De Mey* regarding conceptual change in *The Cognitive Paradigm* [17.28]
3. The use of the Piagetian theory in the cognitive biography of Charles Darwin by *Howard Gruber* [17.29] and the use of some Piagetian ideas to explore the development of concepts in modern physics by *Miller* [17.30]
4. The cognitive formalization of some aspects of Kuhn’s theory about scientific change with the purpose of understanding cases in the history of science by *Andersen et al.* [17.9]
5. The research by *Nersessian* on the psychological nature of change of meaning in science, particularly of the notion of electromagnetic fields in theoretical physics [17.3], and its subsequent elaborations
6. The use of tools provided by cognitive sciences in order to interpret scientific work by Faraday by *Tweney* [17.31]
7. The reconstruction of experiments made by Michael Faraday and the design of conceptual maps of the process by which his vague interpretations became concrete, conveyable scientific concepts by David Gooding; as well as his conception of the emergence of scientific concepts as a process where “hand, eye, and brain” interact [17.16]
8. The cognitive-historical replications by *Tweney* [17.32] and by *Cavicchi* [17.33, 34].

The way in which *Nersessian* understands the cognitive-historical method constitutes a kind of philosophical analysis that integrates several contributions from philosophy, history, and psychology. This is not a purely formal analysis like the ones exalted by logical positivism and questioned by the philosophy of science of the mid-twentieth century. It is, by contrast, an interpretation of historical scientific practices in terms of contemporary cognitive science with the purpose of elaborating a philosophical conception of specific conceptual developments ([17.35, p. 3] and [17.36, p. 163]). A particularity of this approach, as underlined by *Nersessian*, is its reflexiveness [17.1, p. 7]. This means that, from this perspective, the methods, theories, and cognitive categories are used to interpret historical cases and that they are also objects of examination themselves. In fact, these tools may be inadequate to understand the complexity of science, and it is necessary to pay attention to this, in order to identify which changes need to be done on them. Such changes do not only affect the cognitive science of science, but also the cognitive science of ordinary cognition. Consequently, through cognitive-historical analysis, the studies on scientific cognition feedback to the field of cognitive science, thus forming the base for additional cognitive research [17.14, p. 7].

### 17.2.2 Epistemic Virtues and Dimensions of this Approach

References to the dimensions of the cognitive-historical method can be found in multiple articles and books by *Nersessian* [17.1, 14, 22, 24]. I will analyze these dimensions with the purpose of deriving the characteristics that make this perspective an advantageous one to answer the questions related to the creation of scientific concepts. It will be seen that, on the historical side of the method, they are: to enable the epistemic access to conceptual change, and therefore, to the practices of the creation of scientific concepts with historical impact; to enable a deep, detailed study of scientific cognition; to satisfy the requirement of ecological validity, that is, of not altering the studied phenomenon by placing it in an artificial situation; and to retrieve data from the scientific practices in a way that exceeds the verbal accounts. On the cognitive side, one attribute stands out: Making possible interpretations of historical practices in a manner that general conclusions about their nature and function can be drawn.

The *historical dimension* of this method is understood in a wide sense. On the one hand, it is a temporal perspective, that is, it seeks to recover the way in which representational, methodological, and reasoning scientific practices are developed during a long period of

time. Some examples of practices approached in this way are [17.24, p. 194]:

“[...] devising and executing real-world and thought experiments, constructing arguments, inventing and using mathematical tools, creating conceptual innovations, devising means of communicating ideas and practices, and training practitioners.”

This temporal feature of the method is valuable because it makes it possible to obtain knowledge on conceptual change, a scientific phenomenon that is difficult to capture because it is exceptional and it usually implies long lapses. The historical component of the analysis does not constitute a historical narrative, but a detailed investigation of microstructures and micro-processes, more specifically, of representational practices and practices of problem solving [17.1, 14, 24]. On the other hand, the historical dimension of the method is a contextual perspective, that is, it takes into account the community where the scientific practices have been carried out and the cultural resources implied therein. This point of view has the virtue of satisfying the objective of preserving the essence of the phenomenon under investigation pursued by the ecological approach in the psychological investigation. One of the ways to access long-term scientific activities is through historical records. Among the sources are: diaries, laboratory notebooks, publications, correspondence, experimental equipment, drawings, diagrams, lecture notes, and texts. In the case that the investigated procedures are long term and extend into the present, data about them can be obtained in other ways, for example, using field observation and other ethnographic methods (see the next section). Here the main sources of information are the cognitive tools employed in scientific activities and the artifacts produced by them. Consequently, another benefit of the cognitive-historical method is that the information about the practices embedded in the conceptual change is not restricted to scientists' verbal accounts.

The cognitive *dimension* of the method is inscribed in the tradition of psychological epistemology, including works by Locke, Hume, and Quine [17.1, p. 5]. It refers to the employment of cognitive sciences for understanding the scientific practices involved in the creation and change of concepts. It postulates that results, interpretations, and relevant debates on cognitive science would help to understand such scientific practices [17.14, p. 6]. As has been previously mentioned, the cognitive approach adopted by *Nersessian* is the *environmental* one, and this implies considering scientific and engineering thought as a complex *system* that com-

prises material, cultural, and social aspects [17.19]. The *continuum hypothesis* justifies that the achievements of cognitive sciences can be employed to understand scientific practices. This hypothesis refers to the human cognitive capacities and mechanisms of those who make science. It maintains that they are *basically* the same as those of ordinary humans. Therefore, to a great extent, what scientists do, and the constraints they experience, derive from their human cognitive condition [17.1]:

“The underlying presupposition is that the problem-solving strategies scientists have invented and the representational practices they have developed over the course of the history of science are very sophisticated and refined outgrowths of ordinary reasoning and representational processes.”

The *continuum hypothesis* does not negate the fact that there are great differences between scientific and the ordinary cognition. Indeed, scientists have a vast knowledge of a specific domain, have a methodological training, and have learned to metacognitively reflect on and refine the use of cognitive capacities that give them the ability to reason scientifically in carrying out the necessary cognitive functions [17.37, p. 2]. *Nersessian* defends this hypothesis by pointing out that it is not speculative, it is not an *a priori* conjecture, but that it is, on the contrary, a claim based on psychological facts. She also suggests that the understanding of the investigated scientific practices consists in attributing to them a sense which transcends the specific characteristics of the case. This appears to derive from the fact that this sense would be similar to, if not the same as, the one that makes ordinary cognitive practices generalizable. Therefore, the possibility to make *general* descriptions about the nature and processes of scientific activities is opened. In fact, *Nersessian* affirms that such regularities can be abstracted from the *thick descriptions* of the particular case ([17.38, Chap. 1]; [17.14, p. 9]). *Thick description* is a concept from qualitative investigation. It is thought to have been introduced for the first time by Gilbert Ryle. For Ryle [17.39], the *thick* description implied assigning intentionality to one's behavior. The *thick* description interprets behavior within the context and assigns thought and intentionality to the observed behavior. So, for Ryle the *thick* description implies understanding and absorbing the context of the situation or the behavior. Also it signifies assigning current and future intentionality to behavior. Geertz borrows this philosophical term from Ryle in order to describe the ethnography work [17.40]. In this way, abstracting regularities about scientific activities, the problem that historicist philoso-

phers have to face is solved: the one of *how to go from a case study to a more general conclusion*, avoiding the risks of making a hasty generalization [17.1, p. 35]. The cognitive-historical approach, therefore, is satisfactory for understanding the nature of scientific practices and surmounting conclusions established for particular cases.

### 17.2.3 Cognitive Methods to Investigate Conceptual Innovation

Nersessian asserts that the *primary* method to investigate practices of conceptual innovation in science is cognitive-historical analysis (Fig. 17.2). She agrees with other cognitive researchers of science that none of the approaches in use is sufficient by itself to understand those practices and considers, like them, that those methods must be complemented in order to understand the complexity of scientific cognition. For example, *Simon* and *Klahr* assert that [17.41, p. 531]:

“[...] the fundamental thesis in this article is that the findings from these diverse approaches, when considered in combination, can advance our understanding of the discovery process more than any single approach;”

a similar position is taken by *Thagard* [17.18, Chap. 1]. However, at the same time, she considers that not all the employed methodologies are necessarily equal. There is a central feature in cognitive-historical analysis that makes Nersessian consider it fundamental and that distinguishes it from the other methodologies: It enables acquisition of knowledge about the mechanisms that have had a historical impact in science. This quality will be properly understood after examining the other approaches and highlighting in them a characteristic they all share, and which contrasts with the one just mentioned as distinct from cognitive-historical analysis.

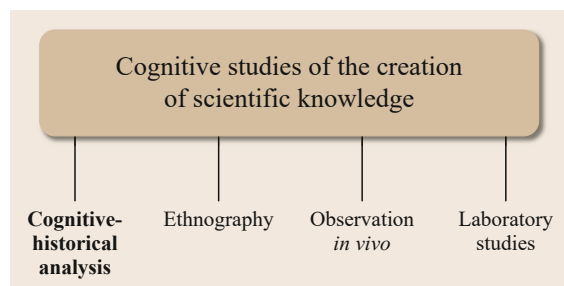
*Ethnographic studies* are based on field work and, in particular, studies about science are usually based

on research made in scientific laboratories [17.42]. Ethnography of science started at the beginning of the 1980s with the sociology of scientific knowledge (SSK) considered as part of social studies of science and technology (STS) – with the work of *Latour* and *Woolgar* [17.21] and *Knorr-Cetina* [17.43], among others. But ethnography of science also is deemed as among the cognitive studies pursued to investigate cognition and its context in mutual relation ([17.44–46]; [17.19, p. 38]). Thus, Nersessian points out that both the way in which scientists understand the problems and the tools they use to solve them depend on a sociocultural context [17.14, p. 7,9]. Specific ethnographic methods are qualitative: They include field observation, collection of artifacts, and field interviews. Each of them is useful to investigate the heuristic of discovery that scientists follow in their daily scientific practice, as they carry it out, that is, in real time and, therefore, to study scientists’ cognitive mechanisms when creating new concepts.

*In vivo* observation is a method suggested by *Dunbar* [17.47, 48]. It arises from the premise that scientists forget many of the important thought processes they use, and that, consequently, there is no record of them in their notes or laboratory notebooks. So he argues that it is necessary to investigate living scientists in order to have information about these processes, that is, by observing them while they perform their work. During observational studies, important activities at the laboratory are recorded, and then the data are codified and interpreted within the frame of psychological constructs.

*Laboratory experiments* are studies of the processes for solving problems of people in artificial situations with the purpose of isolating one or more relevant aspects of science from the real world. In general, they are performed in a psychology laboratory. The experimentation on discovery processes is carried out employing nonscientists as experimental subjects. The principal roles for this experimentation are two: being tools for testing the hypothesis already posed, and being exploratory tools. In this latter case, the experiments are conducted to make certain phenomena appear [17.41, pp. 526–527]. *Klahr* maintains that the experimentation on the discovery processes enables a detailed analysis of the processes of solving problems and that they can be entirely identified and recorded [17.49].

Ethnography, *in vivo* observation, and laboratory experiments have in common that they all provide information about the activities performed at the scientists’ own laboratories, or at laboratories of cognitive sciences during normally brief sessions: hours or days. However, the ethnographic studies mixed with the cognitive-historical method can be extended for years, each of them providing relevant information. An example



**Fig. 17.2** Cognitive approaches of the creation of scientific knowledge

of this is Nersessian's and her collaborators' own research in biomedical engineering laboratories. In this case [17.44, p. 2],

“Ethnographic studies tell us what BME practitioners do in the context of their research and cognitive-historical studies provide insight into how and why these cognitive practices have evolved.”

What is more important is that these approaches are characterized by revealing the creative mechanisms that underlie the practices performed when learning, taking ownership of, and employing *existing concepts*. In other words, they inform about the scientists' cognitive mechanisms that are psychologically creative, that is, mechanisms that produce a conceptual novelty *for them*.

Unlike these approaches, the cognitive-historical method is the only one that can provide information about cognitive practices that are extended in time, thus revealing the mechanisms generating creative concepts in a historical sense, that is, of those concepts previously nonexistent and that have, consequently, historical impact [17.50]. Nersessian adopts Margaret Boden's distinction between psychologically creative ideas and, on the other hand, historically creative ideas. While the first ones “[...] are surprising, or even fundamentally novel, with respect to *the individual mind* which had the idea [...],” the historically creative ideas have not been previously thought by anyone [17.50, p. 43]. This is the reason why Nersessian considers the cognitive-historical method necessary to understand conceptual innovation. Besides, she has in mind that historically creative mechanisms are also psychologically creative, for they generate not only novelties for humanity, but also novelties for the scientists themselves. For this, she considers that ethnography, in vivo observation, and laboratory experimentation are necessary, although in a secondary way, for studying conceptual change. In fact, as these three approaches provide knowledge about scientists' psychologically creative mechanisms, they

can improve the cognitive interpretations of historical episodes in scientific change provided by the cognitive-historical analysis [17.22].

An objection that might be made to the employment of the cognitive-historical method to investigate scientific practices of the creation of concepts is related to a question that any proposal to naturalize the philosophy of science should answer: whether it is legitimate to use scientific knowledge in order to develop a theory of scientific knowledge. A general argument contrary to such a proposal comes from traditional philosophy of science. It states that the use of scientific methods to investigate science is necessarily circular, that it supposes a *petitio principii*, or leads to a regression [17.51, p. 333]. Nersessian defends the use of the cognitive-historical method to develop a theory of the production of scientific knowledge from the argument of circularity. She adopts a point of view similar to the one of the philosophers who practice a naturalized antifundamentalist epistemology, such as *Ronald Giere*, who maintains that [17.12, p. 11]

“The [Cartesian] program of trying to justify science without appeal to any even minimally scientific premises has been going on without conspicuous success for 300 years. One begins to suspect the lack of success is due to the impossibility of the task.”

The circularity implied in the naturalized conception of science, therefore, seems to be inherent to humans and insurmountable for them. But this does not mean that such circularity is vicious. Nersessian supports a virtuous circularity, which could be obtained by putting cognitive and historical interpretations in a state of reflective equilibrium. As has been pointed out earlier, she considers that this reflexivity is a particularity of the cognitive-historical analysis [17.1, p. 7]. Therefore, the studies on scientific cognition provide feedback for the field of cognitive science, thus forming the basis for additional cognitive investigation [17.14, p. 7].

### 17.3 Hypothesis About the Creation of Scientific Concepts

With the cognitive-historical approach, *Nersessian* investigates the cognitive processes that create scientific concepts and succeeds in posing a hypothesis that integrates the dynamic area of a cognitive theory of conceptual change [17.1, p. 5]. This part of the theory *explains* the phenomenon of conceptual change in terms of mechanisms, in a similar way to how dynamics explains movement in physics. This dynamic investiga-

tion requires some representation of scientific concepts, and more broadly, of concepts in general, and, all the more, if one expects to understand why certain mechanisms are very effective in generating them. This is a meta-theoretical question considered as a central one in another of the areas of a cognitive theory of conceptual change, namely, kinematics. This one deals with *describing* the conceptual changes that have occurred

throughout the history of science, in a way similar to that in which in kinematics describes movement in physics.

In Sect. 17.3.1, I will expatiate the dynamic hypothesis proposed by Nersessian to solve the problem of the nature of the practices that create new scientific concepts from the cognitive-historical perspective. In Sect. 17.3.2, I will cite how she understands the power of modeling to generate scientific concepts and refer to her proposal on how to consider the representation of concepts in elaborating this position.

### 17.3.1 Dynamic Hypothesis

The cognitive-historical analysis proceeds by bootstrapping, that is, enables the establishment of hypotheses about conceptual structures and cognitive processes implied in historical cases of scientific investigation, which in turn, serve as support to make new historical studies and cognitive interpretations that provide additional knowledge. Let us see how Nersessian employs it when she poses the *dynamic hypothesis* to solve the problem regarding the nature of the cognitive processes that scientists elaborate in order to articulate new concepts.

The historical studies of several episodes that have led to a conceptual change in science provide information which, according to *Nersessian*, supports the following conclusion: Conceptual innovation does not occur suddenly, but results from *extended problem-solving processes* [17.52, pp. 13–14]:

“[...] If one examines their deeds [of scientists] – their papers, diaries, letters, notebooks – these records support a quite different interpretation in most cases. As I have been arguing for some years, conceptual change results from extended problem-solving processes.”

This way of interpreting historical information fits with the reality that Nersessian includes herself within an epistemological tradition constituted by Dewey, Mead, and Popper, by which science is seen precisely as a problem-solving process. However, in spite of her closeness to these authors, she assumes some distance from them, for their view has a limited range, that is, it does not include the scientific phenomenon of conceptual change [17.1, p. 12]. Nersessian articulates the basic interpretative scheme of the cases of conceptual change as problem-solving processes in psychological terms and, in its elaboration, she uses some concepts that come from Gestalt theory and also from cognitive psychology. A brief reference to them will help to clarify her interpretation.

R. E. Mayer reminds us that, in the psychological literature, a problem consists of a given state, a state of destination, and a set of operators. The problem occurs, according to Gestalt theory, when a situation is in one state, the solver wants it to be in another state, and there are obstacles that prevent a fluid transition from one state to the other. In addition, Mayer asserts that, from a cognitive perspective, the solution of problems is “[...] directed, cognitive processing aimed at finding a way to achieve a goal” and that two phases can often be distinguished in it: the representation of the problem and the solution of the problem. The represented problem may be of various kinds, one of them – of particular relevance for our explanation – is the one of representational problems. There are usually two different ways of finding and carrying out the solution of problems:

1. The sudden appearance of the solution (insight leap), which occurs immediately after a sudden and more suitable restructuring of the representation of the problem – this kind of solution often comes along with the aha! or *Eureka* experience, a subjective feeling of surprise.
2. The process of step-by-step solution – also called the *analytic method* – which consists of finding a strategy and executing a sequence of actions in order to generate a solution to the problem [17.53, pp. 112–113].

*Weisberg* categorizes step-by-step problem solving as an analytic method: “[...] we can categorize the various modes of solving problems that are based on degrees of specificity of knowledge about a problem as *analytic methods*” [17.54, p. 282].

By means of these psychological conceptual tools, Nersessian interprets that, basically, the kind of problem that occurs in conceptual change is a *representational* one, and that the solution implied is a step-by-step one. This means that, on the one hand, this problem consists of a given situation in which a certain phenomenon escapes understanding and the solver does not know how to obtain the new conceptual resources to understand it [17.14, p. xii]; and that, on the other hand, the solution to the representational problem of conceptual change is reached through an heuristic strategy that, as will be immediately apparent, consists of a bootstrapping cycle of modeling, understood as a kind of creative reasoning [17.14, p. 184].

Let us go back to the bootstrapping employment Nersessian’s cognitive-historical method. She realizes that the historical studies of specific cases of conceptual change show, in a ubiquitous way, that scientists use analogy, visual representation, and thought experiments to solve representational problems. She points

out that these three practices have in common that they are ways of modeling, that is, of the construction and manipulation of models. For example, when examining the development of the current concept of the electromagnetic field, she indicates that Faraday articulated the notion of “continuous, progressive transmission of the action” with the help of the concrete visual image of induction “cutting” the strength lines, and with the help of vague analogies between the electric and magnetic actions and *known* progressive phenomena [17.3, pp. 144–145]. Again, Nersessian interprets historical data with cognitive tools and considers that the modeling practices, with which new concepts are constructed, are kinds of reasoning. In doing this, she uses a broad reasoning notion that comes from cognitive psychology, particularly from Johnson-Laird’s semantic conception of reasoning.

Unlike the philosophical traditional notion of reasoning, which only comprises deductive and inductive arguments, Johnson-Laird’s conception enables including kinds of creative reasoning. According to this, much of human reasoning is done through *mental modeling*. When referring to deductive reasoning, *Johnson-Laird et al.* assert that [17.55, p. 3]:

“On the other side, there are those, such as ourselves (see also *Johnson-Laird* and *Byrne* [17.56]) who claim that it is a semantic process that depends on mental models akin to the models that logicians invoke in formulating the semantics of their calculi.”

Following Johnson-Laird, Nersessian maintains that humans retrieve or construct models through which they make inferences about a target problem. The represented structure is supposed to contain parts that can possess an analogous model that also is made of parts. The nature of mental models is, in Peirce’s words, iconic.

The idea that modeling is a kind of reasoning may be applied both to scientific contexts and to ordinary contexts. As she focused on scientific reasoning tasks, Nersessian finds that she needs to extend Johnson-Laird’s conception by widening the domain of mental models. She understands that models are interpretations intended to satisfy the salient constraints of a physical system, process, phenomenon, or situation. This implies that mental models not only comprehend the structural analogues of what is modelled, that is, models which embody representations of spatial, temporal relations, and causal structures, but that they also include functional analogues which are also dynamic in nature. Through her conception of mental modeling, she does not intend to participate in the numerous debates that have been sparked by this notion. That is the reason

why she calls her own hypothesis *minimalist* [17.52, p. 12]:

“To carry out an analysis of model-based reasoning in conceptual change requires only that we adopt a *minimalist* version of a mental modeling hypothesis: that in certain problem solving tasks humans reason by constructing an internal model of the situations, events and processes that in dynamic cases provide the basis for simulative reasoning.”

A property of scientific model-based reasoning is that it does not guarantee the production of a solution, that is, it is not an algorithmic procedure and, because of this, *Nersessian* understands it as heuristic [17.57, pp. 325–326]. The difficulty to produce solutions would be that the models used for reasoning may be unsatisfactory, that is, they may not embody the relevant constraints of the target situation, and not so much that the reasoning may be incorrect [17.52, p. 14].

“In the case of science where the situations are more removed from experience and the assumptions more imbued with theoretical assumptions, there is less assurance that a reasoning process, even if correct, will yield *success*. In the evaluating process, a major criterion for success remains the goodness of fit to the phenomena, but success can also include such factors as enabling the construction of a viable mathematical representation.”

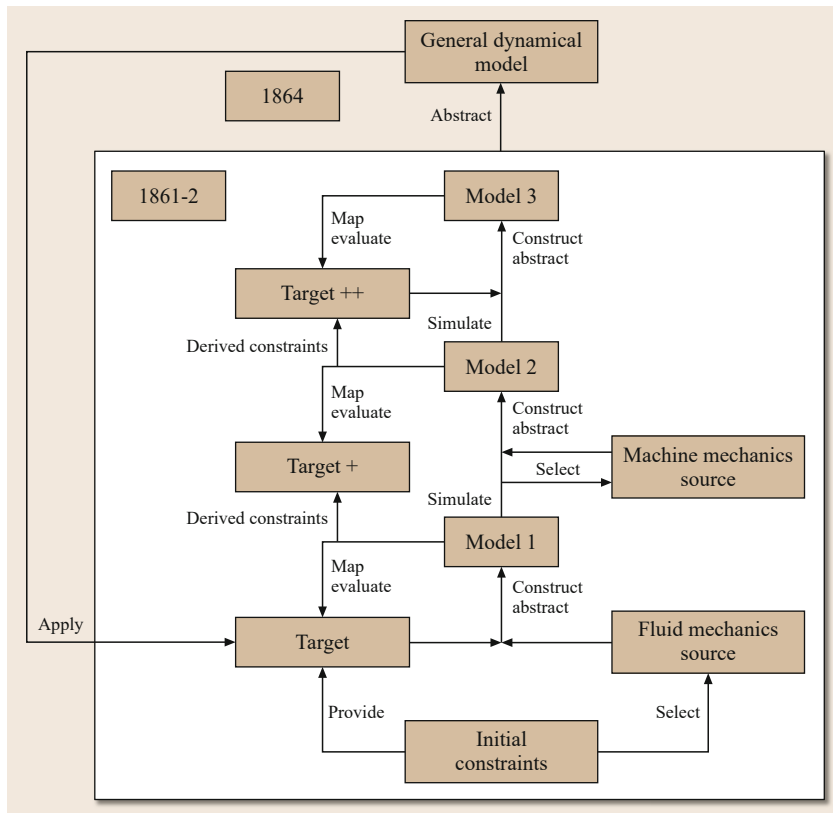
Let us continue with the iterative application of the cognitive-historical method. According to the historical studies of scientific practices, being able to have a model that satisfies the constraints of the target problem frequently involves a cycle of construction, manipulation, evaluation, and adaptation of intermediate models. This is a bootstrapping process. This means that each intermediate model that is constructed achieves a higher satisfaction of the constraints of the target domain, and contributes to constructing the subsequent model. Intermediate models are hybrid, that is, they embody not only the constraints of the target domain, but also those of the respective source domains [17.52, p. 21]. Through the process of construction of satisfactory models, the solution to the representational problems implied in conceptual change cases is achieved. *Nersessian* compares this kind of extended-in-time process with organic phenomena wherein a perfect innovation emerges [17.14, p. ix]:

“Rather, such conceptual innovation, like perfect orchids and flavorful grapes, emerges from lengthy,

organic processes, and requires a combination of inherited and environmental conditions to bud and bloom and reach full development.”

An important feature of the bootstrapping process of creating a model with such reasoning is that it implies selectivity. By means of abstraction and evaluation processes, the irrelevant features are left aside, and attention is focused on the relevant ones according to the problem-solving context. These historical findings differ from those established by the current cognitive theories of analogy. According to them, the models to reason with are already provided by the target source [17.1, p. 20]. Therefore, in order to understand the iterative character of the construction of analogue models in science, it is necessary to modify those cognitive theories. In this way, the reflexive nature of the cognitive-historical method becomes evident. Nersessian deeply analyzes a case that exemplifies the idea that there are modeling processes that generate conceptual innovation. Those are the modeling processes that led Maxwell to make the first derivation of field equations for electromagnetic phenomena. Figure 17.3 shows the contribution of the target, source, and model constraints to these processes of reasoning [17.14, Chap. 2].

The thesis of scientific cognition as model-based reasoning has been developing over more than the last thirty years, due not only to Nersessian’s work, but also to the contribution of authors such as Ronald Giere, Lorenzo Magnani, and Paul Thagard. *L. Magnani* enriched Nersessian’s analysis of model-based reasoning with the reference to the problem of abduction in creative reasoning, also taking advantage of the recent cognitive research on distributed cognition [17.58]. It is fitting to point out the fact that N. Nersessian, since 1998, in collaboration with L. Magnani and P. Thagard, created and promoted the MBR Conferences on Model-Based Reasoning, realizing its seventh convocation in 2015. One of the problems that the thesis of scientific cognition as model-based reasoning presents is that the notion of mental modeling implies that of representation, and the latter has been questioned within the cognitive sciences by the dissenters of the dogma of cognitivism. This *doctrine* of cognitivism is integrated within the representational and the computational theories of the mind [17.59]. As I have mentioned before, Nersessian adopts a moderate environmental perspective that places her on those dissenters’ side. From this environmental approach, she defends the idea of mental modeling, but conceives of it as a procedure carried out



**Fig. 17.3** Maxwell’s modeling process (after [17.14, p. 57])

by a cognitive system constituted by internal representations, frequently coupled with resources from the real world [17.60]. Moreover, together with Lisa Osbeck, she suggests a conception of representations organized in models as practices. These representational practices may be interpreted as distributed, that is, they can be expanded through internal–external traditional domains. These representations are [17.61, Introduction]:

“[...] created and used in the cooperative practices of persons as they engage with natural objects, manufactured devices, and traditions, as they seek to understand and solve new problems.”

In their description of the distribution of representation in scientific cultures, Nersessian and Osbeck employ a language with which they try to convey the co-constitutive nature of culture and cognition, that is, the relation between these two domains in a unique system. In it, the concepts of *cognitive partnering*, *internal–external representational coupling*, and *enactment* are central [17.59, 61].

### 17.3.2 The Power of Model-Based Reasoning

Summing up what was written in the previous part about the hypothesis that Nersessian managed to establish in relation to the creation of scientific concepts, let us say they explain that this scientific practice is based on mechanisms consisting of a modeling iteration. The modeling series ends with the construction of a satisfactory model from which to draw inferences about a target problem. Once she has reached this conclusion about the process of the creation of scientific concepts, Nersessian poses another question. She tries to explain the efficiency of model-based reasoning for creating scientific concepts and, with this in mind, she decided to stipulate a particular concept of concept.

During the 1980s, Nersessian dedicated her studies primarily to the issue related to how to represent concepts in order to understand that conceptual change in science is continuous, gradual, and noncumulative. In this way, she came to describe the representation of a type of concepts as “a set of family resemblances exhibited in its ‘meaning schema’.” But more recently, in [17.14], mentioning the state of the art of the concept representation issue, she stated that there is no agreement either in cognitive science or in philosophy of science about how to conceive of a *concept*. As it is necessary to have some conception about it in order to explain the power of model-based reasoning in the creation of scientific concepts, she propounds one that does not require answer to other problems in debate. In

what follows, I will review the way in which Nersessian treated this issue in order to show the relation it has with her conception of the role of models in creative reasoning.

Nersessian considers the meta-theoretical problem of the representation of concepts as a central theme of the kinematics of conceptual change. This area of the theory of conceptual change aims at determining the *form* of conceptual change, that is, the differences existing between conceptual structures as time goes by, both between conceptual systems and between individual concepts. This task presupposes that a particular representation of concepts is available. On the one hand, the kinds of change that took place between different conceptualizations, the first conceptual structures just mentioned, refer to changes in the form of the organization of the concepts that integrate them. A conceptual system can be analyzed as a network of nodes, where each node corresponds to a concept and each line within the network corresponds to a link between concepts. Within conceptual networks, concepts are organized through links such as kind, property, and relation. These links can be characterized as connections, which indicate that a concept is a kind of another concept, that an object has a property and which express relations, respectively [17.8, pp. 30–31]. Accordingly, the restructuring of the conceptual systems supposes changes related to the concepts that integrate them, and these changes have impact on the other concepts. For instance, changes of hierarchy, changes from properties to relations, and the addition and suppression of concepts. These changes are coordinated, that is, since concepts are interlinked, changes related to a concept have an impact on other concepts.

Nersessian recounts a relevant case of change of conception in the history of science, the one that refers to the representation of movement. She analyzes various phases of this change – the medieval philosophers’, Galileo’s, and Newton’s conceptions of movement – pointing out how some concepts attained a new organization, for instance, the concepts of *movement*, *vacuum*, and *space*, and how the new concept of the *force of gravity* was built. Let us consider a representative sample of the way in which she carries out her examination. She indicates that: *movement* changes in hierarchy within the medieval conceptual structure it is a kind of process, while, within the Galilean conception, it is a state; *gravitas* changes from a property to a relationship – within the medieval mechanics *heaviness* is a property of bodies, while, on the other hand, within Newtonian physics it is a force which acts on the bodies and, as such, a relationship between them; within the Galilean conception, the medieval conception of the distinction of natural/violent movement is abandoned;



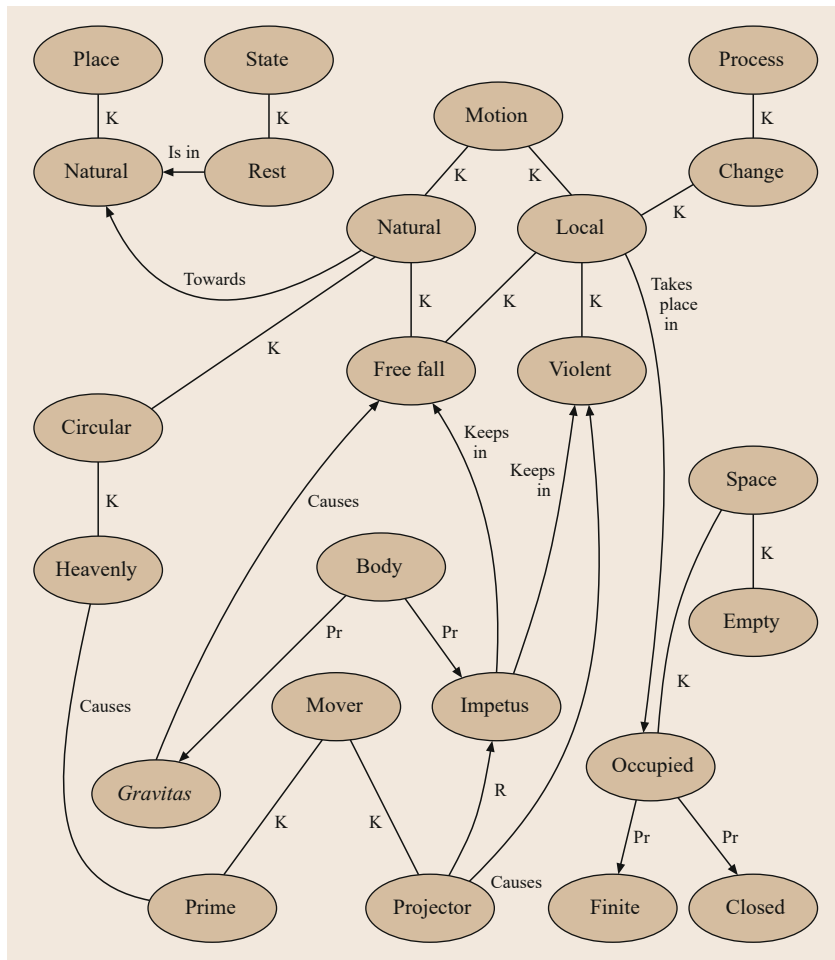
and, finally, within the Newtonian conceptual structure, the *principle of inertia* is added to the theory of mechanics.

Nersessian adopted a system of representation of knowledge, conceptual maps, in order to facilitate the analysis of the changes happening in the various phases. These maps contain conceptual nodes and links between them. For instance, in [17.62, pp. 171–173], she drew three conceptual maps, reproduced in Figs. 17.4–17.6, which represent salient parts of the conceptual structures of the medieval, Galilean, and Newtonian theories of movement. It should be mentioned that Nersessian has made valuable contributions regarding the applicability of knowledge about the way in which conceptual systems in science change to teaching and learning in the field of science [17.62, p. 166].

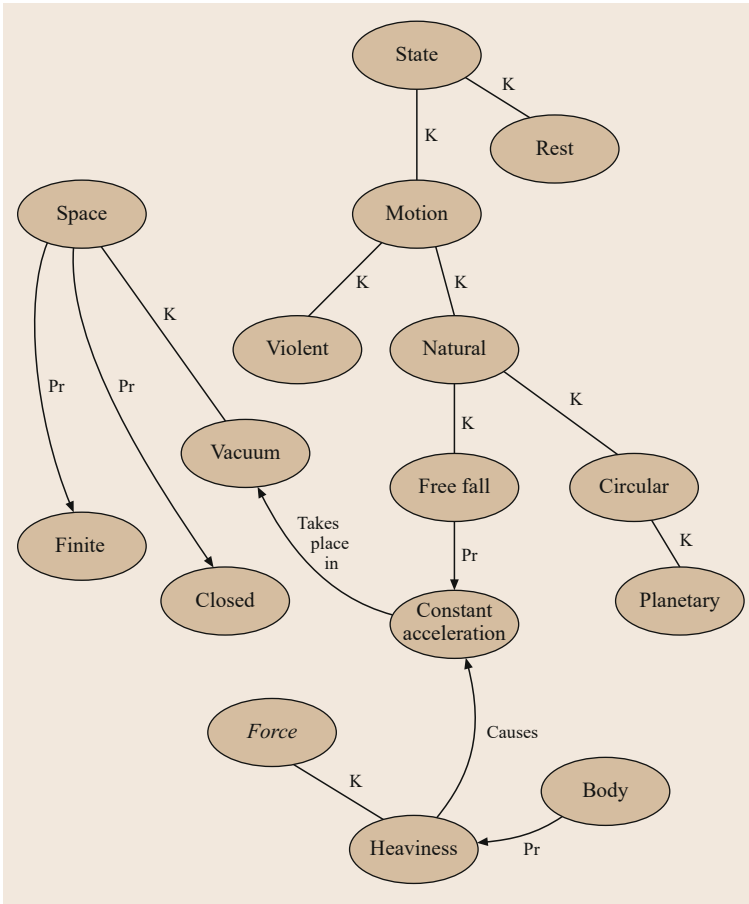
On the other hand, the difference in form that occurs throughout time (in this case not related to complete conceptions, but to an individual concept) refers to the change in meaning between its instances. This kind of

change can be exemplified with the concept of an *electromagnetic field*. Nersessian derived that this concept has undergone three phases. The first one, which she calls *heuristic guide*, encompasses the contributions of Faraday and the first two papers by Maxwell: *On Faraday's Lines of Force* and *On Physical Lines of Force*; the second one, which she calls *elaborative*, comprises the subsequent contributions of Maxwell and those of Lorentz; and the third one, which she calls *philosophical*, that is, a critical reflection on its foundations, encompasses the contributions of Einstein.

To determine the change of an individual concept supposes that a general conception of meaning is available that justifies the existence of an identifiable line of descent among the instances of a concept. In this way, the meta-theoretical problem of establishing the representation or meaning of a concept emerges. Nersessian's early proposal for a general conception of meaning [17.3, Chap. 7] is the following: All the instances of an individual concept fulfill an explana-



**Fig. 17.4** Partial conceptual structure of the medieval theory of motion (after [17.62, p. 171])



**Fig. 17.5** Partial conceptual structure of the Galilean theory of motion (after [17.62, p. 172])

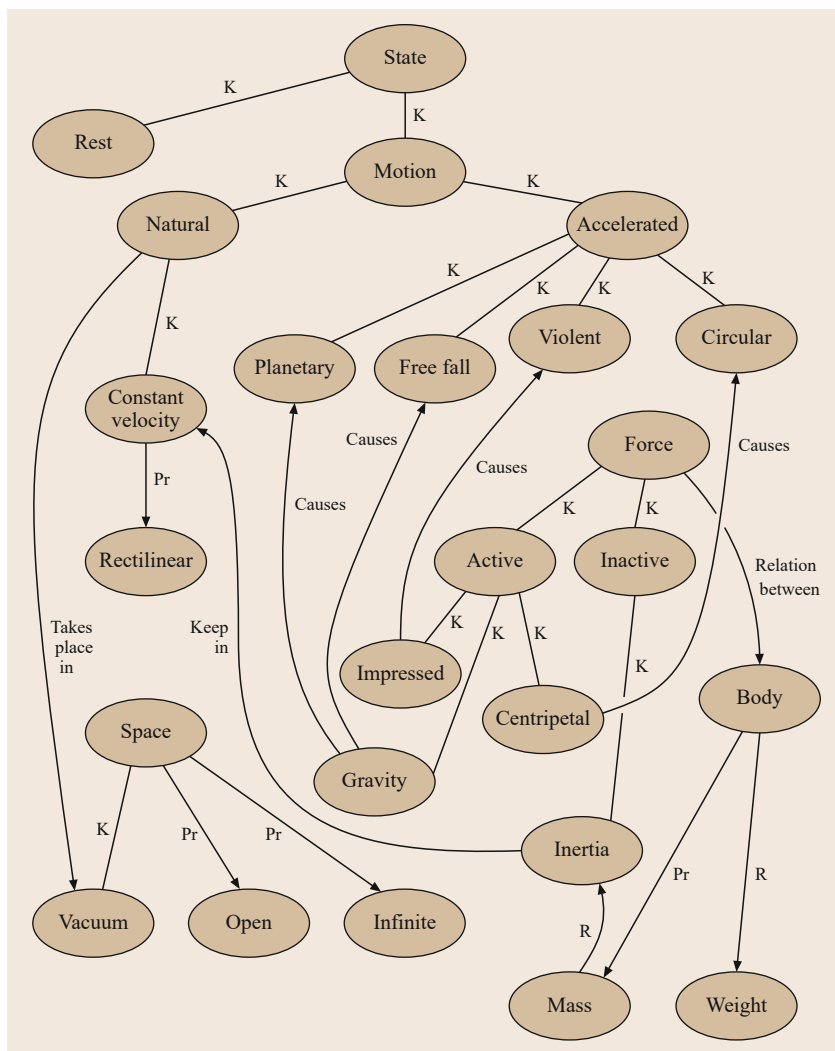
tory/descriptive role in scientific theories. The meaning of a concept or *scheme of meaning* can be understood as a two-dimensional array based on those roles: the dimension that contains a summary of the features of each instance, which can be subsumed under the following factors: *thing, function, structure, and causal power*; and the dimension that contains the development of these features of each instance with the passage of time, which enables one to identify a line of descent between those instances.

In other words, the *instances* of a concept can be represented in both a *synchronic* and a *diachronic* way. Nersessian refers to the synchronic representation of the instances of a concept as a *vector* composed of their salient features. On the other hand, she describes the diachronic representation of the cases of an individual concept as a *vector* expanded to an *array*. Within the array, it is shown how each of the components of the concept's meaning changes over time [17.36, p. 166]. From the representation of the *instances* can arise the representation of a type of concept, which, as I anticipated, is "a set of family resemblances exhibited in

its 'meaning schema'." The representation of a type of concept is an adaptation of the notion of the *prototype* of a concept. Nersessian writes [17.63, p. 161]:

"I have adapted a *prototype* notion of a concept, associated with the work of Eleanor Rosch, to develop a *schema* representation of a scientific concept as an overlapping set of features."

The notion of prototype was elaborated on the basis of the empirical research about concepts carried out by Rosch and her collaborators beginning in 1970. Nersessian found this cognitive view appealing because it enables one to represent the development, the continuity, and the change of concepts in general and of scientific ones in particular [17.36, p. 168]. In fact, the notion of prototype makes it possible to establish a familiarity relationship between the earlier and the later forms of a concept. The probabilistic or prototypes theories about concepts propose that human beings represent a concept by a prototypical example, which is the typical representation of a concept. A prototype in-



**Fig. 17.6** Partial conceptual structure of the Newtonian theory of motion (after [17.62, p. 173])

cludes a list of the features that most probably describe the exemplars of the concept. Some instances of a given concept are better examples than others, depending on the degree in similarity of the object in question to other instances of the concepts or to the prototypical instance. That is why there is a reference to the *graduated structures* of the concepts. Rosch follows Wittgenstein, as, in her conception, a concept is represented by a set of *family resemblances* among the instances placed in the category [17.64]; [17.65, pp. 151–166]. Continuing with the analysis of the example of the concept of *electromagnetic field*, in the three phases of its development, the concept fulfills the role of describing the transmission of electric and magnetic forces and the role of explaining how such continuous and progressive action is possible. On this basis, Nersessian managed to reconstruct the scheme of meaning of the

concept, as summarized in Table 17.1. Here it can be observed that each instance of the concept is linked to the next through *chains of reasoning connections* (COR) (Nersessian borrows this notion from *Dudley Shapere* [17.66]).

In later analyses dealing with the creation of scientific concepts in [17.14], Nersessian discussed briefly the state of the art with respect to the representation of concepts. She wrote that there is no agreement about that, thus [17.14, p. 187]:

“For the present analysis, the format issue can be bypassed by *stipulating* only that whatever the format of a concept, concepts specify constraints for generating members of a class of models.”

Nersessian’s stipulation that concepts specify constraints seems to come from the frame theory about con-

**Table 17.1** Concept of electromagnetic field (after [17.3, p. 158])

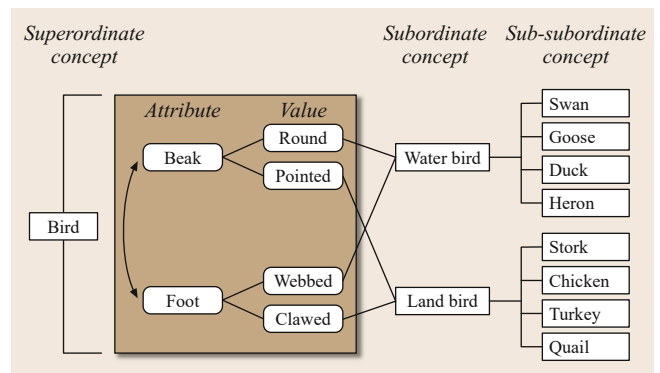
Stuff	Function	Structure	Casual power
Force in region of space (lines are substances? states of <i>aether</i> ?)	Transmits electric and magnetic actions (light? gravity?)	Unknown	Certain electric and magnetic effects
C	C	C	C
O	O	O	O
R	R	R	R
Mechanical processes in quasi-material medium ( <i>aether</i> )	Transmits electric and magnetic actions (now including light)	Maxwell's equations	All electric and magnetic effects, optical effects, radiant heat, etc.
C	C	C	C
O	O	O	O
R	R	R	R
State of immobile <i>aether</i> (nonmechanical)	Same	Same plus Lorentz force	Same
C	C	C	C
O	O	O	O
R	R	R	R
State of space	Same	Same but relativistic interpretation	Same

C  
O – chain-of-reasoning connection  
R

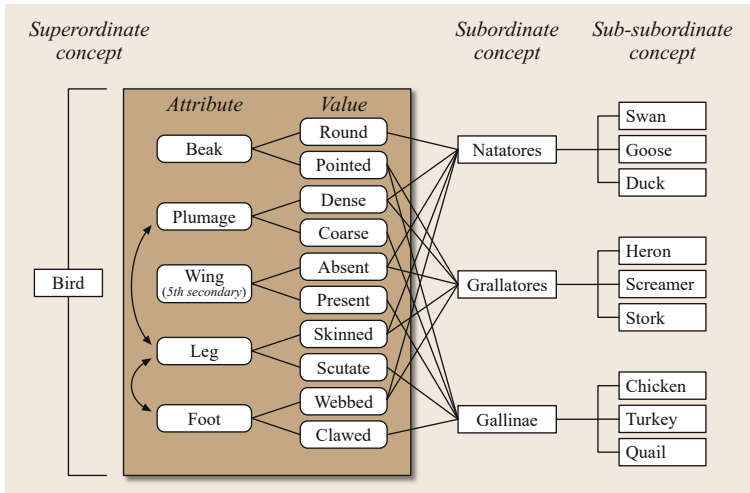
cepts. According to this, the *frame of a concept* is a theoretical representation that organizes all the possible information related to a given concept within a speech community. There are various versions of this conception and one of them is Lawrence Barsalou's *dynamic frames* approach. *Nersessian* judges that this frame perspective has been used successfully in many analysis of the change of taxonomic concepts, but she points out that "[...] the case of science is complicated by the existence of many nontaxonomic concepts, such as 'force' and 'mass'" [17.67, p. 183]. The distinction of taxonomic and nontaxonomic concepts corresponds to the Kuhnian classification between *basic* and *theoretical* concepts [17.68] and, later, to the classification between *normic* and *nomc* concepts [17.69]. While taxonomic, basic or normic concepts are learned by pointing out many of their instances, theoretical or nomic concepts are learned by pointing out complex problem situations to which a law is applied [17.70]. *Nersessian* considers that Barsalou's approach helps to illustrate precisely the idea that concepts specify constraints [17.70]. According to Barsalou, a *frame* is a set of attributes with a multiplicity of values, integrated by structural connections. In general, these attributes hold relationships with each other that are given through the majority of the exemplars of a concept. Barsalou calls these *structural invariants* to these relationships. Furthermore, the values of frame attributes are linked with each other through relationships of dependency. These relations are *constraints*. "Instead, values constrain each other in powerful and complex manners" [17.71, p. 37]; [17.72].

*Nersessian* uses the case expounded by *Hanne Andersen* et al. [17.9, Chap. 4] where, precisely, they employ the notion of dynamic frame – in order to illustrate the idea that concepts specify constraints. The case refers to the representation of the concept of *bird*. This one appears in various ways in the successive taxonomies of Ray, Sundevall, and Gadow. In each one, the corresponding frame of the concept of bird reflects a different set of attributes and different constraints between their values. Figures 17.7–17.9 illustrate the difference of conceptual representation in those three ornithological taxonomies.

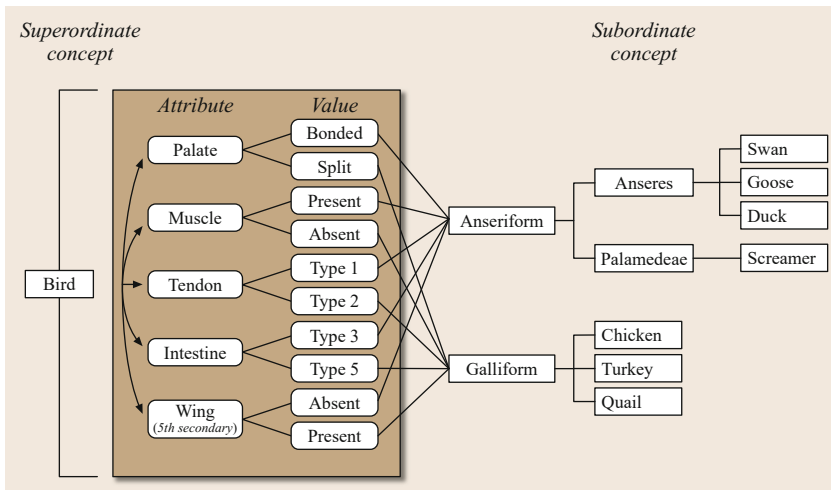
Based on her idea that concepts specify constraints, *Nersessian* describes the concept formation and change as a process of generating new constraints or modify-



**Fig. 17.7** Partial dynamic frame of Ray's concept of bird (1678) (after [17.9, p. 73])



**Fig. 17.8** Partial dynamic frame of Sundevall's concept of bird (1835) (after [17.9, p. 74])



**Fig. 17.9** Partial dynamic frame of Gadow's concept of bird (1893) (after [17.9, p. 88])

ing the existing ones. This construct paved the way to answering philosophical questions such as: How is it that model-based reasoning generates new conceptual representations? And how do models figure in this reasoning and facilitate the reasoning about phenomena?

Model-based reasoning is effective to create new candidate representations because it facilitates the

changes of constraints. By means of the processes of abstraction and integration of constraints from multiple domains in a hybrid model, new combinations of constraints can emerge, and these ones may fit structures and behaviors not represented previously. When scientific change is produced, concepts without precedent in the history of science emerge [17.14, Chap. 6].

## 17.4 Conclusions

From a cognitive-historical approach, Nersessian poses problems relevant to the creation of scientific concepts, thus introducing within the philosophy of science an issue that was traditionally considered not pertinent. Among them, two questions stand out:

1. A fundamental one asks which are the cognitive processes integrated within the environment that scientists develop in forming concepts without precedent in history.
2. The other, derived from the previous one, refers to the reasons that make those mechanisms efficient means to generate new scientific concepts.

As a result, the adoption of a cognitive-historical perspective creates a formulation of the problem that can produce a satisfactory answer. Throughout the chapter, it has been demonstrated that cognitive-historical analysis gathers together features that make it a very valuable tool for that: It makes it possible to obtain information about innovative historical scientific practices; it enables one to study cognitive processes and structures implied in those practices; it aims to investigate creative processes within their own context to avoid distortion; it retrieves data from various kinds of sources; it establishes general conclusions about creative scientific practices; and, finally, it is the method – within cognitive studies – that gives information about historically innovative processes that evolve over long periods of time.

It is possible to appreciate the fertility of the approach through the examination of certain hypothesis about the creation of scientific concepts that Nersessian succeeded in establishing. She obtains information

about historically creative scientific activities and interprets that information in terms of the cognitive sciences. So, on the one hand, she establishes that, through a cycle of bootstrapping modelings, scientists solve representational problems; those modeling processes are genuinely kinds of creative reasoning about a target problem; and model-based reasoning is a heuristic strategy. On the other hand, based on a conception of concepts that stipulates that these specify constraints, she states that the concept formation and change are processes of constraint generation or modification, and she defends the hypothesis that proves that model-based reasoning is effective for creating new candidate representations because they facilitate the change of constraints.

**Acknowledgments.** I wish to thank Alicia E. Gianella for the support she has always given me, and for her inexhaustible willingness to listen to and discuss my embryonic ideas. I am also extremely grateful to the encouragement Nancy Nersessian gave me to continue with my research, when I had the opportunity to meet her at the Conference *Logic, Reasoning and Rationality* held in September 2010 at the Universiteit Gent, Belgium. I highly appreciate the valuable comments of Carlos Oller to a first draft of this chapter, as well as Sandra Meta's help with the final version in English. Finally, I wish to specially mention Lorenzo Magnani for the confidence he granted me, and to the C.I.E.C.E. (Centro de Investigación en Epistemología de las Ciencias Económicas de la Facultad de Ciencias Económicas – Universidad de Buenos Aires), which has given me a space for productive intellectual exchange.

## References

- |   |  |
|---|--|
| <p>17.1 N.J. Nersessian: Cognitive models of science. In: <i>How Do Scientists Think? Capturing the Dynamics of Conceptual Change in Science</i>, ed. by R. Giere (Minnesota Press, Minneapolis 1992)</p> <p>17.2 H. Andersen, K. Hepburn: Scientific change. In: <i>Internet Encyclopedia of Philosophy</i>, ed. by J. Fieser, B. Dowden, <a href="http://www.iep.utm.edu/s-change">http://www.iep.utm.edu/s-change</a> (2011)</p> <p>17.3 N.J. Nersessian: <i>Faraday to Einstein: Constructing Meaning in Scientific Theories</i> (Kluwer, Dordrecht, Boston, London 1984)</p> <p>17.4 T. Arabatzis, V. Kindi: The Problem of conceptual change in the philosophy and history of science. In: <i>International Handbook of Research on Conceptual Change</i>, ed. by S. Vosniadou (Routledge, New York 2008) pp. 345–373</p> | <p>17.5 T.S. Kuhn: <i>The Structures of Scientific Revolutions</i> (Chicago Univ. Press, Chicago 1965)</p> <p>17.6 T.S. Kuhn: What are scientific revolutions? In: <i>The Probabilistic Revolution, Ideas in History</i>, Vol. I, ed. by L. Kruger, L.J. Daston, M. Heidelberger (MIT Press, Cambridge 1987) pp. 7–22</p> <p>17.7 N. Nersessian, H. Andersen: Conceptual change and incommensurability: A cognitive-historical view. In: <i>Danish Yearbook of Philosophy</i> 32, ed. by F. Collin (Museum Tusulanum Press, Copenhagen 1997)</p> <p>17.8 P. Thagard: <i>Conceptual Revolutions</i> (Princeton Univ. Press, Princeton 1992)</p> <p>17.9 H. Andersen, P. Barker, X. Chen: <i>The Cognitive Structure of Scientific Revolutions</i> (Cambridge Univ. Press, Cambridge 2006)</p> |
|---|--|

- 17.10 M. Milkowski, K. Talmont-Kaminski (Eds.): *Regarding the Mind, Naturally: Naturalist Approaches to the Sciences of the Mental* (Cambridge Scholars Publishing, Newcastle upon Tyne 2013)
- 17.11 L. Fleck: *Genesis and Development of a Scientific Fact* (Benno Schwabe, Basel 1935)
- 17.12 R.N. Giere: *Explaining Science: A Cognitive Approach* (Univ. Chicago Press, Chicago 1998)
- 17.13 R. Giere: The cognitive study of science. In: *The Process of Science*, ed. by N.J. Nersessian (Martinus Nijhoff, Dordrecht Boston Lancaster 1987)
- 17.14 N. Nersessian: *Creating Scientific Concepts* (MIT Press, Cambridge London 2008)
- 17.15 T. Nickles: Normal science: From logic to case-based. In: *Thomas Kuhn*, ed. by T. Nickles (Cambridge Univ. Press, Cambridge 2003)
- 17.16 D.W. Gooding: *Experiment and the Making of Meaning* (Springer, New York 1990)
- 17.17 J. Rouse: Understanding scientific practices. Cultural studies of science as a philosophical program. In: *The Science Studies Reader*, ed. by M. Biagioli (Routledge, New York 1999)
- 17.18 P. Thagard: *The Cognitive Science of Science: Explanation, Discovery, and Conceptual Change* (MIT Press, Cambridge 2012)
- 17.19 N. Nersessian: Interpreting scientific and engineering practices: Integrating the cognitive, social, and cultural dimensions. In: *Scientific and Technological Thinking*, ed. by M. Gorman, R. Tweney, D. Gooding, A. Kincannon (Erlbaum, New Jersey 2005)
- 17.20 M. Gorman, A. Kincannon, M.M. Mehalik: Spherical horses and shared toothbrushes: Lessons learned from a workshop on scientific and technological thinking. In: *Discovery Science*, ed. by K.P. Jantke, A. Shinoara (Springer, Berlin, Heidelberg, New York 2001)
- 17.21 B. Latour, S. Woolgar: *Laboratory Life: The Construction of Scientific Facts* (Princeton Univ., Princeton 1979)
- 17.22 N.J. Nersessian: Conceptual change: Creativity, cognition, and culture. In: *Models of Discovery and Creativity*, ed. by J. Meheus, T. Nickles (Springer, Dordrecht, Heidelberg, London, New York 2009)
- 17.23 R. Giere: Cognitive approaches to science. In: *A Companion to the Philosophy of Science*, ed. by W.H. Newton Smith (Blackwell, Oxford 2000)
- 17.24 N.J. Nersessian: Opening the black box: Cognitive science and history of science, *OSIRIS* **10**, 194–211 (1995)
- 17.25 R. Tweney: Cognitive–historical approaches to the understanding of science. In: *Handbook of the Psychology of Science*, ed. by G.J. Feist, M.E. Gorman (Springer, New York 2013)
- 17.26 D. Kulkarni, H.A. Simon: The processes of scientific discovery: The strategy of experimentation, *Cognitive Science*, **125**, 139–176 (1988)
- 17.27 J. Schragar, P. Langley (Eds.): *Computational models of scientific discovery and theory formation*, D. Kulkarni and H. A. Simon, “Experimentation in machine discovery” (Morgan Kaufmann, San Mateo 1990)
- 17.28 M. De Mey: *The Cognitive Paradigm* (D. Reidel, Dordrecht 1982)
- 17.29 H.E. Gruber: *Darwin on Man: A Psychological Study of Scientific Creativity* (Dutton, New York 1974)
- 17.30 A.I. Miller: *Imagery in Scientific Thought: Creating 20th Century Physics* (Birkhauser, Boston 1984)
- 17.31 R. Tweney: Psychology of science and metascience. In: *A Framework for the Cognitive Psychology of Science*, ed. by B. Gholson, A. Houts, R.M. Neimeyer, W. Shadish (Cambridge Univ. Press, Cambridge 1989)
- 17.32 R. Tweney: Discovering discovery: How Faraday found the first metallic colloid, *Perspectives on Science* **14**(1), 97–121 (2006)
- 17.33 E. Cavicchi: Experimenting with magnetism: Ways of learning of Joann and Faraday, *American Journal of Physics* **65**, 867–882 (1997)
- 17.34 E. Cavicchi: Experiences with the magnetism of conducting loops: Historical instruments, experimental replications, and productive confusions, *American Journal of Physics* **71**(2), 156–167 (2003)
- 17.35 U. Feest, F. Steinle (Eds.): *Scientific Concepts and Investigative Practice*, Vol. 3 (de Gruyter, Berlin Boston 2012)
- 17.36 N.J. Nersessian: A cognitive–historical approach to meaning in scientific theories. In: *The Process of Science*, ed. by N.J. Nersessian (Martinus Nijhoff, Dordrecht, Boston, Lancaster 1987)
- 17.37 D.C. Minnen, N.J. Nersessian: Exploring science: The cognition and development of discovery processes, *Am. Psychol. Assoc.* **48**(3), 360–363 (2003)
- 17.38 C. Geertz: *The Interpretation of Cultures: Selected Essay* (Basic books, New York 1973)
- 17.39 G. Ryle: *Collected Papers: Volume II: Collected Essays (1929–1968)* (Hutchinson, London 1971)
- 17.40 J.G. Ponterotto: Brief note on the origins, evolution, and meaning of the qualitative research concept Thick Description, *Qual. Rep.* **11**, 538–549 (2006)
- 17.41 D. Klahr, H.A. Simon: Studies of scientific discovery: Complementary approaches and convergent findings, *Psychol. Bull.* **125**, 524–543 (1999)
- 17.42 D.J. Hess: Ethnography and the development of science and technology studies. In: *Sage Handbook of Ethnography*, ed. by P. Atkinson, A. Coffey, S. Delamont, J. Lofland, L. Lofland (SAGE, Thousand Oaks 2001)
- 17.43 K. Knorr-Cetina: *The Manufacture of Knowledge* (Pergamon, New York 1981)
- 17.44 N.J. Nersessian, W.C. Newstetter, E. Kurz-Milcke, J. Davies: A mixed–method approach to studying distributed cognition in evolving environments, *Proc. ICLS Conf. (ICLS, Seattle 2002)*
- 17.45 M. MacLeod, N.J. Nersessian: Coupling simulation and experiment: The bimodal strategy in integrative systems biology, *Stud. Hist. Philos. Sci. C* **44**, 572–584 (2013)
- 17.46 M. MacLeod, N.J. Nersessian: The creative industry of integrative systems biology, *Mind Soc.* **12**, 35–48 (2013)
- 17.47 R.J. Sternberger, L. Davidson, K. Dunbar (Eds.): *Mechanisms of Insight, How Scientist Really Rea-*

- son: *Scientific Reasoning in Real World Laboratories* (MIT Press, Cambridge 1995)
- 17.48 K. Dunbar: How scientists think: On-line creativity and conceptual change in science. In: *Creative Thought: An Investigation of Conceptual Structures and Processes*, ed. by T.B. Ward, S.M. Smith, J. Vaid (American Psychological Association, Washington 1997)
- 17.49 D. Klahr: *Exploring Science. The Cognition of Development of Discovery Processes* (MIT Press, Cambridge 2010)
- 17.50 M.A. Boden: *The Creative Mind: Myths and Mechanisms* (Routledge, London 2004)
- 17.51 R.N. Giere: Philosophy of science naturalized, *Philos. Sci.* **52**, 331–356 (1985)
- 17.52 N.J. Nersessian: Model based reasoning in conceptual change. In: *Model-Based Reasoning in Scientific Discovery*, ed. by L. Magnani, N.J. Nersessian, P. Thagard (Kluwer Academic, New York 1999)
- 17.53 R.E. Mayer: Problem solving and reasoning. In: *Learning and Cognition in Education*, ed. by V.G. Aukrust (Elsevier, Oxford 2011)
- 17.54 R.W. Weisberg: *Creativity. Understanding Innovation in Problem Solving, Science, Invention and the Arts* (Wiley, New Jersey 2006)
- 17.55 P.N. Johnson-Laird, V. Girotto, P. Legrenzi: Mental models: A gentle guide for outsiders, *Sistemi Intelligenti* **9**, 63–83 (1998)
- 17.56 P.N. Johnson-Laird, R.M. Byrne: *Deduction* (Lawrence Erlbaum, Hillsdale 1991)
- 17.57 L.R. Novick, M. Bassok: Problem solving. In: *The Cambridge Handbook of Thinking and Reasoning*, ed. by K. Holyoak, B. Morrison (Cambridge Univ. Press, Cambridge 2005)
- 17.58 L. Magnani: *Abductive Cognition. The Epistemological and Eco-Cognitive Dimensions of Hypothetical Reasoning* (Springer, Berlin, Heidelberg 2009)
- 17.59 L.M. Osbeck: Transformations in cognitive science: Implications and issues posed, *J. Theor. Philos. Psychol.* **29**, 16–33 (2009)
- 17.60 L.M. Osbeck, K.R. Malone, N.J. Nersessian: Dissenters in the sanctuary evolving frameworks in mainstream cognitive science, *Theory Psychol.* **17**, 243–264 (2007)
- 17.61 L.M. Osbeck, N.J. Nersessian: The distribution of representation, *J. Theory Soc. Behav.* **36**, 141–160 (2006)
- 17.62 N.J. Nersessian: Conceptual change in science and in science education, *Synthese* **80**, 163–183 (1989)
- 17.63 N. Nersessian: Conceptual change. In: *A Companion to Cognitive Science*, ed. by W. Bechtel, G. Graham (Wiley, Malden 1999) pp. 157–166
- 17.64 E. Rosch, C.B. Mervis: Family resemblances: Studies in the internal structure of categories, *Cogn. Psychol.* **7**, 573–605 (1975)
- 17.65 M. Chapman, R.A. Dixon: *Meaning and the Growth of Understanding: Wittgenstein's Significance for Developmental Psychology* (Springer, Berlin, Heidelberg 1987)
- 17.66 D. Shapere: Meaning and scientific change. In: *Mind and Cosmos*, ed. by R.G. Colodny (Univ. of Pittsburgh Press, Pittsburgh 1966) pp. 41–85
- 17.67 N.J. Nersessian: Kuhn, conceptual change, and cognitive science. In: *Thomas Kuhn*, ed. by T. Nickles (Cambridge Univ. Press, Cambridge 2003)
- 17.68 T.S. Kuhn: Metaphor in science. In: *Metaphor and Thought*, ed. by A. Ortony (Cambridge Univ. Press, Cambridge 1979)
- 17.69 T.S. Kuhn: Afterwords. In: *World Changes*, ed. by P. Horwich (MIT Press, Cambridge 1993) pp. 311–342
- 17.70 H. Andersen, N.J. Nersessian: Nomic concepts, frames, and conceptual change, *Philosophy of Science* **67**, 224–241 (2000)
- 17.71 L.W. Barsalou: Frames, concepts, and conceptual fields. In: *Frames, Fields, and Contrasts: New Essays in Semantic and Lexical Organization*, ed. by A. Lehrer, E.F. Kittay (Routledge, New York London 2009)
- 17.72 L. Barsalou, C. Hale: Components of conceptual representation. From feature lists to recursive frames. In: *Categories and Concepts: Theoretical Views and Inductive Data Analysis*, ed. by I. Van Mechelen, J. Hampton, R. Michalski, P. Theuns (Academic Press, Waltham 1993), <http://philpapers.org/rec/VANCAC-5>



# 18. Physically Similar Systems – A History of the Concept

Susan G. Sterrett

The concept of *similar systems* arose in physics and appears to have originated with Newton in the seventeenth century. This chapter provides a critical history of the concept of *physically similar systems*, the twentieth century concept into which it developed. The concept was used in the nineteenth century in various fields of engineering (Froude, Bertrand, Reech), theoretical physics (van der Waals, Onnes, Lorentz, Maxwell, Boltzmann), and theoretical and experimental hydrodynamics (Stokes, Helmholtz, Reynolds, Prandtl, Rayleigh). In 1914, it was articulated in terms of ideas developed in the eighteenth century and used in nineteenth century mathematics and mechanics: equations, functions, and dimensional analysis. The terminology *physically similar systems* was proposed for this new characterization of similar systems by the physicist Edgar Buckingham. Related work by Vaschy, Bertrand, and Riabouchinsky had appeared by then. The concept is very powerful in studying physical phenomena both theoretically and experimentally. As it is not currently a part of the core curricula of science, technology, engineering, and mathematics (STEM) disciplines or philosophy of science, it is not as well known as it ought to be.

The concept of *similar systems* is one of the most powerful concepts in the natural sciences, yet one of the most neglected concepts in philosophy of science today. The concept of similar systems was developed specifically for physics, and its use in biology has generally been in terms of plant and animal physiology; hence, the term *physically similar systems* is often used. It remains an open research question whether, and how, the concept of similar systems might be applied to sciences other than physics, such as ecology, economics, and anthropology.

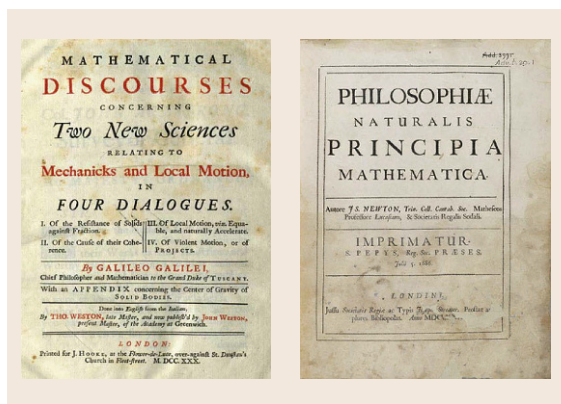
This chapter is devoted to providing a history of the concept of physically similar systems. It also aims, in doing so, to increase the understanding and apprecia-

18.1	<b>Similar Systems, the Twentieth Century Concept</b> .....	379
18.2	<b>Newton and Galileo</b> .....	380
18.2.1	Newton on Similar Systems .....	380
18.2.2	Galileo .....	381
18.3	<b>Late Nineteenth and Early Twentieth Century</b> .....	383
18.3.1	Engineering and Similarity <i>Laws</i> .....	384
18.3.2	Similar Systems in Theoretical Physics: Lorentz, Boltzmann, van der Waals, and Onnes .....	386
18.3.3	Similar Systems in Theoretical Physics ..	391
18.4	<b>1914: The Year of Physically Similar Systems</b> .....	397
18.4.1	Overview of Relevant Events of the Year 1914 .....	398
18.4.2	Stanton and Pannell.....	398
18.4.3	Buckingham and Tolman .....	399
18.4.4	Precursors of the <i>Pi-Theorem</i> in Buckingham's 1914 Papers .....	406
18.5	<b>Physically Similar Systems: The Path in Retrospect</b> .....	408
	<b>References</b> .....	409

tion of the concept of *similar systems* in philosophy. In addition to being neglected in philosophy of science, the concept of *similar systems* is also often not fully understood even when it is mentioned.

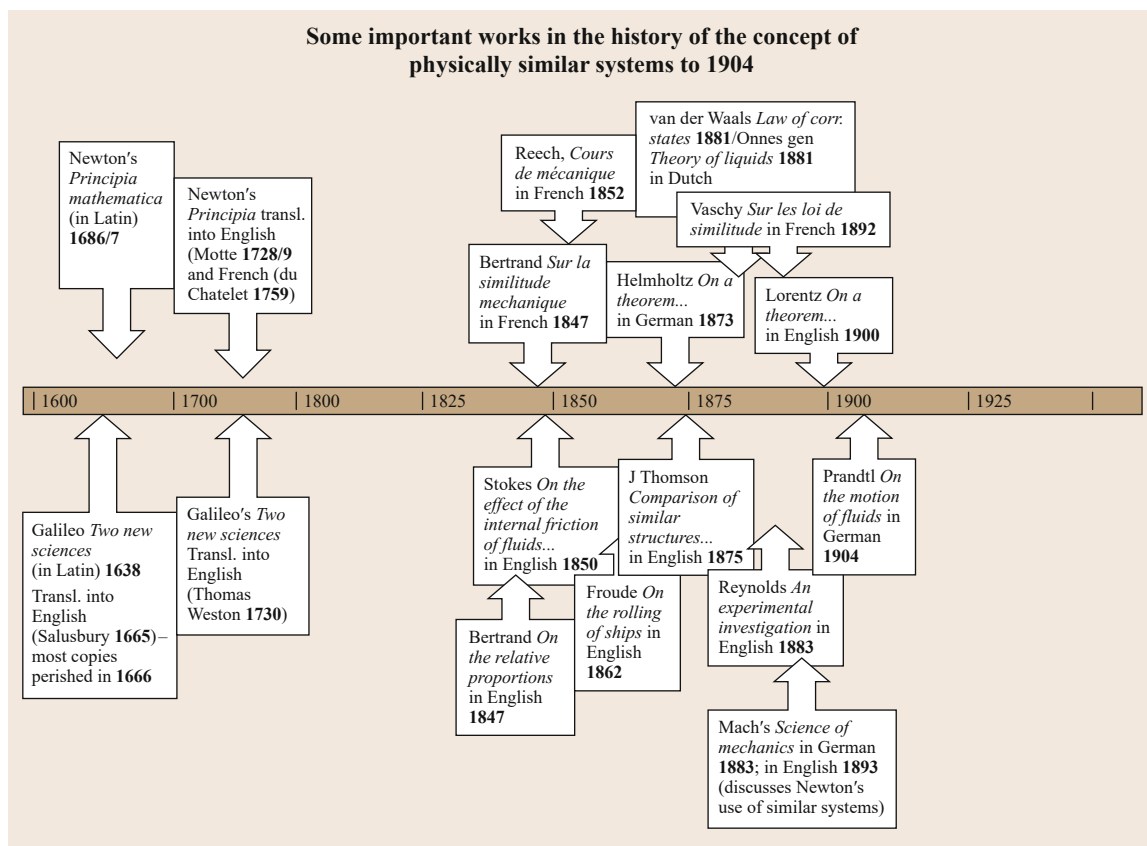
The concept of similar systems has been useful in developing methods for drawing inferences about the values of specific quantities in one system from observations on another system. Some know of the concept only in this derivative way, via applications to specific questions in physics, biology, or engineering.

The fact that it has such useful applications has sometimes led to an underappreciation of the fundamental nature, immense power, and broad scope of the concept. Yet, its utility in practical matters of determin-

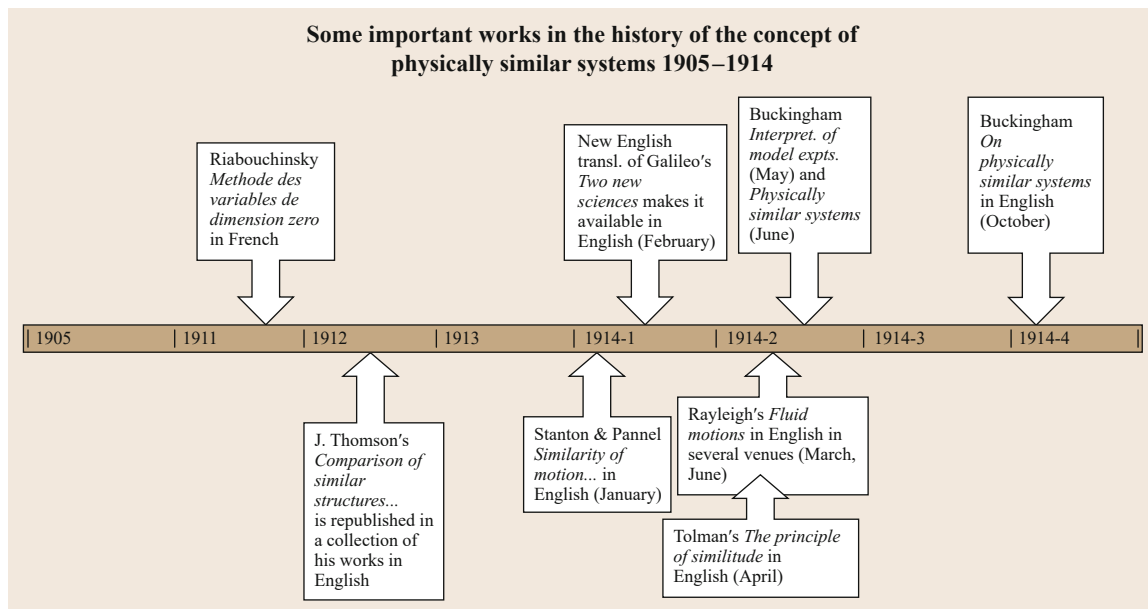


**Fig. 18.1** Newton seems to have been the first to use the term *similar systems* in his *Principia Mathematica*, but Galileo seems to have employed a closely kin idea in his reasoning in *Two New Sciences*

ing or predicting the value of a particular otherwise unobservable quantity is an important feature of the concept. It is due at least in part to the utility of methods involving the concept of similar systems in providing answers to some otherwise intractable problems that natural philosophers in the Renaissance such as Galileo Galilei and Isaac Newton reasoned using some version of the concept (Fig. 18.1), and, later, in the late nineteenth and early twentieth centuries, that scientists further developed it. Thus, the understanding of the concept developed over centuries (Figs. 18.2; 18.3). I will use the twentieth century understanding of *similar systems* to characterize the concept first, then go back to some early precursors from which it was developed and follow the path up to the twentieth century characterization of it. This history of the concept, though admittedly not exhaustively complete, should help clarify its role in reasoning and drawing inferences.



**Fig. 18.2** This timeline (not to scale) illustrates that the concept of similar systems is credited to Renaissance era thinkers Galileo and Newton, and was revived in the second half of the nineteenth century, when it was extended to chemistry, electromagnetic theory, heat, and thermodynamics



**Fig. 18.3** This timeline (not to scale) shows there was a lot of discussion about and interest in issues regarding similarity in 1914 and the years immediately preceding. In 1914 the term *physically similar systems* comes into use

## 18.1 Similar Systems, the Twentieth Century Concept

The landmark year in clarifying and articulating the concept of physically similar systems was 1914. There were two papers with *Physically Similar Systems* in the title that year by *Edgar Buckingham*: one in July (*Physically Similar Systems*) in the *Journal of the Washington Academy of Sciences* [18.1] and another in October 1914 (*On Physically Similar Systems: Illustrations of the Use of Dimensional Equations*) in *Physical Review* [18.2]. Though the latter one is well known and highly cited, and the former one little known, I think that it is the former, that is, the much shorter July 1914 piece, that represents a crucial link or advance, conceptually speaking. The October 1914 Buckingham paper is often credited for the theorem it contains, which is ironic: as Buckingham emphasized numerous times in later papers, a version of the theorem itself had been proven years before. His articulation and discussion of the notion of *physically similar systems*, however, were unusually reasoned and more general than any others accompanying the proof of the theorem.

For now, I just wish to characterize the concept as it is currently understood; for that, we look to the well-known October 1914 *Physical Review* paper [18.2]. The paper opens with a section “The Most General Form of Physical Equations,” which is about describ-

ing a relation that holds among physical quantities of different kinds, by an equation. This is followed by a section introducing and making use of the principle of dimensional homogeneity, entitled “Introduction of Dimensional Conditions.” After exhibiting those points in an example, comes “The General Form to Which Any Physical Equation is Reducible” which states as “a general conclusion from the principle of dimensional homogeneity” that [18.2, p. 350]

“If a relation subsists among any number of physical quantities of  $n$  different kinds, and if the symbols  $Q_1, Q_2, \dots, Q_n$  represent one quantity of each kind, while the remaining quantities of each kind are specified by their ratios  $r', r'', \dots$ , etc., to the particular quantity of that kind selected, then: any equation which describes this relation completely is reducible to the form

$$\Psi(\Pi_1, \Pi_2, \dots, \Pi_i, r', r'', \dots) = 0.”$$

As this form of the equation will be key in defining the notion of similar systems, let us give it a proper name; I’ll call it the *Reduced Relation Equation of 1914*. The number of  $\Pi$ ’s in this equation is the difference between:

“the number of fundamental units required in an absolute system for measuring the  $n$  kinds of quantity, and  $n$ , the kinds of quantity [involved in the relation].”

The function  $\Psi$  is not defined in this form of the equation, but that is perfectly fine; we still consider it an equation – it’s just an equation in which the form of the function is not specified. The equation states, basically, that such a function relating the  $\Pi$ ’s and  $r$ ’s does exist, and the conclusion is that this equation, the *Reduced Relation Equation of 1914*, is another form of the original physical equation, that is, that any physical equation can be reduced to this form. Next follows a short section illustrating how this conclusion can be applied to the same example given earlier in the paper to determine the relationships between some specific quantities in an elegant and particularly useful way. All this is done prior to, and independently of, defining the notion of physically similar systems.

It is in the section entitled “Physically Similar Systems,” the sixth section of the paper, that the notion of similar systems is first presented. Referring to the equation in his paper shown above, which I have called

the *Reduced Relation Equation of 1914*, Buckingham writes that “we may develop from it the notion of similar systems;” he develops it as follows [18.2, p. 353]:

“Let  $S$  be a physical system, and let a relation subsist among a number of quantities  $Q$  which pertain to  $S$ . Let us imagine  $S$  to be transformed into another system  $S'$  so that  $S'$  corresponds to  $S$  as regards the essential quantities. There is no point of the transformation at which we can suppose that the quantities cease to be dependent on one another: hence we must suppose that some relation will subsist among the quantities  $Q'$  in  $S'$  which correspond to the quantities  $Q$  in  $S$ . If this relation in  $S'$  is of the same form as the relation in  $S$  and is describable by the same equation, the two systems are *physically similar* as regards this relation.”

This is the notion of *physically similar systems* still currently in use today. It was first articulated in 1914 by the physicist Edgar Buckingham. But it did not arise from Buckingham’s cogitations out of the blue. For its precursors, we have to go back to the Renaissance.

## 18.2 Newton and Galileo

### 18.2.1 Newton on Similar Systems

Newton seems to have been the first to use the term *similar systems*. He uses it more than once, but the text usually associated with the concept of similar systems is in Book 2, Proposition 32, where he writes [18.3, p. 327]:

“Suppose two similar systems of bodies consisting of an equal number of particles, and let the correspondent particles be similar and proportional, each in one system to each in the other, and have a like situation among themselves, and the same given ratio of density to each other; and let them begin to move among themselves in proportional times, and with like motions (that is, those in one system among one another, and those in the other among one another.) And if the particles that are in the same system do not touch one another, except in the moments of reflection; nor attract, nor repel each other, except with accelerative forces that are inversely as the diameters of the correspondent particles, and directly as the squares of the velocities: I say, that the particles of those systems will continue to move among themselves with like motions and in proportional times.”

In his *Science of Mechanics*, Mach refers to Newton’s concept of similar systems in the context of his own discussion of oscillatory motion [18.4, p. 203]. Mach’s critical-historical work on mechanics was written to be accessible to the nonspecialist; his critique is informative of the understanding of similarity and similar systems at that time. After generalizing one of his own conclusions, Mach remarks: “The considerations last presented may be put in a very much abbreviated and very obvious form by a method of conception first employed by Newton.” He does not quite accept Newton’s use of the term similar system there, though [18.4, p. 166]

“Newton calls those material systems *similar* that have geometrically similar configurations and whose homologous masses bear to one another the same ratio. He says further that systems of this kind execute similar movements when the homologous points describe similar paths in proportional times.”

Mach admires Newton’s methodology here, but he points out an issue with Newton’s use of the term *similar* [18.4, p. 166]:

“Conformably to the geometrical terminology of the present day we should not be permitted to call me-

chanical structures of this kind (of five dimensions) *similar* unless their homologous linear dimensions as well as the times and the masses bore to one another the *same* ratio.”

I gather that what Mach is saying is that the notion of similar in use at the time he is writing is the notion of geometrical similarity, in which there is a kind of shrinking or enlarging of *every* linear quantity of *each* dimension by the *same* ratio (for geometrical similarity, there would usually not be more than three dimensions). That is, I believe he means that, if we are talking about a three-dimensional machine, similarity amounts to shrinking or enlarging quantities of each linear dimension *by the same ratio* while keeping the machine and all its parts exactly the same shape, that is, while preserving every ratio of linear quantities within the same machine. Now, of course, areas and volumes will bear a different ratio to their homologues than quantities of the linear dimensions do (e.g., if the ratio is 1 : 3 for the linear dimension, it will be 1 : 9 for an area and 1 : 27 for a volume), but the similarity can be defined in terms of the linear dimensions alone. That is how geometrical similarity works. Mach says, I think that a strict application of the notion of geometric similarity would require that the ratio between a quantity and its homologous quantity be the same for all five of the dimensions that Newton mentions for his case, and that the situation imagined in Newton’s proposition does not satisfy that constraint.

However – and what is significant and interesting – *Mach* does not say that Newton is wrong here; rather, what he says is that what Newton was doing is better understood in Mach’s day in terms of affine transformations [18.4, p. 204]:

“The structures might more appropriately be termed *affined* to one another.

We shall retain, however, the name phoronomically [kinematically] *similar* structures, and in the consideration that is to follow leave the masses entirely out of account.”

It is clear that *Newton* was interested in more than this, that he wanted to employ the notion of similar systems to reason about forces, too; in fact, he does so in the remarks that follow the quote above ([18.3, pp. 327–328], [18.5, pp. 766–768]). However, in leaving the masses out of the account, Mach picks out from Newton’s work what he wishes to endorse, and shows how the points he endorses ought to be understood in the terminology of the nineteenth century. *Mach* shows how to understand phoronomically (kinematically) similar structures for the topic of oscillation he has been discussing [18.4]:

“In two such similar motions, then, let the homologous paths be  $s$  and  $\alpha s$ , the homologous times be  $t$  and  $\beta t$ ; whence the homologous velocities are  $v = s/t$  and  $\alpha v = \alpha/\beta s/t$ , the homologous accelerations  $\phi = 2s/t^2$  and  $\epsilon\phi = \alpha/\beta^2 2s/t^2$ .

Now all oscillations which a body performs under the conditions above set forth with any two different amplitudes 1 and  $\alpha$ , will be readily recognized as *similar* motions.”

Thus, in spite of noting that *similar* generally means *geometrically similar* at the time he was writing, Mach indulges Newton in the use of the adjective *similar* to indicate phoronomically (kinematically) similar structures, which are, properly speaking (in the terminology of Mach’s day), not related by *similarity* but by *affinity* (that is, by affine transformations). After showing how elegantly theorems about centripetal motion can be obtained by such means, he remarks [18.4, p. 205]:

“It is a pity that investigations of this kind respecting mechanical and phoronomical *affinity* are not more extensively cultivated, since they promise the most beautiful and most elucidative extensions of insight imaginable.”

Thus, Mach sees the great power of the notion of similar systems. In terms of clarification of the notion itself, though, which is the topic of this article, Mach’s attention in his critique of Newton is on the *similar* in *similar systems*; he does not here discuss criteria for something counting as a *system*.

Newton is recognized for the concept today, as he has been throughout all of the nineteenth and twentieth centuries. In their *Similarity of Motion in Relation to the Surface Friction of Fluids* paper in early 1914, *Stanton* and *Pannell* credit George Greenhill with pointing out that the idea that relations “applicable to all fluids and conditions of flow” existed was “foreshadowed by Newton in Proposition 32, Book II of the *Principia*” [18.6, p. 199]. *Zahm*’s 1929 report *Theories of Flow Similitude* [18.7] also credits Newton for a method of “dynamically similar systems,” citing Newton’s Propositions 32 and 33. Also, in many more recent works, including [18.8, p. 86ff], [18.9, pp. 39–41], and [18.5, p. 766].

### 18.2.2 Galileo

Although Newton seems to have been the first to use the term *similar systems*, *Galileo*’s reasoning certainly used a notion of similar systems akin to, if not prescient of, Newton’s in discussing not only the motions of the bob

of a pendulum, but also the more complicated behavior of machines and structures with mass; this is especially clear in his *Dialogues Concerning Two New Sciences*. Galileo's dialogue begins with Salviati (usually taken to be the voice of Galileo), recounting numerous examples of a large structure that has the same proportions and ratios as a smaller structure but that is not proportionately strong. In these opening pages of the dialogue, Salviati explains to a puzzled Sagredo that "if a scantling can bear the weight of ten scantlings, a [geometrically] similar beam will by no means be able to bear the weight of ten like beams" [18.10, m.p. 52–53]. The phenomenon of the effect of size on the function of machines of similar design holds among natural as well as artificial forms, Salviati explains: "just as smaller animals are proportionately stronger or more robust than larger ones, so smaller plants will sustain themselves better" [18.10, m.p. 52–53].

Perhaps the most well known of Salviati's illustrations is about giants [18.10, m.p. 52–53]:

"I think you both know that if an oak were two hundred feet high, it could not support branches spread out similarly to those of an oak of average size. Only by a miracle could nature form a horse the size of twenty horses, or a giant ten times the height of a man – unless she greatly altered the proportions of the members, especially those of the skeleton, thickening the bones far beyond their ordinary symmetry."

Although Galileo's work opens with the wise participant in the dialogue reminding the others of the reasons for the lack of giant versions of naturally occurring life forms, it soon proceeds to the case of a *valid* use of a small (artificial) machine to infer the behavior of a large (artificial) machine. But *the basis for the similarity is not merely geometric similarity*. Later in this same work of Galileo's, Sagredo makes use of Salviati's statement that the "times of oscillation" of bodies [18.10, m.p. 139] suspended by threads of different lengths "are as the square roots of the string lengths; or we should say that the lengths are as the doubled ratios, or squares, of the times." From this, Sagredo uses one physical pendulum to infer the length of another physical pendulum [18.10, m.p. 140]:

"Then, if I understood correctly, I can easily know the length of a string hanging from any great height, even though the upper attachment is out of my sight, and I see only the lower end. For if I attach a heavy weight to the string down here, and set it in oscillation back and forth; and if a companion counts a number of its vibrations made by another move-

able hung to a thread exactly one braccio in length, I can find the length of the string from the numbers of vibrations of these two pendulums during the same period of time."

The reasoning that Sagredo uses to infer the length of one pendulum (the larger) from another (the smaller) is based upon the constancy of the value of a certain ratio involving the length and the frequency of a pendulum's oscillations. What Sagredo derives from the constancy of that ratio for all pendulums is a *law of correspondence* telling him how to find the corresponding length in the large pendulum from the length of the small (or vice versa) and the number of oscillations of the two pendulums observed during the same time period. (The time period itself during which the oscillations are observed is not needed; what is needed is only the (square of the) ratio of the number of oscillations of the two pendulums.) He works out an example [18.10, pp. 140]:

"[...] let us assume that in the time my friend has counted twenty vibrations of the long string, I have counted two hundred forty of my thread, which is one braccio long. Then after squaring the numbers 20 and 240, giving 400 and 57 600, I shall say that the long string contains 57 600 of those units [misure] of which my thread contains 400; and since my thread is a single braccio, I divide 57 600 by 400 and get 144, so 144 braccia is the length of the string."

Salviati (the voice of Galileo) responds approvingly to Sagredo's claim that this method will yield the length of the string: "Nor will you be in error by a span, especially if you take a large number of vibrations." This is reasoning much like Newton's use of similar systems, in that one pendulum is regarded as being similar to another pendulum, so that the period of oscillation and length of one of the pendulums is homologous to the period of oscillation and length of the other.

Of course, Galileo's reasoning here is not presented as a general method, as it is specific to pendulums, whereas Newton's notion of similar systems is. Nor do we find in Galileo's discussion here any explicit criteria for something being a machine that could serve to delineate the sorts of things on which this kind of reasoning could be used. However, Galileo's discussion does make clear that the two quantities that are considered homologous – the *time of vibration* and the length of the pendulum string – are fixed features of a pendulum, in contrast to other quantities such as the amplitude of the oscillations, or the weight of the bob [18.10, p. 141]:

"Take in hand any string you like, to which a weight is attached, and try the best you can to increase or

diminish the frequency of its vibrations; this will be a mere waste of effort. On the other hand, we confer motion on any pendulum, by merely blowing on it [...] This motion may be made quite large [...] yet it will take place only in accord with the time appropriate to its oscillations.”

Thus, each of the two quantities – length of the string, time of vibration – of a given pendulum determines the other. The point germane to the topic of the history of similar systems, though, is this: *every pendulum is related to every other pendulum* by a law of correspondence. The law of correspondence relates each of these two quantities in one pendulum to its homologue in another pendulum. I think that we can see this as akin to how Newton conceived similar systems to be related: by a law of correspondence between quantities in one system and their homologous quantities in the similar system. Only the length of the string and the time of vibration show up as homologous properties in comparison of the two pendulums. Thus, Galileo makes a point of distinguishing between quantities that characterize a given pendulum (length of string; time of oscillation) and quantities that do not (amplitude of oscillation; weight of bob), in addition to making the point about how *some* behaviors of *all*

pendulums are related to each other by a *law of correspondence*.

Because the point is so often missed, it may be helpful to state it a slightly different way. Clearly, Galileo sees that in a pendulum’s behavior, the quantities that characterize a pendulum’s behavior are related to each other in a fixed (though nonlinear) relation, as evidenced by his remarks about the time of oscillation of a pendulum being determined by the length of its string. Yet, rather than illustrating that one can use this relation to figure out the value of one quantity associated with a certain pendulum by measuring another quantity associated with *that same pendulum*, what Galileo is doing here is using a completely different method of inference: establishing a *law of correspondence* between two different pendulums. Then, from an observation of one quantity obtained experimentally on another pendulum chosen or constructed for the purpose, the law of correspondence he has established is invoked to *infer* the value of *the homologous quantity* in the pendulum. (In the passage from Galileo quoted above, the method was used to infer the length of one pendulum from the length of another pendulum.) It is the articulation of this method that justifies including Galileo along with Newton in a history of the concept of physically similar systems [18.11].

## 18.3 Late Nineteenth and Early Twentieth Century

By the late nineteenth century, mechanics and the mathematics used in it had changed dramatically from Newton’s – at least in terms of many of the mathematical methods used. The concept of mechanical similarity survived these major changes, though, and quite easily accommodated itself to the more advanced mathematics developed for mechanics. In fact, the notion of mechanical similarity was developed further, and more rigorously, into different kinds of similarity in mechanics – geometrical, kinematical, and dynamical – and extended to other areas of physics that had become quantitative, such as heat and electricity. The concept of similar systems survived, too, sometimes explicitly, sometimes only implicitly and in practice. More problematically, during the nineteenth century, the term was sometimes used to refer to something other than the rigorous notions associated with the term that were being developed in physics.

The advances in mathematics and physics to which the concept of similar systems and, along with it, the concept of similarity, were rather easily incorporated were not merely superficial matters such as the use of a different notation for calculus. By the late nineteenth

century, there was widespread use of the more advanced mathematical methods that had been developed: partial differential equations and associated analysis methods for continuum mechanics, hydrodynamics, gas theory, electricity, and magnetism. During the eighteenth century, there had been many advances in mathematics and mechanics that transformed the methods of inquiry used into ones we would be at home with even today. The question of what constitutes a system shifts from asking not only how to decide when a configuration of bodies constitutes a system (Newton and Galileo seem to have thought in terms of systems of that sort), to also being able to ask what features of a function (or equation) indicate that the relations between quantities that it expresses have also delineated a system. For, it is functions that the eighteenth century gave mechanics, and functions represented or expressed “relations among quantities in nature,” as *Hepburn* puts it [18.12, p. 129]. As noted in Sect. 18.1, when Buckingham articulated the concept of *physically similar systems* in 1914 [18.2], he did so by providing the “most general form of an equation,” and, as seen in the excerpt quoted above, he did so by describing that form in terms of an

equation using an unknown function

$$\Psi(\Pi_1, \Pi_2, \dots, \Pi_i, r', r'', \dots) = 0,$$

that is, the equation I have called the *Reduced Relation Equation of 1914*.

*Buckingham* did his doctoral work at the very end of the nineteenth century. Where were people employing or talking about the notion of similar systems during the late nineteenth century? By then, some notion of similar systems was known in theoretical physics, where it was occasionally explicitly discussed using the term *similar systems*, as well as in many branches of engineering, where it was involved, albeit sometimes implicitly or obliquely, in experimental investigations. Then, too, there were activities and investigations that did not fit neatly into one or the other of these categories, or straddled them. How did various thinkers producing these works think about and express the concepts associated with mechanical similarity and similar systems?

### 18.3.1 Engineering and Similarity Laws

#### Similar Structures

In engineering and science of the nineteenth century, the main notion invoked when reasoning with similar machines or systems was that of a *similarity law* or a *similarity principle*. *James Thomson* (1822–1892) (brother of *William Thomson*, *Lord Kelvin* (1824–1907)) gave an influential paper in 1875 entitled *Comparison of Similar Structures as to Elasticity, Strength, and Stability* [18.13] that tried to identify and lay out the methodology involved in the engineering design of structures such as bridges and buildings, but he used some other interesting examples such as obelisks and umbrellas, too. Thomson's examples are often about how to vary some quantity such that two structures of different sizes are similar in one of these respects I refer to as behavioral: that is, elasticity, strength, or stability. Thomson's paper was built upon and expanded in 1899 (by *Barr* [18.14]) and again in 1913 (by *Torrance* [18.15]).

The principle *James Thomson* identified was meant to be general. Yet, there were still different *kinds* of comparisons. In his 1875 paper, which became more widely available when his collected works were published in 1912, Thomson distinguished between two kinds of comparisons of similar structures, which, he said, were “very distinct, and which stand remarkably in contrast each with the other.” One kind of comparison of similar structures is “in respect to their elasticity and strength for resisting bending, or damage, or breakage by similarly applied systems of forces.” The other, contrasting kind was “comparisons of similar structures as

to their stability, when that is mainly or essentially due to their gravity [weight] or, as we may say, to the downward force which they receive from gravitation” [18.16, p. 362].

*Thomson* offered a “comprehensive but simple and easily intelligible principle” for the first kind of comparison: Similar structures, if strained similarly within limits of elasticity from their forms when free from applied forces, must have their systems of applied forces, similar in arrangement and of amounts, at homologous places, proportional to the squares of their homologous linear dimensions. His reasoning in establishing this principle is a deductive argument special to solid mechanics, the mechanics of deformable bodies. To establish this we have only to build up, in imagination, both structures out of similar small elements or blocks, alike strained, with the same intensity and direction of stress in each new pair of homologous elements built into the pair of objects [18.16, pp. 362–363]. These small elements or blocks are imagined to be so small in relation to the overall body that the stresses in them can be considered homogenous throughout the element or block. This is how the principle is derived, but the point of emphasis for both scientific understanding and engineering practice was that “similar structures of different dimensions must not be similarly loaded [...] if they are to be stressed with equal severity.” In saying that the structures must not be similarly loaded, he draws attention to the part of the principle that says that the loads in the two similar structures must vary by the squares of their linear dimensions, rather than by the simple multiplicative factor that the linear dimensions do.

This was commonly what was meant at the time by a *similarity principle* or, sometimes *similarity law* or *law of similarity*. Each one covered a certain class of cases. The point of the *principle* was usually to state how one variable – for example, density, stiffness – was to be varied as another, such as length, was varied. One form such reasoning took was to show how the ratio of variables of one type varied as a ratio of another type of variable did: for instance [18.17, p. 136],

“If the scale ratio for any two orifices, i. e., the ratio of any two corresponding linear dimensions, is  $S$ , the ratio of the areas of corresponding elements of the orifices will be  $S^2$ , while if similarly situated with respect to the water surface, their depths are proportional to  $S$ .”

However, sometimes the similarity law or principle for a certain kind of behavior was stated simply as a ratio, the implication being that that ratio was invariant for similar systems; setting the ratio equal to 1 and rearranging terms yielded the relations between



quantities that must be maintained in order to achieve the similarity of that type.

### Similar Interactions: A Law of Comparison for Model Ships

One of the most well-known engineering advances employing similarity and, implicitly, the notion of similar systems, was William Froude's (1810–1879) solution of significant, urgent, and previously unsolved problems in ship design for the British Admiralty ([18.18, p. 279], [18.5, 19, 20]). In the design of ships for stability and speed, not only does gravitational force enter into the consideration of a structure's behavior, but the ship's interaction with the water in which it is sitting or moving must also be considered.

Froude's reasoning about the stability of ships involved examining the motion of a pendulum in a resistive fluid [18.21, pp. 5ff, 15ff, 61]: the same question Newton addressed when he presented the proposition in which he introduced the idea of similar systems. Schaffer points out that, although the statement does not appear in the final version of the *Principia*, Newton had written that “if various shapes of ships were constructed as little models and compared with each other, one could test cheaply which was best for navigation” [18.22, p. 90].

Unlike Newton, Froude does not seem to analyze the notion of similar systems in thinking about a pendulum in a resistive medium. However, the idea of relating quantities in one physical situation to those in another is predominant in Froude's work; it is, in fact, the topic of his main contributions to the problem of the efficient design of large ships driven by propellers. As Zwart has pointed out [18.5], the naval architect John Scott Russell had already constructed and tested many small models, but his experience had convinced him that the little models, though they had provided him with much pleasure, could provide no help in determining how large ships behaved. The exchange between Russell and Froude following Froude's reading of his 1874 paper was recorded in a transcript and so is available today [18.23], showing that the problem of how to extrapolate observations on the behavior of small models of ships when placed in water to the behavior of full size ships was considered unsolved when Froude took it on ([18.5, p. 15], [18.20, pp. 128–130], [18.19, 23]). Hagler also notes that Froude's confidence that the smaller model ships (some of which were over 20 feet long) could be used to infer the behavior of larger full-scale ships was based in part on Rankine's investigations on streamlines. Froude explicitly discusses Rankine's work in his 1869's “The State of Existing Knowledge on the Stability, Propulsion and Seagoing Qualities of Ships” [18.20]. He convinced the Admiralty to fund

the construction of an experimental water tank to carry out the experiments he proposed. His methods for extrapolating from smaller, scale models of ships in his water tank to the full size ship were vindicated when the Admiralty conducted full-scale tests on the *HMS Greyhound* and Froude was able to compare the measurements taken on the full size *Greyhound* with those he had taken on his 1/16 model of the *HMS Greyhound* in his experimental tank. His *Law of Comparison* was soon adopted for all further ship design not only by the British Admiralty, but also by the US Navy, which constructed the Experimental Model Basin in Washington, DC in the 1890s. The Experimental Model Basin was constructed under the leadership of David Watson Taylor. Hagler [18.20] provides a good discussion of David Watson Taylor's writings on ship design; Taylor shows how the methodology used by the US in almost all its naval design work in the first half of the twentieth century is ultimately traceable to this work Froude did in the nineteenth century.

Froude similarity was developed specifically for the purpose of using model experimentation for ship design. As with the similarity laws in mechanics, Froude similarity can be expressed in terms of a ratio, the Froude number, which is a dimensionless parameter. Though no notion of similar systems is defined, a nascent notion of similar systems was involved in practice, since the similarity of situations is established when the Froude numbers for each of the two situations are equal. One formulation of the Froude number is  $v/\sqrt{gL}$ , where  $v$  is a velocity,  $L$  is a length, and  $g$  is the gravitational acceleration. The application of Froude similarity requires expertise; which velocity and characteristic length are relevant depends on the phenomenon being investigated. We can see from the form of the Froude dimensionless ratio, however, that quantities do not all scale linearly, much less by the same linear factor. Another point of note is that, as Froude similarity compares *homologous forces* as well as *homologous motions*, it is a kind of *dynamic* similarity, not merely a *kinematic* similarity.

### Bertrand and Reech: The French Connection Between Newton and Froude

Many have pointed out that Froude took over results due to others, naming in particular French engineering professor Ferdinand Reech and French mathematician Joseph Bertrand, both of whom wrote on similarity methods in mechanics ([18.24, p. 141ff], [18.25, p. 381], [18.26, p. 15], and [18.18, p. 279]). The extent to which this is true has been debated [18.24], but none deny that Froude holds a unique place as an experimentalist whose accomplishments advanced both the field of hydraulics and the industry of marine architecture.

Ferdinand Reech (1805–1884) publishing in 1852 on topics he had lectured about much earlier, explicitly followed Newton’s approach, discussing and deriving principles about how to relate observations of velocities and motions of one ship to other ships of different sizes. Like Newton, he considered bodies and forces on them, though he employed the term *similar system* in his discussions when deriving laws of comparison [18.27]. It is Joseph Bertrand who seems to have taken a conceptual step beyond Newton, though he heaps quite a great deal of credit for his work upon Newton, as though he is doing little more than showing the consequences of Newton’s theorems about similar systems.

*Joseph Bertrand* (1822–1900) produced many textbooks and treatises, including *Sur la similitude en mecanique*. He also published, in English in 1847, a sort of manifesto advocating that “persons occupied with the study of mechanics” attend to the theorem about similitude he derives using nineteenth century methods in mechanics, but for which he credits Newton. Of Newton’s theorem about similar systems in the *Principia*, he writes [18.28, p. 130]

“This theorem constitutes a real theory of similitude in mechanics. It will be seen, that any system being given, there exists an infinite number of possible systems, which may be regarded as similar to it; and that, instead of a single kind of similitude, as in geometry, we may suppose four, viz., those of length, time, forces, and masses; each of these is, according to Newton’s theorem, a consequence of the other three.”

*Bertrand* then went on in that same paper of 1847 to explain that he had

“endeavoured to substitute [...] a proposition founded upon dynamic equations, and which does not differ mainly from the form employed by M. Cauchy to deduce from the equations of the movement of elastic bodies the laws of the vibrations of similar bodies, [...] but this theorem of M. Cauchy, although analogous to that of Newton, cannot be regarded as a corollary of the same;”

using this instead, he deduces applications to laws of oscillation, centripetal force, speed of propagation of sound in various gases, and “a theorem relating to turbines” [18.28, p. 130]. *Bertrand*’s concern seems to be twofold: (i) to get people in the field of mechanics to appreciate the power of the theory (or principle) of similitude in providing solutions to otherwise insoluble problems, and, (ii) to get people who use model experiments to understand the appropriate precautions that must be taken in designing experiments using small models to prevent errors that can be anticipated us-

ing the theory. He explains how the notion of similar systems, though it may look rather limited, is in fact sometimes indispensable, that is, for problems not susceptible to a mathematical solution [18.28, p. 131]:

“It is true that only proportional results can be deduced from [the principle]; and that, consequently, it will only serve to solve a question, when another of an analogous nature and of an equivalent analytical difficulty shall have been solved. It may, however, be of great utility to determine in certain cases the analogy which exists between the movements of the two systems, even supposing each of them not to be susceptible of strict theoretical determination.”

He gives an example of the usefulness of the principle: the performance of “experiments on a small scale” to ascertain “the value of a mechanical invention, which is too expensive to put in operation on a large scale” [18.28, p. 131]. What is interesting is that in this same paper where he is advocating the use of the principle, he also discusses the kind of conundrums that arise in attempting to apply it to complicated cases such as a small-scale model of a locomotive; he cites an example of “an error which it is impossible to avoid, but which it is very essential to know.” This 1847 paper published in England is, thus, a call to improving engineering practice by attending to theoretical derivations in mechanics, that is, the theory of similitude. (*Bertrand* refers to it in the 1847 paper as the Cauchy theorem, which seems rather modest, for *Cajori* describes *Bertrand* as deriving “the principle of mechanical similitude” from “the principle of virtual velocities” [18.25, p. 380]. I mention *Bertrand*’s 1847 paper here for its use of late eighteenth and nineteenth century mechanics.)

### 18.3.2 Similar Systems in Theoretical Physics: Lorentz, Boltzmann, van der Waals, and Onnes

Mechanical similarity held an important place among some researchers in theoretical physics in the late nineteenth century as well. The notion of similar systems was often employed in theories about the relationship of microscopic configurations to macroscopic phenomena, sometimes explicitly. Sometimes the term *similar systems* was extended beyond the normal use it had had up to that time, too.

#### Lorentz

By the turn of the century, *Henrik Lorentz* (1853–1928) would note that “The consideration of similar systems

has already proved of great value in molecular theory,” as it had allowed Kamerlingh Onnes “to give a theoretical demonstration of van der Waals’s law of corresponding states” [18.29]. The experimental confirmation of that law, Lorentz wrote, “has taught us that a large number of really existing bodies may, to a certain approximation be regarded as similar.”

Lorentz had already developed a notion of corresponding states for use in electrodynamics by 1900. The context in which he made the observation above, though, was his paper *The Theory of Radiation and the Second Law of Thermodynamics*, in which he was concerned with the question of the similarity in the structure of different bodies that would be mandated by thermodynamics [18.29, p. 440]. It would take us too far afield to explain everything that Lorentz was trying to do in this paper; here we restrict our discussion to what concept of *similar systems* Lorentz employed or seems to have had in mind.

Lorentz’ idea of *similar systems* involves starting with one system and then constructing a second one from the first. Lorentz writes of “comparing two systems”; what he says is that the systems he compares are: “[...] in a wide sense of the word, *similar*, that is, such that, for every kind of geometrical or physical quantity involved, there is a fixed ratio between its corresponding values in the two systems, [...]” [18.29]. It is not clear on what basis he justifies being able to say that “We shall begin by supposing that, in passing from one system to another, the dimensions, masses and molecular forces may be arbitrarily modified,” as this seems to require a certain kind of independence among the things being modified. He argues that “if the second system, as compared with the original one, is to satisfy Boltzmann’s and Wien’s laws,” that “we shall find that the charges of the electrons must remain unaltered.”

He first describes a certain system  $S$  that includes a “ponderable body” enclosed in a space. Some of the features of  $S$  are delineated (he ascribes an “irregular molecular motion” and the “power of acting on one another with certain molecular forces” to the particles making up the body, for instance, and adds that some are electrically charged) but other features are not (there may be other (molecular) forces of another kind, acting on the electron) [18.29, p. 443]. The description of the “really existing” system  $S$  is meant to pick out something that actually exists, in contrast to the system  $S'$ , which “perhaps will be only an imaginary one” [18.29, p. 444]. To complete the description of the state of  $S'$ , “we indicate, for each of the physical quantities involved, the number by which we must multiply its value in  $S$ , in order to obtain its value in  $S'$  at corresponding points and times.” He then explores the constraints on

these numbers; some are constrained by laws of motion, but others are not. This leaves him free to “imagine a large variety of systems  $S'$ , similar to  $S$ , and which must be deemed possible as far as our equations of motion are concerned” [18.29, p. 445].

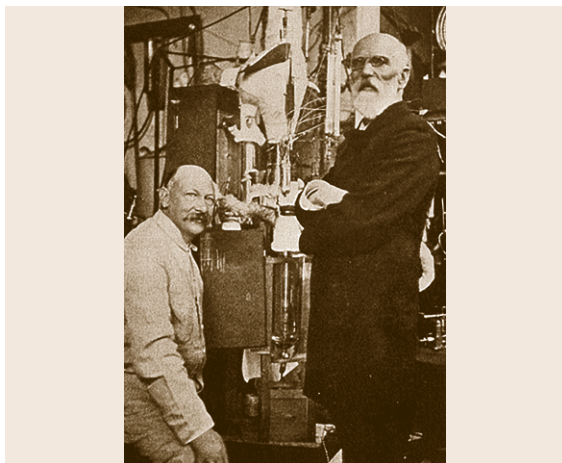
Lorentz uses the notion of similar systems to explore the constraints on theory, as opposed to using theory to state how one can construct a system  $S'$  to be similar to a certain system  $S$ , in order to make inferences about one of the systems based upon observations about the other. This seems a different use of the notion than Galileo or Newton made of it; it also allows the contemplation of unprecedented kinds of similarity. It may, Lorentz realizes, even give rise to systems of a different ontological status; he explains why that, too, may be useful [18.29, pp. 447–448]:

“It might be argued that two bodies existing in nature will hardly ever be similar in the sense we have given to the word, and that therefore, if  $S$  corresponds to a real system, this will not be the case with  $S'$ . But this seems to be no objection. Suppose, we have formed an image of a class of phenomena, with a view to certain laws that have been derived from observation or from general principles. If, then, we wish to know, which of the features of our picture are essential and which not, i. e., which of them are necessary for the agreement with the laws in question we have only to seek in how far these latter will still hold after different modifications of the image; it will not at all be necessary that every image which agrees in its essential characteristics with the one we have first formed corresponds to a natural object.”

Thus, Lorentz’s exploratory use of similar systems in fields beyond mechanics was motivated by the example of van der Waals’ and Onnes’ highly successful results using mechanical similarity to derive new theoretical results.

### Van der Waals and Onnes

In his 1881 *General Theory of Liquids*, Onnes argued that van der Waals’ “law of corresponding states”, which had just been published the previous year, could be derived from scaling arguments, in conjunction with assumptions about how molecules behaved. Van der Waals was impressed with the paper, and a long friendship between the two ensued (Fig. 18.4). Van der Waals was awarded the Nobel Prize in Physics in 1910 for “The equation of state for gases and liquids” [18.30], and Onnes was awarded it in 1913 [18.31], for “Investigations into the properties of substances at low temperatures, which have led, amongst other things, to the preparation of liquid helium.” In his lecture delivered for the occasion, Onnes highlighted the connection between



**Fig. 18.4** Heike Kamerlingh Onnes and Johannes Dederik van der Waals in the laboratory with the helium *liquefactor*. Onnes used the theory of corresponding states along with experimental data on one substance to predict the conditions at which helium would liquefy

his investigations into properties of substances at low temperatures and similarity principles [18.32, p. 306]:

“[...] From the very beginning [...] I allowed myself to be led by van der Waal’s theories, particularly by the law of corresponding states which at that time had just been deduced by van der Waals.

This law had a particular attraction for me because I thought to find the basis for it in the stationary mechanical similarity of substances and from this point of view the study of deviations in substances of simple chemical structure with low critical temperatures seemed particularly important.”

What is special about the low temperatures Onnes needed to achieve in order to liquefy helium is that, according to the kinetic theory of gases on which van der Waals’ equation of state was based, there would be much less molecular motion than in the usual kinds of cases considered. Onnes’s approach in looking for the foundation of the law of corresponding states has a slightly different emphasis than the kinetic theory of gases. Boyle’s law (often called the ideal gas law) and van der Waals’ equation were based on investigating the relationship between the microscale (the molecular level) and the macroscale (the properties of the substance, such as temperature and density). But Onnes was instead looking at the foundation for the similarity of states. Like van der Waals, he looked to mechanics and physics for governing principles, but Onnes pointed out that it was also useful to look at principles of sim-

ilarity. At low enough temperatures, where the motion of the molecules was not the predominant factor, the relevant principles of similarity would be principles of static mechanical similarity, as opposed to dynamical similarity.

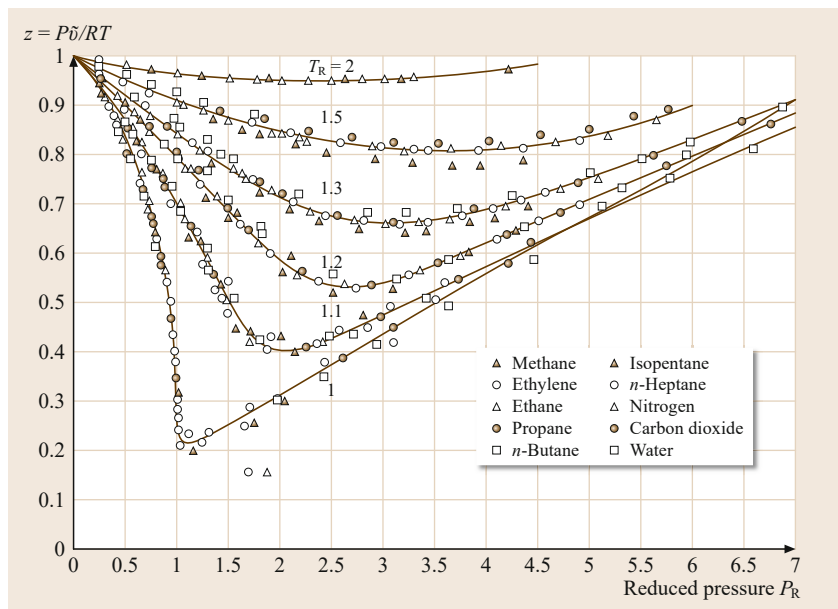
The criterion for similarity Onnes developed arose out of investigations into the transition from one regime to another. This had been the case in work in hydrodynamics, too. In Osborne Reynolds work, discussed below, it was the critical point at which fluid flow underwent a transition from laminar to turbulent flow (or, in his terminology, from “lamellar” to “eddying” flow [18.6, p. 200]) that led to the identification of the dimensionless parameter that later became known as the Reynolds number. The Reynolds number is in a way a criterion of similarity, in that fluid systems with the same Reynolds number will be in the same flow regime, regardless of the fluid. So it was with thermodynamics, Onnes showed: the critical point at which a substance undergoes a transition from the gaseous state to the liquid state led to the identification of a criterion of similarity of states that held for all substances.

Van der Waals was interested in the continuity of states and used the critical values of pressure, volume, and temperature in a brilliant way to normalize them. He defined “reduced pressure”, “reduced volume”, and “reduced temperature” to yield an equation of state in which none of the parameters that are characteristic of a particular substance appear. As *Levelt Sengers* notes [18.33, p. 25], “This is a truly remarkable result.” The equation of state is [18.33, p. 25]:

“universal; all characteristics of individual fluids have disappeared from it or, rather, have been hidden in the reduction factors. The reduced pressures of two fluids are the same if the fluids are in *corresponding states*, that is, at the same reduced pressure and volume.”

This is an important part of the history of similar systems in that the principle of corresponding states allowed the production of curves (Fig. 18.5) representative of all substances from experiments on a particular substance [18.33, p. 26]:

“The principle of corresponding states [...] frees the scientist from the particular constraints of the van der Waals equation. The properties of a fluid can now be predicted if only its critical parameters are known, simply from correspondence with the properties of a well characterized reference fluid. Alternatively, unknown critical properties of a fluid can be predicted if its properties are known in a region not necessarily close to criticality, based on the behavior of the reference fluid.”



**Fig. 18.5** This graph of empirical results illustrates the theory of corresponding states that van der Waals presented in 1880; fluids in *corresponding states* have the same *reduced pressure*: the data for different substances all *fall on top of each other*. Onnes later derived it using a law of mechanical similarity he ascribed to Newton

Onnes used this insight about corresponding states to set up an experimental apparatus to liquefy helium, which has an extremely low critical temperature. What is so exciting about his story is that he had to rely on the law of corresponding states to estimate the critical temperature so that he would know where to look – that is, so that he would know what conditions to create in order for helium to liquefy. What is especially relevant to the history of the notion of *physically similar systems* is that he did more than just use van der Waals’ law of corresponding states. He also gave a foundation for it that was independent of the exact form of van der Waals’ equation and did not depend on results in statistical mechanics. Instead, he used *mechanical similarity* [18.33, p. 30]:

“Kamerlingh Onnes’s (1881) purpose is to demonstrate that the principle of corresponding states can be derived on the basis of what he calls the principle of similarity of motion, which he ascribes to Newton. He assumes, with van der Waals, that the molecules are elastic bodies of constant size, which are subjected to attractive forces only when in the boundary layer near a wall, since the attractive forces in the interior of the volume are assumed to balance each other [...] He realizes this can be valid only if there is a large number of molecules within the range of attraction [...] Onnes] considered a state in which  $N$  molecules occupy a volume  $v$ , and all have the same speed  $u$  (no Maxwellian distribution!). The problem is to express the external pressure  $p$ , required to keep the system of moving

particles in balance, as a function of the five parameters. He solves this problem by deriving a set of scaling relations for  $M$ ,  $A$ ,  $v$ ,  $u$ , and  $p$ , which pertain if the units of length, mass, and time are changed.”

Onnes provides a criterion for corresponding states based on these scaling relations, along with assumptions about what the molecular-sized objects are like. *Levelt Sengers* remarks [18.33, p. 30]:

“Two fluids are in corresponding states if, by proper scaling of length, time and mass for each fluid, they can be brought into the same “state of motion.” It is not clearly stated what he means by this, but he must have had in mind an exact mapping of the molecular motion in one system onto that of another system if the systems are in corresponding states.”

*Levelt Sengers* illustrates what being in the same “state of motion” means “in modern terms” [18.33, p. 30]:

“[...] suppose a movie is made of the molecular motions in one fluid. Then, after setting the initial positions and speed of the molecules, choosing the temperature and volume of a second fluid appropriately, and adjusting the film speed, a movie of the molecular motion in a second fluid can be made to be an exact replica of that in the first fluid.”

Appeal to such imagined visual images is very much in keeping with nineteenth century science, and one can see here an attempt to generalize Newton’s use

of similar systems in the *Principia* to thermodynamics. Onnes used the principle of corresponding states for more than visualizing, though, and, even, for more than theorizing; he used it to show how one could make a prediction about one fluid from knowledge about another. *Wisniak* explains [18.34, p. 569]

“Kamerlingh Onnes proposed to use the law of corresponding states to examine the possibility of cooling hydrogen further by its own expansion. He then used this law to predict from the known experience with oxygen what was to be expected from the apparatus for the cooling of hydrogen: [quoting Onnes:] But let us return to the thermodynamically corresponding substances. If two such substances are brought in corresponding engines and if these engines are set in motion with corresponding velocities, then they will run correspondingly as long as there is given off a corresponding quantity of heat in the corresponding times by the walls of the machine.”

Thus, Onnes has introduced not just corresponding motions and times, as in mechanical similarity, but also corresponding quantities of heat. *Wisniak* continues [18.34, p. 569]

“He [Onnes] then introduced the notion of thermodynamically corresponding operations to argue that if then in a model, working with oxygen, after a given time a given volume of liquid oxygen is found, there will be obtained in the corresponding hydrogen apparatus after the corresponding time a corresponding volume of liquid hydrogen.”

By *model* here, Onnes clearly means physical model, and the model includes the contained gases such as oxygen and hydrogen. The model is an actual physical model: a physical setup, an actual, physical machine. By the end of the nineteenth century, the physics of machines included the thermodynamics of machines. And, as in Newton and Galileo’s day, one could talk both about imagined similar systems, and about actual similar machines.

### Maxwell and Boltzmann

As several scholars have noted, Ludwig Boltzmann (1844–1906) mentioned *similar systems* in his investigations into the theory of gases, too. It’s been noted that, in his 1884 and 1887 papers, Boltzmann [18.35, pp. 56–57]:

“tried to deepen the foundation of the new theory [that was to become known as statistical mechanics] by introducing the concept of *Ergoden* – meaning a collection (ensemble) of similar systems (of gas

molecules) having the same energy but different initial conditions.”

*Stephen G. Brush*, also citing Boltzmann’s 1884 and 1887 papers, remarks that [18.36, pp. 75–76]:

“There has been considerable confusion about what Maxwell and Boltzmann really meant by ergodic systems. It appears that they did not have in mind completely deterministic mechanical systems following a single trajectory unaffected by external conditions; [ . . . ]

In fact, when Boltzmann first introduced the words *Ergoden* and *ergodische*, he used them not for single systems but for the collections of similar systems with the same energy but different conditions. In the papers published in 1884 and 1887, Boltzmann was continuing his earlier analysis of mechanical analogies for the Second Law of Thermodynamics, and also developing what is now (following J. Willard Gibbs) known as *ensemble theory*. Here again, Boltzmann was following a trail blazed by Maxwell, who had introduced the ensemble concept in his 1879 paper. But while Maxwell never got past the restriction that all systems in the ensemble must have the same energy, Boltzmann suggested more general possibilities, and Gibbs ultimately showed that it is most useful to consider ensembles in which not only the energy but also the number of particles can have any value, with a specified probability.”

What these commentators on Boltzmann are referring to in mentioning the influence of *Maxwell* are Maxwell’s remarks in his *On Boltzmann’s Theorem on the average distribution of energy in a system of material points* [18.37]. There, *Maxwell* wrote, speaking of the case “in which the system is supposed to be contained within a fixed vessel” [18.37, pp. 715ff]:

“I have found it convenient, instead of considering one system of material particles, to consider a large number of systems similar to each other in all respects except in the initial circumstances of the motion, which are supposed to vary from system to system, the total energy being the same in all. In the statistical investigation of the motion, we confine our attention to the number of these systems which at a given time are in a phase such that the variables which define it lie within given limits. [Emphasis in italic added.]

If the number of systems which are in a given phase (defined with respect to configuration and velocity) does not vary with the time, the distribution of the systems is said to be *steady*.”

It is not clear how the use of the notion of similar systems here, i.e., in forming ensembles in thermodynamics in order to study their behavior statistically, might be related to either Newton's notion of similar systems or the notion involved in the principle of corresponding states. It is certainly a use of similar systems that is very different from using one system experimentally to infer the values of quantities in another. So, if, as Brush's comment implies, Boltzmann was thinking of more general kinds of similar systems, it seems he was no longer restricting the notion of similar systems to systems that are behaviorally similar to each other with respect to motions, and he was not restricting its use to the use of one system or machine to infer the behavior of another.

Yet Boltzmann's departure from Newton's use of the term similar systems was almost certainly not a matter of confusion on Boltzmann's part about the notion in the sense Newton had used it, for *Boltzmann's* encyclopedia entry on models [18.38] shows that he was well aware of, and respected the distinctive nature of, the use of experimental models of machines, in which one machine is specially constructed in order to infer the behavior of another. Boltzmann, in fact, associates the latter kind of model with Newton's insights.

On the approach in which physical models constructed with our own hands are actually a continuation and integration of our process of thought, *Boltzmann* says in that encyclopedia article (*Model*) [18.38]:

“physical theory is merely a mental construction of mechanical models, the working of which we make plain to ourselves by the analogy of mechanisms we hold in our hands.”

In contrast, *Boltzmann* explicitly described experimental models as of a different sort than the kind with which he was comparing mental models, and explained why they must be distinguished [18.38]:

“A distinction must be observed between the models which have been described and those experimental models which present on a small scale a machine that is subsequently to be completed on a larger, so as to afford a trial of its capabilities. Here it must be noted that a mere alteration in dimensions is often sufficient to cause a material alteration in the action, since the various capabilities depend in various ways on the linear dimensions. Thus the weight varies as the cube of the linear dimensions, the surface of any single part and the phenomena that depend on such surfaces are proportionate to the square, while other effects – such as friction, expansion and condition of heat, etc., vary according

to other laws. Hence a flying-machine, which when made on a small scale is able to support its own weight, loses its power when its dimensions are increased. The theory, initiated by Sir Isaac Newton, of the dependence of various effects on the linear dimensions, is treated in the article Units, Dimensions Of.”

The use of a flying machine to illustrate the point was not incidental; in his *On Aeronautics*, Boltzmann urged research into solving the problem of flight, and expressed his opinion that experimentation with kites was the appropriate approach. The complexities of air-flow over an airplane wing, he said, were too difficult to study using hydrodynamics [18.39, p. 256]. Yet, the basis for extrapolating from experiments on a kite or flying machine from one observed situation to another, unobserved, situation (even with a machine of the same size) owes something to hydrodynamics. The dimensionless parameters yielding the appropriate correspondences between homologous quantities for kites and flying machines were provided by Helmholtz's innovative use of the equations of hydrodynamics.

### 18.3.3 Similar Systems in Theoretical Physics

#### Stokes and Helmholtz

Hermann von Helmholtz (1821–1894), like Ludwig Boltzmann and so many other physicists of the nineteenth century, contributed to the scientific literature on research into flight. Some of these contributions took the form of investigations concerning the earth's atmosphere. Six of the 20 papers in the important and selective 1891 anthology *The Mechanics of the Earth's Atmosphere: A Collection of Translations* by Cleveland Abbe [18.40] were by Helmholtz; one of these was his 1873 *On a Theorem Relative to Movements That Are Geometrically Similar in Fluid Bodies, Together with an Application to the Problem of Steering Balloons* [18.41, 42]. It is the only one of Helmholtz's papers in that volume that explicitly addresses an application to the problem of flight. What is relevant to the history of the concept of similar systems is the kind of reasoning he uses in the paper.

*Helmholtz's* starting point is “the hydrodynamic equations” which, he argues, can be considered “the exact expression of the laws controlling the motions of fluids” ([18.41, p. 67], [18.42]). What about the well-known contradictions between observations and the consequences of the equations? Those, he argues, are only apparent contradictions, which disappear once the phenomenon of “surfaces of separation” is no longer neglected [18.41, p. 67]; his *On Discontinuous Motions*

in *Liquids* [18.43, 44], also included in the same collection of translations, aims to establish their existence.

The *Discontinuous Motions* paper [18.43] is an extraordinarily interesting contribution to the methods of reasoning by analogy between fluid currents, electrical currents, and heat currents. For, the paper begins by pointing out that “the partial differential equations for the interior of an incompressible fluid that is not subject to friction and whose particles have no motion of rotation” are precisely the same as the partial differential equations for “stationary currents of electricity or heat in conductors of uniform conductivity” [18.43, p. 58]. Yet, he notes, even for the same configurations and boundary conditions, the behavior of these different kinds of currents can differ. How can this be? It would be easy to assume that the difference is a matter of the equations being, in the case of hydrodynamics, an “imperfect approximation to reality,” possibly due to friction or viscosity. Yet, Helmholtz argues, various observations indicate that this is not plausible. Instead, he proposes, the difference in behavior between fluid currents on the one hand and electrical and heat currents on the other is due to “a surface of separation” that exists or arises in the case of the fluid. In some situations, “the liquid is torn asunder,” whereas electricity and heat flows are not. Though the main point of the paper is to propose his detailed account of what happens in the liquid to cause this difference (the pressure becomes negative), it is interesting, especially in the context of nineteenth century, that Helmholtz discusses a case in which physical entities described by the same partial differential equations do *not* behave in the same way. Yet, once the existence of discontinuous motions in fluids is recognized, *Helmholtz* says, the contradictions that “have been made to appear to exist between many apparent consequences of the hydro-dynamic equations on one hand and the observed reality on the other” will then “disappear” [18.41, p. 67].

The problem with the hydrodynamic equations is not that they are wrong, for they are not; they are “the exact expressions of the laws controlling the motions of fluids”. The problem is that [18.41, p. 67]

“it is only for a relatively few and specially simple experimental cases that we are able to deduce from these differential equations the corresponding integrals appropriate to the conditions of the given special cases.”

So, the hydrodynamic equations are impeccable; it’s their solution that is the problem. Simplifying is not going to work, either, since in some cases “the nature of the problem is such that the internal friction [viscosity] and the formation of surfaces of discontinuity can not be neglected.” These surfaces of discontinuity present

a very fundamental problem to finding a neat solution, too, for [18.41, p. 67]

“The discontinuous surfaces are extremely variable, since they possess a sort of unstable equilibrium, and with every disturbance in the whirl they strive to unroll themselves; this circumstance makes their theoretical treatment very difficult.”

Theory being of very little use in prediction here, [18.41, p. 68]

“we are thrown almost entirely back upon experimental trials, [...] as to the result of new modifications of our hydraulic machines, aqueducts, or propelling apparatus.”

That was how things stood but, *Helmholtz* says, there is another method, one that is neither a matter of prediction from theory nor an experimental trial of the machine whose behavior one wishes to predict. His description deserves to be read closely [18.41, p. 68]:

“In this state of affairs [the insolubility of the hydrodynamic equations for many cases of interest] I desire to call attention to an application of the hydro-dynamic equations that allows one to transfer the results of observations made upon any fluid and with an apparatus of given dimensions and velocity over to a geometrically similar mass of another fluid and to apparatus of other magnitudes and to other velocities of motion.”

The method Helmholtz is referring to, which he presented in this now-classic paper in 1873, thus differs from deducing predictions from theory in the same way that Newton’s notion of similar systems and Galileo’s use of one pendulum to inform him about another differ from deducing predictions from theory: theory is involved in the inference, but the *way* that theory is involved is to allow someone to “transfer the results of observations” made on one thing (system, machine, mass of fluid, apparatus) over to another thing (system, machine, mass of fluid, apparatus).

The way Helmholtz proceeds to establish this different “application of the hydro-dynamic equations” appeals to a formalism not available to either Galileo or Newton, though: “[t]he equations of motion in the Eulerian form introducing the frictional forces, as is done by Stokes.” Although Helmholtz does not use the term “similar system” here, *Stokes* did use it, in his *On the Effect of the Internal Friction of Fluids on the Motion of Pendulums*, presented in 1850 [18.45, p. 19]. In that paper, before attempting a solution of some flow equations, *Stokes* first examined “the general laws which follow merely from the dimensions of the several terms which appear in the equations.” To do this,



Stokes had employed “similar systems” [18.45, pp. 16–17]:

“Consider any number of similar systems, composed of similar solids, oscillating in a similar manner in different fluids or in the same fluid. Let  $a, a', a'' \dots$  be homologous lines in the different systems;  $T, T', T'' \dots$  corresponding times, such for example as the times of oscillation from rest to rest. Let  $x, y, z$  be measured from similarly situated origins, and in corresponding directions, and  $t$  from corresponding epochs, such for example as the commencements of oscillations when the systems are beginning to move from a given side of the mean position.”

Then, Stokes says that the form of the equations shows that the equations being satisfied for one system will be satisfied for all the systems, if certain relations between the quantities in those equations are met, which he lays out. He adds the condition needed in order for the systems to be dynamically similar; then, if we “compare similarly situated points,” the motions in the systems will also be similar, and the “resultants [of pressure of the fluids on the solids] in two similar systems are to one another” in a certain ratio that he shows how to obtain. Stokes does not end there; the paper contains further discussion about establishing similarity between the two systems, having to do with how the fluids are confined. This much about *Stokes* should give a general idea of how he conceived of and used the notion of “similar systems” [18.45, p. 19].

Helmholtz’ approach probably owes much to Stokes; *David Cahan’s* study *Helmholtz and the British Scientific Elite: From Force Conservation to Energy Conservation* identifies Stokes as one of the British elites with whom Helmholtz built a relationship during the 1850s and 1860s [18.46]. Helmholtz does refer to Stokes, to be sure, but there is also something creative in what he does in his own paper. Helmholtz turns the idea of how the Eulerian equations for flow are related to similar systems around, so that he sees how one might, in principle at least, use the equations in conjunction with model experiments on ships to inform us about how to predict and direct the motions of balloons (dirigibles).

The discussion and derivation of the conclusions *Helmholtz* reached for all the cases he considered in his 1873 paper [18.41] are too long to summarize here, but a few points can be mentioned:

1. Helmholtz’s strategy is to consider two given fluids and use the hydrodynamic equations to infer the way or ways in which their quantities must be related. For the first fluid, the directions of its

coordinate axes are designated  $x, y,$  and  $z$ ; the components of velocity associated with them are designated  $u, v,$  and  $w$ . The time  $t$ , fluid density  $\varepsilon$ , pressure  $p$ , and coefficient of friction  $k$  (viscosity) are also named, which allows him to construct the equations of motion of the first fluid in the Eulerian form. The second fluid is then given designations of  $U, V, W$  for the components of velocity (in coordinate axes  $X, Y, Z$ ), the pressure  $P$ , the fluid density  $E$ , and the viscosity constant by  $K$ . Three additional constants  $q, r,$  and  $n$  are named, so that the quantities in the second fluid can then be related to the designated quantities in the first fluid such that the quantities in the second fluid will also satisfy the equations of motion that were constructed for the first fluid. For example, the densities of the two fluids are related by  $E = r\varepsilon$ ; their coefficients of friction are related by  $K = qk$ ; and the velocity components, by  $U = nu, V = nv,$  and  $W = nw$ . Then, the pressures must be related by  $P = n^2rp + \text{constant}$ , and the times in the two fluids must be related by  $T = qt/n^2$ . Putting the terms for the quantities of the second fluid expressed in terms of the quantities of the first fluid into the equations of motion for the first fluid shows that they satisfy those equations.

2. The nature of the two fluids determines how their densities and coefficients of friction are related to each other, so two of the three constants,  $q$  and  $r$ , are determined. Helmholtz then considers various kinds of cases (e.g., compressible versus incompressible, cohesive versus noncohesive (liquid vs gaseous fluids), certain boundary conditions, whether friction can be neglected), and what they permit to be inferred about the third undetermined constant  $n$ . The paper contains a variety of interesting remarks, some of great practical significance, about how other quantities of the two fluids (e.g., velocity of sound) must be related to each other.
3. When Helmholtz comes to addressing the practical problem mentioned in the title: “driving balloons forward relative to the surrounding air,” he uses not two masses of air in which two different air balloons are situated, but rather: for the second fluid, *a mass of air in which an air balloon is situated*, and, for the first fluid, *a mass of water in which a ship is situated*. He writes: “our propositions allow us to compare this problem [driving balloons forward relative to the surrounding air] with the other one that is practically executed in many forms, namely, to drive a ship forward in water by means of oar-like or screw-like means of motion [...] we must [...] imagine to ourselves a ship driven along under the surface. Such a balloon which presents

a surface above and below that is congruent with the submerged surface of an ordinary ship scarcely differs in its powers of motion from an ordinary ship” [18.41, p. 73]. Then, letting “the small letters of the two above given systems of hydrodynamic equations refer to water and the large letters to the air,” he examines the practical conditions under which he can “apply the transference from ship to balloon with complete consideration of the peculiarities of air and water.”

Helmholtz’s discussion contains many subtle points concerning what would need to be considered if actually building the kind of ship needed to model an air balloon. As he indicates, the practical considerations involved in applying the method are not trivial and can sometimes even be prohibitive; nevertheless, the point is that the approach he outlines permits one to make a proper analysis of any such comparison, or “transference” using the hydrodynamic equations, and can sometimes yield a solution when the hydrodynamic equations are insoluble [18.41, p. 74]. Evidence of the influence and significance of this particular paper of Helmholtz’s into the twentieth century appears in *Zahm’s Theories of Flow Similitude* [18.7]. Zahm identifies three methods, one with Isaac Newton, one with Stokes and Helmholtz, and one with Rayleigh. The sole paper by Helmholtz cited there is this paper of 1873 [18.41].

The significance to the history of physically similar systems is that Helmholtz’s account of his method involves a differential equation, that the equation is so central to the account, and that how it is involved is stated so clearly. What is not stated very clearly is whatever it is that plays the role of system; sometimes Helmholtz seems to be saying that the transference is from one mass of fluid to another; other times, that it is between the objects situated within the fluid. If we denote whatever ought to play that role by the term system, though, we would say that, in Helmholtz’s analysis, the hydrodynamic equations are not only the core of the criterion for allowing the “transference” of results [18.41, p. 74] observed in one situation to another, but they indirectly give a criterion for, and thus specify, what a system is, that is, what the similarity in *similar systems* is between. If we use the term system this way, then it is implicit in Helmholtz’s account that a system is the mass and its configuration (including anything situated within the mass), with boundary conditions, to which the partial differential equation applies. We might also take note of the fact that what the equation applies to is in equilibrium (though not necessarily static equilibrium). The governing differential equations are important, too, in the

specification of what quantities need to be considered in the analysis.

Yet, Helmholtz is careful not to overreach concerning what can be deduced from the *form* of an equation; as he points out in his *Discontinuous Motions* paper [18.43] when investigating the example of fluid being “torn asunder”: *just because a certain situation is governed by an equation of the same form as another equation governing a different situation, does not in itself guarantee that the two situations will exhibit analogous behavior* – even when the configuration and boundary conditions are also analogous. It is for the confluence of all these points that I consider Helmholtz’ 1873 paper [18.41] such a major contribution to the history of the concept of similar systems.

### Reynolds

Osborne Reynolds’ (1842–1912) work and influence on similarity was immense, but it was by no means his only major achievement [18.47]. Unless one has invested the time required to read a significant part of his work, any evaluation of his achievements and influence will sound like hyperbole. I mention here only his most significant contribution relevant to the history of the concept of similar systems.

The decisive difference Reynolds made in the notion of similar systems was to show that it applied beyond well-behaved regimes. In fact, he showed, it applied during the transition between well-behaved regimes and chaotic ones. And, not only that, but that the critical point of transition between well-behaved (laminar flow) and chaotic (turbulent flow) regimes could be characterized, and characterized by a parameter that was independent of the fluid. *Stokes* put it well in the statement he made in his role as President of the Royal Society on the occasion of presenting a Royal Medal to Reynolds on November 30, 1888 [18.45, p. 234]:

“In an important paper published in the *Philosophical Transactions* for 1883, [Osborne Reynolds] has given an account of an investigation, both theoretical and experimental, of the circumstances which determine whether the motion of water shall be direct or sinuous, or, in other words, regular and stable, or else eddying and unstable. The dimensions of the terms in the equations of motion of a fluid when viscosity is taken into account involve, as had been pointed out, the conditions of dynamical similarity in geometrically similar systems in which the motion is regular; but when the motion becomes eddying it seemed no longer to be amenable to mathematical treatment. But Professor Reynolds has shown that the same conditions of similarity

hold good, as to the average effect, even when the motion is of the eddying kind; and moreover that if in one system the motion is on the border between steady and eddying, in another system it will also be on the border, provided the system satisfies the above conditions of dynamical as well as geometrical similarity.”

Stokes does not here use the term *similar systems*, but that is what he means in using the grammatical construction: “if in one system [...], in another system it will also [...], provided the system satisfies the above conditions of dynamical as well as geometrical similarity.” What this means is that there are some (experimentally determined) functions of a certain (dimensionless) parameter that describe the behavior of fluids, whatever the fluid. The parameter is not a single measured quantity such as distance, velocity, or viscosity; rather, it is a ratio involving a number of quantities (e.g., density, velocity, characteristic length, and viscosity). The ratio is without units, as it is dimensionless. Reynolds is often cited for coming up with the criterion of dynamical similarity, but obviously, the idea predated his work, as Stokes’ statement recognizes. Rather, what *Reynolds* did that was so decisive for the future of hydrodynamics (and aerodynamics) was, as he explained in a letter to Stokes, that there was a *critical value (or values)* for “what may be called the parameter of dynamical similarity [the dimensionless parameter mentioned earlier, which is now known as Reynolds number]” [18.48, p. 233].

In the excerpt from his statement quoted above, Stokes puts his finger on why what *Reynolds* did was so significant in terms of a fundamental understanding of fluid behavior, but *Reynolds*’ 1883 paper also had practical significance for research in the field as well. *Stokes* continued [18.49, p. 234]:

“This is a matter of great practical importance, because the resistance to the flow of water in channels and conduits usually depends mainly on the formation of eddies; and though we cannot determine mathematically the actual resistance, yet the application of the above proposition leads to a formula for the flow, in which there is a most material reduction in the number of constants for the determination of which we are obliged to have recourse to experiment.”

It is not surprising that interest in applying the methods of similar systems grew in the subsequent years.

### Prandtl

*Prandtl*’s work in experimental hydrodynamics and aerodynamics is singularly prominent in work done in

the field in Germany in the twentieth century. *Ludwig Prandtl* (1873–1953) was an ex-engineer-turned-professor in the Polytechnic at Hanover conducting research on air flow when he presented a paper at the Third International Congress of Mathematicians in 1904: *Motion of Fluids with Very Little Viscosity* [18.50]. It did not make much of a splash – except with Felix Klein, then a prominent mathematician at the University of Göttingen. In his paper, *Prandtl* laid out a plan to treat flow around bodies. What he proposed was that the problem be analyzed into several distinct questions [18.50]:

1. What happened at the boundary of the *skin* that formed against the body, and what happened on each side of it, that is
2. What happened in the fluid on the side of the boundary that was within the *skin*, and
3. What happened in the fluid on the other side of the boundary, within the main fluid stream.

*Prandtl* showed that, in the mainstream, the mathematical solutions that were obtained by neglecting viscosity could be applied to even these real fluids. In the part of the flow under the *skin* formed around the body, however, viscosity did have to be taken into account. And, crucially, what happened in the mainstream – the formation of vortices – set conditions for what happened on the other side of the boundary, via setting boundary conditions at the interface between the two layers. Klein saw the potential of *Prandtl*’s approach and brought him to a post in Göttingen right away [18.11].

In Göttingen, *Prandtl* then made use of the knowledge that had been developed about hydrodynamical similarity, using a water tank for some of his most famous experiments. Rather than towing an object in the water, though, *Prandtl* used a water wheel to move the fluid in the water tank, much like fans were being used to push air through wind tunnels which by then were replacing the whirling arm or moving rail-car apparatuses used earlier in aerodynamical research. *Prandtl*’s results for airfoils were based on hydrodynamical similarity and, hence, on the concept of dynamically similar systems. His approach went beyond that, too, including fundamental questions he addressed by combining mathematical solutions and experimental results in an uncommon kind of synthesis. *William Lanchester* in England also employed dynamic similarity and authored significant works about his theoretical and experimental research in aerodynamics; his visit to *Prandtl* in 1908 may have contributed somewhat to *Prandtl* developing these ideas, since *Prandtl* was in a position to understand *Lanchester*’s work, and appreciate its significance [18.11].

### Rayleigh

Lord Rayleigh (John William Strutt) (1842–1919) became a proponent of dynamic similarity in Great Britain. The context of his advocacy of the method was part scientific, part political. The scientific part was an appreciation for the significance of dynamical similarity in effective research; the political part was a feeling that Britain ought not be left behind in aeronautical research. His political, social, and professional prominence put him in a position to be an effective advocate. He was the first president of the British Advisory Committee on Aeronautics, founded in 1909. Its first report includes his *Note as to the Application of the Principle of Dynamical Similarity* [18.51]; he introduces the topic by first citing Lanchester for one application of the principle of dynamical similarity, then noting his own communications of “a somewhat more general statement which may be found to possess advantages.” The next year, 1910–1911, the committee’s annual report included two papers on dynamical similarity, one of them by Rayleigh, under the *General Questions in Aerodynamics* section of the report [18.52]. In 1911–1912, the annual report mentions plans for experiments on an airship to determine its resistance “by towing tests in the William Froude National Tank” [18.53]. Under a section, *The Law of Dynamical Similarity and the Use of Models in Aeronautics*, the report notes its significance to all their research [18.52]:

“The theory relating to dynamical similarity explained by Lord Rayleigh and Mr. Lanchester in the first of the Annual Reports of the Committee is of fundamental importance in all applications of the method of models to the determination of the forces acting on bodies moving in air or in water.”

The next year, the annual report noted that [18.54]:

“Much evidence has now been accumulated in favour of the truth of the law of dynamical similarity to which attention was drawn by Lord Rayleigh and Mr. Lanchester in the first Report of this Committee.”

In June of 1914, the journal *Nature* featured a kind of survey paper, *Fluid Motions*, based on “a discourse delivered at the Royal Institution on March 20” by Rayleigh [18.55]. Here, we see Rayleigh actively campaigning for wider appreciation and use of the principle, which he credits Stokes with having “laid down in all its completeness.” We know that Stokes explicitly used the notion of similar systems in developing and explaining the use of the principle, so it is fair to say that Rayleigh means his discussion and use of it to be consistent with Stokes’ notion of similar systems.

In this paper, Rayleigh pointed out that it appeared that viscosity was important in many cases where it was so small that it seemed improbable that it should matter. When viscosities were low, as in water, one would not expect that the actual value of viscosity would be a significant factor in water’s qualitative behavior. As explained above, Osborne Reynolds’ results on fluid flow in pipes had shown that it is; Reynolds began to suspect that viscosity was important even in water when he observed unexpected changes in fluid flow as the temperature was varied. Since viscosity varies with temperature, he investigated the effect of viscosity and found that it was indeed important for fluid flow through pipes, even for nonviscous fluids such as water. Rayleigh added that Reynolds also investigated cases where viscosity was the “leading consideration,” as Rayleigh put it, in remarking that “It appears that in the extreme cases, when viscosity can be neglected and again when it is paramount, we are able to give a pretty good account of what passes, it is in the intermediate region, where both inertia and viscosity are of influence, that the difficulty is the greatest” [18.55]. This is the lead-in to his advocacy for the law of dynamic similarity: “But even here we are not wholly without guidance.” What is this guidance? He continues [18.55, p. 364]:

“There is a general law, called the law of dynamical similarity, which is often of great service. In the past this law has been unaccountably neglected, and not only in the present field. It allows us to infer what will happen upon one scale of operations from what has been observed at another.”

Rayleigh also notes: “But the principle is at least equally important in effecting a comparison between different fluids. If we know what happens on a certain scale and at a certain velocity in *water*, [emphasis in the original] we can infer what will happen in *air* on any other scale, provided the velocity is chosen suitably.” This is, of course, the point Helmholtz had made in 1873. Rayleigh notes that the point applies only in the range where the velocities are small in comparison to the velocity of sound [18.55].

Rayleigh gives an example of a use of the principle which permits one observation or experiment to be regarded as representative of a whole class of actual cases: that is, the class of all the other cases to which it is similar, even though the cases may have very different values of measurable quantities such as velocity. The important fact about the situation is expressed by the formula for the dimensionless parameter, which picks out the cases to which it is similar [18.55, p. 364]:

“It appears that similar motions may take place provided a certain condition be satisfied, viz. that the

product of the linear dimension and the velocity, divided by the kinematic viscosity of the fluid, remain unchanged.”

Put more specifically, the important feature of a particular situation is the value of this dimensionless parameter; what Rayleigh is saying is that, even in cases of a different fluid, so long as this dimensionless product is the same (and, of course, that one is in the applicable velocity range for which it was derived), the motions will be similar.

One might think that, by 1914, when the use of wind tunnels had become recognized as essential to practical aeronautical research, this principle would have become accepted and would no longer be in question, at least among aeronautical researchers. But if Rayleigh’s estimation of the state of the profession is correct, apart from Lanchester’s work, this was not so, even as late as March of 1914; he says that:

“although the principle of similarity is well established on the theoretical side and has met with some confirmation in experiment, there has been much hesitation in applying it, [. . .]”

He especially mentions problems in its acceptance in aeronautics due to skepticism that viscosity, which is extremely small in air, should be considered an important parameter:

“In order to remove these doubts it is very desirable to experiment with different viscosities, but this is not easy to do on a moderately large scale, as in the wind channels used for aeronautical purposes.”

Rayleigh tries to persuade the reader of the significance of the effects of viscosity on the velocity of fluid flow by relating some experiments he performed with a cleverly designed apparatus in his laboratory. The apparatus consisted of two bottles containing fluid at different heights, connected by a tube with a constriction, through which fluid flowed due to the difference in “head”, or height of fluid, in the two bottles [18.55, p. 364]. The tube with the constriction contained fittings that allow the measurement of pressure head at the constriction, and on either side of it. To investigate the effects of viscosity, Rayleigh varied the temperature of the fluid, which changes the fluid viscosity, and he observed how the velocity of the fluid flowing between the two bottles was affected. The kind of relationship he establishes and uses is of the form Galileo employed in reasoning from one pendulum to another. In other words, he worked in terms of ratios (ratios of velocities, ratios of viscosities, ratios of heads), and he employed the fact that some ratios are the square root of others [18.55]. He took the experimental results he reported in this 1914 paper to conclusively settle the question of the relevance of viscosity to fluid motions. This is an example of the kind of exploratory work that can be involved in order to answer one of the questions needed in order to use the principle of similarity properly: What quantities are relevant to the behavior of interest (in the range of interest)? Although the researcher’s experience and judgment are involved, sometimes new experiments should be, and are, conceived and carried out to help determine the question.

Rayleigh delivered this “Discourse” in early 1914 [18.55]. 1914 was a very special year for the concept of similar systems, and deserves a section all its own.

## 18.4 1914: The Year of *Physically Similar Systems*

In terms of an advance in the understanding and formalization of physically similar systems, 1914 was a landmark year, just as 1850 (*Stokes’* paper [18.45]), 1873 (*Helmholtz’s* paper [18.41]), and 1883 (*Reynolds’* paper [18.56]) would still be nineteenth century landmarks in any history of the concept of dynamical similarity. Going back to earlier eras, many would also consider the dates 1638 (*Galileo’s Two New Sciences* [18.10]) and 1673 (*Newton’s Principia* [18.3]) significant to the concept of similar systems. My review above suggests additions to the above list of dates in the nineteenth century that should be recognized as important in the history of the concept of similar systems: the years around 1880 (*van der Waals* paper [18.57]) and 1881 (*Onnes’* paper [18.32]). The role of the no-

tion of similar systems in both the development and the understanding of the principle of corresponding states in physical chemistry should enjoy far more recognition among philosophers of science than it has to date, and perhaps Lorentz ought to be included, too, for his recognition of the importance of the method of similar systems. A strong argument could also be made for including a date commemorating one of Froude’s influential achievements in the nineteenth century list.

In contrast, however, dates for the papers by Maxwell and Boltzmann using the term *similar systems* should not be included on this list, in my view. This exclusion is not a lack of generosity, but an effort at clarification. Their use of the term *similar system* in statistical mechanics, a term that already had a fairly

well-defined meaning in the theories of mechanical similarity and dynamical similarity, may have caused, or at least contributed to, confusion about the concepts of *similar system* and similarity as they are used in connection with mechanical and dynamical similarity. As we shall see, confusions about these concepts came to a head in 1914; perhaps it is no coincidence that at least one source of the confusion was a proposal by someone known for his work in statistical thermodynamics.

#### 18.4.1 Overview of Relevant Events of the Year 1914

In the part of 1914 leading up to *Buckingham's* landmark paper in October 1914 [18.2] that developed the notion of physically similar systems, hardly a month went by without some major work concerning similarity and similar systems appearing (Fig. 18.3):

- In January 1914, *Stanton* and *Pannell* publish a major compendium of work [18.6] done at Britain's National Physical Laboratory over the previous four years, *Investigation into Similarity of Motions*
- In February 1914, a much anticipated English translation of *Galileo's Two New Sciences* [18.10] was published.
- In March 1914, *Rayleigh* delivered his lecture *Fluid Motions* [18.55] at the Royal Institute (March 20, 1914).
- In April 1914, *Richard Chace Tolman's* *The Principle of Similitude* appears in *Physical Review* [18.58], and *Rayleigh's* *Fluid Motions* [18.55] is published in the periodical *Engineering*, 97 (April 8, 1914).
- In May 1914, *Buckingham* gives a paper on *The Interpretation of Model Experiments* to the Washington Academy of Sciences [18.59].
- In June 1914, *Rayleigh's* review article *Fluid Motions* was published in *Nature* [18.55].
- In July 1914, *Buckingham's* *Physically Similar Systems* was published in *Journal of the Washington Academy of Science* [18.1].
- In October 1914, *Buckingham's* *Physically Similar Systems: Illustrations of the Use of Dimensional Equations* [18.2].

And sometime during 1914, *Philipp Forchheimer's* *Hydraulik* [18.60] was published, which contains a section on *The Law of Similarity* (Das Ähnlichkeitgesetz). *Hydraulik* becomes a highly regarded compendium and reference work on Hydraulics for many decades afterward. In the concluding paragraph of the section on the law of similarity, *Forchheimer* writes that every hy-

draulic equation that fulfills the law of similarity can be expressed in the form of an equation consisting of an unidentified function  $F$  of three dimensionless ratios set equal to an unidentified constant. He indicates that the law of similarity is shown to be merely a special case of the general law according to which all the terms of any of the equations of importance in mechanics, need to be of equal dimension, inasmuch as the law of similarity treats one body as a prototype, and the others as copies of it.

#### 18.4.2 Stanton and Pannell

In January of 1914, *Stanton* and *Pannell* read their paper *Similarity of Motion in Relation to the Surface Friction of Fluids* [18.6] to the Royal Society of London. *Stanton* was superintendent of Britain's National Physical Laboratory (NPL) Engineering Department. The paper was a compendium of the work done there on similarity, and had been submitted to the Society in December 1913. It begins with references to *Helmholtz's* and *Stokes's* work using equations for non-ideal fluid flow, refers to *Newton's Principia* on similar motions, and uses *Rayleigh's* equation for fluid resistance. It explains that *Stanton* and *Pannell's* work involves investigating “the conditions under which similar motions can be produced under practical conditions.” The work had been carried out due in part to interest in the possibilities of using small-scale models in wind tunnels for engineering research. With one exception, they began, the experimental study of similar motions of fluids was very recent [18.6, p. 200]:

“Apart from the researches on similarity of motion of fluids, which have been in progress in the Aeronautical Department of the National Physical Laboratory during the last four years, the only previous experimental investigation on the subject, as far as the authors are aware, has been that of *Osborne Reynolds* [...]”

*Stanton* and *Pannell* cite several of *Reynolds's* major discoveries:

1. that there is a critical point at which fluid flow suddenly changed from “lamellar motion” to “eddy motion” [18.6, p. 200]
2. that the critical velocity is directly proportional to the kinematical viscosity of the water and inversely proportional to the diameter of the tube, and
3. that for geometrically similar tubes, the dimensionless product: (critical velocity)  $\times$  (diameter)/(kinematic viscosity of water) is constant.

*Stanton* and *Pannell* also noted a complication: surface roughness needed to be taken into account; this is

a matter of geometry on a much smaller scale making a difference. However, the overall approach of the use of dimensionless parameters to establish similar situations was still seen to be valid, as indicated by their extensive experiments [18.6, p. 201]:

“From the foregoing it appears that similarity of motion in fluids at constant values of the variable  $vd/\nu$  [velocity  $\times$  diameter/kinematic viscosity of water] will exist, provided the surfaces relative to which the fluids move are geometrically similar, which similarity, as Lord RAYLEIGH pointed out, must extend to those irregularities in the surfaces which constitute roughness. In view of the practical value of the ability to apply this principle to the prediction of the resistance of aircraft from experiments on models, experimental investigation of the conditions under which similar motions can be produced under practical conditions becomes of considerable importance, [...] By the use of colouring matter to reveal the eddy systems at the back of similar inclined plates in streams of air and water, photographs of the systems existing in the two fluids when the value of  $vd/\nu$  was the same for each, have been obtained, and their comparison has revealed a remarkable similarity in the motions.”

In referring to the dimensionless parameter  $vd/\nu$  as a *variable*, what Stanton and Pannell meant was that their equation for the resistance  $R$  includes a function of this dimensionless parameter, that is, resistance  $R = (\text{density}) \times (\text{velocity})^2 \times (\text{some function of } vd/\nu)$ . As they put it,  $R = \rho v^2 F(vd/\nu)$ , where  $F(vd/\nu)$  indicates some unspecified function of  $vd/\nu$ . Hence,  $vd/\nu$  is a variable in the sense that the relation for resistance includes an unspecified function of  $vd/\nu$ . It is also a variable in a more practical sense: it can be physically manipulated.

Stanton and Pannell presented this relation as a consequence of the principle of dynamical similarity (in conjunction with assumptions about what “the resistance of bodies immersed in fluids moving relatively to them” depends on. Evidently, it was Rayleigh who suggested the generalization; they cite *Rayleigh’s* contribution in the Report to the Advisory Committee for Aeronautics, 1909–1910 [18.51, p. 38]. Rayleigh had there spoken of the possibility of taking a more general approach than current researchers were taking in applying the “principle of dynamical similarity.”

In presenting the results they obtained at the National Laboratory in the paper, it is noteworthy that the results are presented in graphs where one of the variables plotted is the term  $R/\rho v^2$ , which is just another expression for the unspecified function, and is dimensionless. What this implies is that the laboratory

experiments are not conceived of in terms of the values of individual measurable quantities such as velocity but in terms of the value of a dimensionless parameter.

Rayleigh, too, presented a kind of survey paper in early 1914, as mentioned above. In that March 1914 paper [18.55], *Rayleigh* noted that the principle of dynamical similarity “allows us to infer what will happen upon one scale of operations from what has been observed at another.” That is, one use of the principle is to use an observation or experiment as representative of a whole class of actual cases: all the other cases to which it is similar, even though the cases may have very different values of measurable individual quantities such as velocity. The important fact of the situation is the dimensionless parameter just mentioned [18.55]:

“It appears that similar motions may take place provided a certain condition be satisfied, viz. that the product of the linear dimension and the velocity, divided by the kinematic viscosity of the fluid, remain unchanged.”

A consequence of this fact is that, even in cases of a different fluid, so long as this dimensionless product is the same, the motions will be similar: no mention of the fluid! Not only is this striking claim correct, but it is responsible for a particularly useful application of Stanton and Pannell’s work, of which they were well aware: tests done on water can be used to infer behavior about systems where the fluid is air. Not because air and water are similar – the relevant fluid properties are very different, in fact – but because the dimensionless parameter relating a number of the features of the fluid and of the situation is the same. Air and water are about as different as can be [18.6, p. 202]:

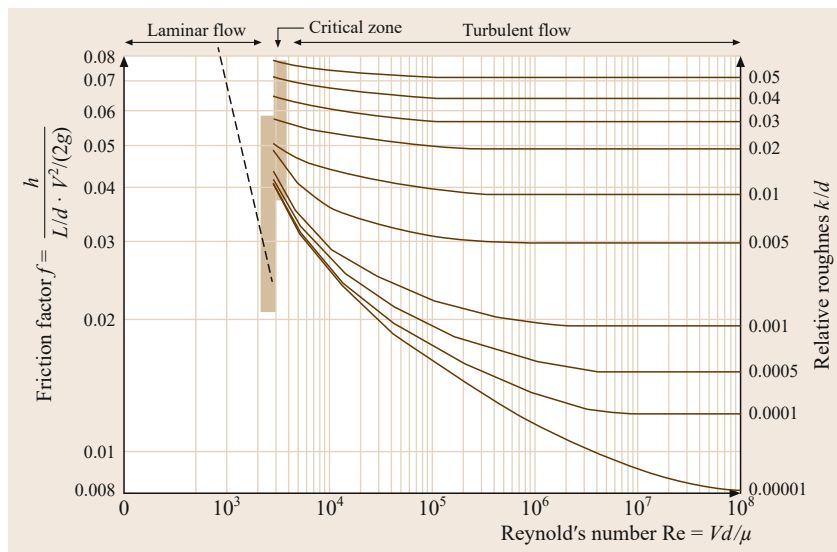
“The fluids used in the majority of the experiments have been air and water. The physical properties of these are so widely different that observations on others are hardly necessary [...]”

Just as the theorem of corresponding states in physical chemistry allowed the construction of a function such that the values for many different kinds of fluids all fell on the same line, so here, too: that the function of the variable  $vd/\nu$  is the same for air, water, and oil is experimentally illustrated by Fig. 18.6 from the paper.

### 18.4.3 Buckingham and Tolman

#### Buckingham’s Background in 1914

Edgar Buckingham (1867–1940) was a physicist who had been working at the National Bureau of Standards in Washington, DC, since 1906. He had little previous experience or background in aeronautics when he began working on issues related to aeronautical research. His



**Fig. 18.6** The approach in Stanton and Pannell’s paper of 1913/1914 expressing experimental results in terms of Reynolds number remains the fundamental approach even today. The chart above, a Moody diagram, illustrates that the fluid behavior for flow in pipes can be expressed in a manner that is independent of the nature of the fluid and other specifics of the configuration

involvement arose as a consequence of efforts afoot to establish a government agency devoted to aeronautical research in the United States, modeled on the British Advisory Committee for Aeronautics; one spot was allocated for a physicist from the National Bureau of Standards [18.61]. How did it end up that it was Buckingham, then, who authored the paper that has become such a landmark in hydrodynamics and aerodynamics? In a letter to Rayleigh in 1915, *Buckingham* explained the origins of his 1914 paper *On Physically Similar Systems: Illustrations of the Use of Dimensional Equations* [18.62]:

“Some three or four years ago, having occasion to occupy myself with practical hydro- and aerodynamics, I at once found that I needed to know more about the method [of dimensions] in order to use it with confidence for my own purposes. Since you and the few others who have made much use of the method of dimensions have generally referred to it somewhat casually as to a subject with which everyone was familiar, I supposed that the hiatus in my education would be easily filled.”

But it was not [18.62]:

“[...] upon looking through your collected papers, the *Sound* [probably a reference to Rayleigh’s *Theory of Sound*], Stokes’s papers, and a few standard books such as Thompson and Tait [*Principles of Mechanics*] and Routh’s *Rigid Dynamics* I was amazed at my failure to find any simple but comprehensive exposition of the method which could be used as a textbook. [...] Each one of your nu-

merous applications of the method seemed perfectly clear, and yet their simplicity gave them the appearance of magic and made the general principle rather elusive.”

It is noteworthy that Buckingham mentions looking at the main *mechanics* textbooks used in Britain, rather than engineering texts. Approaching aerodynamics from the point of view of a physicist was consistent with the kind of community in which Buckingham worked and had been educated. He had earned an undergraduate degree in physics at Harvard University (graduating in 1887) and a doctorate in physics from Leipzig in 1894. Descriptions of him as *an engineer* or *physicist-engineer* as mentioned in *Maila Walter’s* book [18.8] are somewhat misleading. After a few years as a physics professor, Buckingham worked as a physicist at US government agencies; first at the USDA Bureau of Soils (where he did very original theoretical work, applying energy methods), and then at the National Bureau of Standards [18.11]. Involving physicists on aerodynamical research planning made sense, and it also helped cultivate a more prestigious image of a research institution concerned with aerodynamics in 1914. *Buckingham* seemed aware of this, as evidenced by his remark to Rayleigh about the latter’s *Nature* article on the principle of dynamical similarity; he wrote Rayleigh that [18.62]:

“a note, such as the one in *Nature* of March 18th, which has your authority behind it, has an effect far more important in the present state of affairs than any detailed exposition of the subject, however good, because physicists will be sure to read it.”



One of Buckingham's special areas of expertise within physics was thermodynamics. He did not view thermodynamics as merely a subspecialty in physics, though, but rather as an enlightened view of science in which thermodynamics encompassed all of classical mechanics. In his 1900 book, *Outline of a Theory of Thermodynamics*, Buckingham had written [18.63, p. 16]:

“Thermodynamics [...] aims at the study of all the properties or qualities of material systems, and of all the forms of energy which they possess. It must, therefore, be held, in a general sense, to include pure dynamics, which is then to be looked upon as the thermodynamics of systems of which a number of nonmechanical properties are considered invariable. For *thermodynamics*, in this larger sense, the more appropriate name *energetics* is often used, the word *thermodynamics* being reserved to designate the treatment of problems which are directly concerned with temperature and heat.”

Buckingham's approach toward formalizing physics in his 1900 book on the foundations of thermodynamics had been to make the formalism he proposed as flexible as possible, and to build as few assumptions into it as possible. In generalizing the existing science of dynamics, he chose to regard as variable certain properties that are often considered invariable in dynamics. As Buckingham obtained his doctorate in Leipzig under Wilhelm Ostwald, a friend of Boltzmann who was often engaged with him in discussions and debates about foundational issues in science, Buckingham was familiar with debates in philosophy of science [18.11]. Buckingham developed (if he had not already had) a penchant for asking foundational questions, too; in his new role of advisor on research into aeronautics, he set for himself the task of discerning the foundations of the methods he saw being used in aeronautical research.

#### Buckingham's Papers at the Washington Academy of Sciences in 1914

By the middle of 1914, Buckingham had figured out some things about the foundations of the methods used in aerodynamical research. As his note to Rayleigh indicates, he had been concentrating on understanding how “the method of dimensions,” or dimensional analysis, was employed in aerodynamical and hydrodynamical research. On May 23, 1914, he presented a paper entitled *The interpretation of experiments on models* to the Washington Academy of Sciences in Washington, DC, of which he was a member; 27 people were present, and four discussed the paper afterward [18.59]. The account published in the academy's journal stated that:

“The speaker began by deducing a general theorem regarding the form which physical equations must have in order to satisfy the requirement of dimensional homogeneity.”

Dimensional homogeneity is an exceedingly general requirement of an equation; if the terms in an equation have any units (as equations in physics do), the equation is not really considered an equation if it does not meet the requirement of dimensional homogeneity. Thus, this deduction is of something very fundamental in physics; it is about the logic of equations. The account continues [18.59]:

“The theorem may be stated as follows: If a relation subsists among a number of physical quantities, and if we form all the possible independent dimensionless products of powers of those quantities, any equation which describes the relation is reducible to the statement that some unknown function of these dimensionless products, taken as independent arguments, must vanish.”

The antecedent of the theorem is extremely general: “If a relation subsists among a number of physical quantities [...]”; what is striking is that the antecedent of the theorem is not a requirement that the relation mentioned *be known*, only that it *exist*. The theorem was described as a “general summary of the requirement of dimensional homogeneity.” The report on Buckingham's talk added that the method of determining the number and forms of the independent dimensionless products was explained. There is no mention of similar systems in the journal's account of this May 1914 talk, but it does add that the theorem “may be looked at from various standpoints and utilized for various purposes,” and that “several illustrative examples” were given showing the “practical operation of the theorem” [18.59].

In July 1914, the academy's journal featured a short, six-page paper by Buckingham. The topic identified was more general than model experiments, and this time it did mention *similar systems*; in fact, the paper was titled *Physically Similar Systems*. That *Buckingham* meant the July paper to be seen as a generalization of the earlier paper on the interpretation of model experiments was indicated in the closing sentence of the paper [18.2, p. 353]:

“A particular form of this theorem, known as the principle of *dynamical similarity* is in familiar use for the *interpretation of experiments on mechanical models*; but the theorem is equally applicable to problems in heat and electromagnetism” (emphasis added).”

Like the May 1914 talk, the short July 1914 paper is notable for the generality of its approach. It did not imply that there were any set fundamental quantities, nor how many there were. It did not talk about physics, even. It spoke of quantities, relations between quantities, and equations. It is spare and elegant. It begins [18.1]:

“Let  $n$  physical quantities,  $Q$ , of  $n$  different kinds, be so related that the value of any one is fixed by the others. If no further quantity is involved in the phenomenon characterized by the relation, the relation is complete and may be described by an equation of the form  $\Sigma MQ_1^{b_1} Q_2^{b_2} Q_3^{b_3} \dots Q_n^{b_n} = 0$ , in which the coefficients  $M$  are dimensionless or pure numbers”

He makes it clear that it is a matter of choice which units are to be regarded as fundamental ones.

“Let  $k$  be the number of fundamental units needed in an absolute system for measuring the  $n$  kinds of quantity. Then among the  $n$  units required, there is always at least one set of  $k$  which are independent and not derivable from one another, and which might therefore be used as fundamental units, the remaining  $(n - k)$  being derived from them.”

Together, these allow him to say how the quantities *other than* those that are taken to be among the  $k$  fundamental quantities *are related to* those fundamental quantities. Denoting the fundamental units by  $[Q_1]$  through  $[Q_k]$  – in this July 1914 paper he sometimes uses the square brackets indicate the *units* of the enclosed quantity – and the remaining  $(n - k)$  units that are derived from them by  $[P_1]$ ,  $[P_2]$ , and so on up to  $[P_{n-k}]$ , we get  $(n - k)$  equations that relate the units of the  $(n - k)$   $P$ s to the units of the  $k$   $Q$ s. Putting these requirements in terms of dimensions rather than units allows one to apply the requirement of dimensional homogeneity – doing so for each of the fundamental units gives  $k$  equations; each of the  $k$  equations is a result of setting the exponents of one of the units to zero. It can then be shown that the number of independent dimensionless parameters  $\Pi_i$ s is  $(n - k)$  [18.1].

The generality of the treatment here marks this work on similar systems by Buckingham’s off from the earlier work by *Stokes* in 1850 [18.45] and *Helmholtz* in 1873 [18.41]. Whereas *Stokes* spoke of “similar systems, composed of similar solids, oscillating in a similar manner” and of comparing [18.45]:

“similarly situated points in inferring from the circumstance that [the relevant hydrodynamical equations] are satisfied for one system that they will be satisfied for all [the other similar] systems,”

*Buckingham* spoke of an undetermined function whose arguments were dimensionless parameters and he spoke of varying the quantities ( $Q$ s and  $P$ s above) in ways that “are not entirely arbitrary but subjected to the  $(n - k - 1)$  conditions that [certain] dimensionless  $\Pi_i$ ’s remain constant” [18.1].

Putting it in other terms, Buckingham characterized systems as similar in terms of a (nonunique) set of invariants. His emphasis is on the principle of dimensional homogeneity, which is really about the *logic* of the *equations* of physics. The concept of similar systems arises from reflecting on how the principle of dimensional homogeneity might actually be put to use, what it might allow one to infer. After the paper’s opening pages, in which he laid out the observations about the nature of equations that express relations in nature (i. e., wherein the value of one quantity is fixed by the others) stated above, he writes:

“The chief value of the principle of dimensional homogeneity is found in its application to problems in which it is possible to arrange matters so that the [dimensionless ratios]  $r$ ’s and the [dimensionless parameters]  $\Pi$ ’s of [the set of linear equations relating the  $P$ ’s to the  $Q$ ’s and the (unknown) function  $\phi$  of the dimensionless  $r$ ’s and  $\Pi$ ’s] remain constant,”

so that the unknown function  $\phi$  takes on a fixed value, thus giving a definite relation between the  $P$ s and  $Q$ s in terms of the value of the unknown function  $\phi$ . As he remarks, the point is not that dimensional analysis *provides* the function  $\phi$  or even the *value*  $\phi$  takes on once the values of the invariants are set. Rather, the principle allows one to express the relations between quantities in terms of  $\phi$ , which has a fixed value if all its arguments (the dimensionless parameters) are fixed. Hence, doing an experiment on one case yields the relation for all the cases in which the dimensionless parameters that are the arguments of  $\phi$  have the same value, even if the individual quantities from which those parameters are formed are all different.

Though Buckingham was, he said, only aiming to give a clear treatment of the same idea that *Stokes* and others had stated, a lot had happened in mathematics and physics (especially in physical chemistry and thermodynamics), in the intervening decades. In their works on similar systems, *Stokes* and *Helmholtz* worked with physical equations, the partial differential equations of fluids and fields; *Buckingham*, as a physicist, was certainly cognizant of and competent in working with them, too, but in the July 1914 paper on similar systems, he worked with (more abstract) dimensional equations. The goal here, in this lean paper that featured no examples or applications, was to get straight on things that

(so far as he was aware) had not yet been articulated by others who had employed the method. He would later write to Rayleigh about these first papers on the method [18.62]:

“I had therefore [...] to write an elementary textbook on the subject for my own information. My object has been to reduce the method to a mere algebraic routine of general applicability, making it clear that Physics came in only at the start in deciding what variables should be considered, and that the rest was a necessary consequence of the physical knowledge used at the beginning; thus distinguishing sharply between what was assumed, either hypothetically or from observation, and what was mere logic and therefore certain.

The resulting exposition is naturally, in its general form, very cumbersome in appearance, and a large number of problems can be handled vastly more simply without dragging in so much mathematical machinery.”

His exposition treats of a system  $S$  characterized very abstractly: “The quantities involved in a physical relation pertain to some particular physical system which may usually be treated as of very limited extent” [18.1, p. 352]. The system constructed to be similar to it, likewise, is described very formally [18.1, p. 352]:

“Let  $S'$  be a second system into which  $S$  would be transformed if all quantities of each kind  $Q$  involved in [the equation expressing the physical relation pertaining to the system] were changed in some arbitrary ratio, so that the  $r$ 's for all quantities of these kinds remained constant, while the particular quantities  $Q_1, Q_2, \dots, Q_k$  changed in  $k$  independent ratios.”

After completing the specification of the constraints on how the quantities change in concert with each other so that  $S'$  also satisfies the relation: “Two systems  $S$  and  $S'$  which are related in the manner just described are similar as regards the physical relation in question.” [18.1, p. 352]

The exposition may have been cumbersome, but the point is elegant and spare: the constraints that must be satisfied in constructing the system  $S'$  are just these: to keep the value of the dimensionless parameters that appear in the general form of the equation – the arguments of the function  $\phi$  – the same in  $S'$  as in  $S$ . So, what is crucial is to identify a set of dimensionless parameters that can serve as the arguments of the undetermined function  $\phi$ . For Buckingham, unlike for some predecessors writing about similar systems or dynamic similarity, the method underlying the construction of

physically similar systems is not a method peculiar to mechanics; it applies to *any equation describing a complete relation that holds between quantities*.

#### Richard Chace Tolman's *Principle of Similitude*

Meanwhile, another physicist in the United States was publishing on similitude, too, though with considerably less rigor. Richard Chace Tolman (1881–1948) was an assistant professor of the relatively new field of physical chemistry at the University of California when Onnes won the Nobel Prize for his work in physical chemistry on the liquefaction of helium; Onnes delivered his Nobel Prize Lecture in December 1913 [18.31, 64]. As noted above, *Onnes* had aimed to “demonstrate that the principle of corresponding states can be derived on the basis of what he calls the principle of similarity of motion, which he ascribes to Newton” [18.32].

*Tolman* published *The Principle of Similitude* in the March 1914 *Physical Review*, in which he proposed the following [18.58, p. 244]:

“The fundamental entities out of which the physical universe is constructed are of such a nature that from them a miniature universe could be constructed exactly similar in every respect to the present universe.”

*Tolman* then (he claimed) showed that he could derive a variety of laws, including the ideal gas law, from the principle of similitude he had proposed, proceeding in somewhat the same way as *Onnes* had proceeded in showing that the principle of corresponding states was a consequence of mechanical similarity. *Tolman* seemed to appeal to a criterion that the two universes should be observationally equivalent [18.58, p. 245]

“[...] let us consider two observers,  $O$  and  $O'$ , provided with instruments for making physical measurements.  $O$  is provided with ordinary meter sticks, clocks and other measuring apparatus of the kind and size which we now possess, and makes measurements in our present physical universe.  $O'$ , however, is provided with a shorter meter stick, and corresponding altered clocks and other apparatus so that he could make measurements in the miniature universe of which we have spoken, and in accordance with our postulate obtain exactly the same numerical results in all his experiments as does  $O$  in the analogous measurements made in the real universe.”

He brings up some other considerations, some from physics (Coulomb's law), some from the theory of dimensions, and then tries to show how various physical relations, such as the ideal gas law, can be deduced from simple physical assumptions and his proposed principle

of similitude. For relations involving gravitation, however, a contradiction arises; his response is to use the contradiction as motivation to propose a new criterion for an acceptable theory of gravitation. He concludes that his proposed principle is a new relativity principle: the “principle of the relativity of size” [18.58, p. 255].

*Tolman* believes that, in his paper, he has laid out transformation equations that specify the changes that have to be made in lengths, masses, time intervals, energy quantities, etc., in order to construct a miniature world such that [18.58, p. 255]:

“If, now, throughout the universe a simultaneous change in all physical magnitudes of just the nature required by these transformation equations should suddenly occur, it is evident that to any observer the universe would appear entirely unchanged. The length of any physical object would still appear to him as before, since his meter sticks would all be changed in the same ratio as the dimensions of the object, and similar considerations would apply to intervals of time, etc. From this point of view we can see that it is meaningless to speak of the absolute length of an object, all we can talk about are the relative lengths of objects, the relative duration of lengths of time, etc. The principle of similitude is thus identical with the principle of the relativity of size.”

*Tolman*’s suggestion differs from the concept of similar systems mentioned so far, though the difference may not be obvious. Others working on similar systems where quantities or paths were homologous between similar systems noted that there were limits of applicability; they recognized the fact that there are ranges in which size matters (e.g., surface tension matters disproportionately at small scales [18.21]; the restriction in *Helmholtz*’ 1873 paper that velocities must be small with respect to the velocity of sound [18.41], Reynolds’ recognition of the role of “mean range” of molecules in transpiration [18.11]). *Helmholtz* even explicitly discussed the practical difficulties of constructing models of a different size than the configuration modeled, raising the question of whether in some cases it may not be possible to do so [18.41]. *Tolman* not only does not recognize such limits; he suggests making the denial that they exist a principle of physics. It seems pretty clear that *Tolman* is here modeling his exposition on Einstein’s 1905 paper on the special theory of relativity. *Tolman* proposes that the relativity of size be regarded along the lines of the relativity of motion: in his paper on special relativity, Einstein had considered it a principle that observers cannot tell one state

of unaccelerated motion from another; *Tolman* proposes to do the same for the statement that observers not be able to distinguish an appropriately constructed model universe from the actual one [18.58], if inhabiting it as an appropriately transformed being and using appropriately constructed or transformed instruments. There is a confusion in *Tolman*’s reasoning. While it is quite natural to say that a desirable principle of nature, and a desirable constraint on measuring systems, is that it should not matter to the project of pursuing truth that one observer in the actual world is using one system of measurement and another observer in the actual world is using another system of measurement, *Tolman* seems here to be confusing that requirement with a requirement that miniature universes constructed from the materials of the actual universe be indistinguishable from the actual, full size, universe by the miniature observers inhabiting those miniature universes.

#### Buckingham’s *Physical Review* Paper and Reply to Richard Chace Tolman

It’s rather obvious that the notion of similar systems – one system being transformed into another system  $S'$  in such a way that it “corresponds” to  $S$  (“as regards the essential quantities”) – is relevant to evaluating the claim *Tolman* made in his 1914 *Principle of Similitude* paper [18.58] that the universe could be transformed overnight into an observationally indistinguishable miniature universe. The notion of similar systems is also relevant to *Stanton* and *Pannell*’s *Similarity of Motion* paper [18.6], in that it is a more general treatment of the methodology of model testing (“the principle of dynamical similarity” [18.6, p. 201]) given there. In the next paper that *Buckingham* wrote on the topic [18.2], in addition to presenting the generalized treatment found in the July 1914 version of *Physically Similar Systems*, he addressed both these related topics on which major papers had appeared in the earlier part of the year: experimental models and *Tolman*’s claims about the possibility of an observationally indistinguishable miniature universe. The October 1914 *Physical Review* featured *Buckingham*’s *On Physically Similar Systems: Illustrations of the Use of Dimensional Equations*; his manuscript is dated June 18th of that year [18.2].

In his 1914 *Physical Review* paper [18.2], *Buckingham* says that his purpose in presenting how the notion of physically similar systems can be developed from the principle of dimensional homogeneity in that paper was to provide the background against which to respond to *Tolman*’s proposed “principle of similitude” [18.2, p. 356]. He makes several points relevant to addressing *Tolman*’s proposal for a new principle in physics in

developing “the notion of physical similarity” and “the notion of physically similar systems”:

1. It is only “the phenomenon characterized by the relation [expressed by the equation whose existence was assumed at the start]” that “occurs in a similar manner” in both systems: “we say that the bodies or systems are *similar with respect to this phenomenon* (emphasis added).” *Buckingham* specifically points out that systems that are “said to be *dynamically similar*” might not be similar “as regards some other dynamical relation”; two dynamically similar systems might not “behave similarly in some different sort of experiment.”
2. There is a more general conception of similarity than dynamical similarity, and it too “follows directly from the dimensional reasoning, based on the principle of homogeneity.”
3. Tolman’s proposed *Principle of Similitude* is not clearly stated, but inasmuch as *Buckingham* understands it, it seems to him “merely a particular case” of the theorem *Buckingham* presents in the paper. *Buckingham* reasons as follows: The way *Tolman* proceeds is to select four specific independent kinds of quantity (length, speed, quantity of electricity, electrostatic force), subjects these four kinds of quantity to four arbitrary conditions, then finds the conditions that some other kinds of quantities are subject to “in passing from the actual universe to a miniature universe that is physically similar to it” [18.2, p. 356]. I take *Buckingham*’s point to be that, inasmuch as what *Tolman* is concluding is correct, it can be concluded using the principle of dimensional homogeneity without the aid of the “new” principle that *Tolman* proposed in his March 1914 *Physical Review* paper.

Having already remarked that the notion of similar systems used in constructing and using a model propeller is generalizable beyond mechanics, he then goes on to show how the principle involved in doing so – the “method of dimensions” [18.65, p. 696] – applies in problems ranging from electrodynamics (energy density of a field, the relation between mass and radius of an electron, radiation from an accelerated electron) to thermal transmission, and, finally, at a higher level, to the kind of *bird’s-eye view* question to which his interest tended to migrate: “the relation of the law of gravitation to our ordinary system of mechanical units.”

The question he asks about the role of the law of gravitation in determining units of measure is a bit different. It is about the number of *fundamental units*, and the question *Buckingham* asks can be put in terms of similar systems: if it is, in fact, true that in mechanics

three fundamental units suffice to describe mechanical phenomena (more if thermal and electromagnetic phenomena are to be described), then it would be correct to conclude that [18.2, pp. 372–373]:

“[A] purely mechanical system may be kept similar to itself when any three independent kinds of mechanical quantity pertaining to it are varied in arbitrary ratios, by simultaneously changing the remaining kinds of quantity in ratios specified by [the constraint of dimensional homogeneity . . .] For instance, we derive a unit of force from independent units of mass, length, and time, by using these units in a certain way which is fixed by definition, and we thereby determine a definite force which is reproducible and may be used as a unit. Now by Newton’s law of gravitation it is, in principle, possible to derive one of the three fundamental units of mechanics from the other two.”

*Buckingham* then describes a laboratory experiment from which a unit of time can be derived from units of mass and length – *if* one assumes Newton’s law of gravitation to hold. To be clear: *Buckingham* is granting that people have sometimes reduced the number of fundamental units to two, such as when a unit of time is derived from units for mass and length, when working on specific problems. What he is concerned to show is that, in order to do so, they have had to use assumptions about the law of gravitation. He is not unaware that the current state of physics indicates Newton’s law of gravitation is not the final word, and is pointing out the role that a law of gravitation plays in such reductions of the number of fundamental units to two. Put in terms of similar systems, the question is: How many degrees of freedom do we have in constructing a system  $S'$  that is similar to  $S$ ? How many quantities can be varied in an arbitrary ratio when we transform  $S$  into  $S'$ , a system that is physically similar to it?

*Buckingham* points out that, even in the domain of mechanics, it depends on what phenomenon the relation between quantities characterizes. As he emphasized, the notion of physical similarity and physically similar systems involves only similarity *with respect to* a specified relation. (Recall that the analysis started with the quantities involved in a given equation, where that equation describes a relation that relates a certain number of kinds of quantities such that any one was determined by all the others, and the relation characterized a phenomenon of interest.) In developing a general methodology, *Buckingham* had considered all possible relations that could exist among the given kinds of quantities. In the most general case, the law of gravitation is a constraint on how quantities are related. Recognizing this additional constraint reduces the

number of independent quantities by one. However, he explains, such generality is not always required in practice [18.2, p. 374]:

“But if for ‘all possible relations’ we substitute ‘all relations that do not involve the law of gravitation,’ we may ignore the law and proceed as if it were non-existent.”

This can actually be done in many cases, he says, since [18.2, p. 374]:

“in practice, physicists are seldom concerned with the law of gravitation: for all our ordinary physical phenomena occur subject to the attraction of an earth of constant mass and most of them occur under such circumstance that the variation of gravity with height is of no sensible importance.”

However, for precise geodesy and astronomy, one needs to be explicit about the law of gravitation.

Buckingham’s answer to the question Tolman’s paper raises about the possibility of constructing observationally indistinguishable miniature universes, thus, bifurcates into two cases, depending on whether or not the phenomenon that we are interested in observing in the miniature universe is influenced by the law of gravitation. If not, then it might not be impossible to construct a miniature universe, as Tolman suggests, that will be similar to the universe (as regards that phenomenon.) On the other hand, if the phenomenon is influenced by the law of gravitation, more things must be taken into account:

“the gravitational forces in the miniature universe must bear to the corresponding gravitational forces in the actual universe a ratio fixed by the law of gravitation.”

He points out that the effect of the law of gravitation on the phenomena of interest shows up in the process of constructing similar systems. If we erroneously try to independently choose three units rather than letting the third be determined by the first two fundamental units chosen, we run into trouble because the measured values for corresponding speeds and forces will not correspond to the values in the actual universe – unless, that is, the third unit is allowed to be fixed by the law of gravitation in terms of the first two.

The points about physically similar systems, systems of units, and the law of gravitation seem to be questions in the logic of physics. Yet, the main claim of Buckingham’s papers on physically similar systems can actually be stated in terms of a theorem about the *symbolism of relations* between physical quantities.

This is seen in the “convenient summary” with which he concludes the paper [18.2, p. 376]:

“A convenient summary of the general consequence of the principle of dimensional homogeneity consists in the statement that any equation which describes completely a relation subsisting among a number of physical quantities of an equal or smaller number of different kinds, is reducible to the form  $\Psi(\Pi_1, \Pi_2, \dots, \Pi_i, \text{etc.}) = 0$  in which the  $\Pi$ ’s are all the independent dimensionless products of the form  $Q_1^x Q_2^y \dots$ , etc. that can be made by using the symbols of all the quantities  $Q$ .”

The equation  $\Psi(\Pi_1, \Pi_2, \dots, \Pi_i, \text{etc.}) = 0$  in the quote from Buckingham above is what I called *The Reduced Relation Equation of 1914* in Sect 18.1 of this chapter.

#### 18.4.4 Precursors of the Pi-Theorem in Buckingham’s 1914 Papers

This chapter is devoted to the history of the notion of physically similar systems. Buckingham’s 1914 papers are considered a landmark in the development of our current notion of physically similar systems, due to the articulation of what a physically similar system is and how it is related to the symbolism used to express relations in physics. First, Buckingham showed that *The Reduced Relation Equation of 1914* followed from the principle of the homogeneity of a physical equation. Then, he showed how the notion of *physically similar systems* could be developed from it.

However, since Buckingham’s name has since become attached to the so-called *pi-theorem*, and the full contents of his 1914 papers are often ignored, being inaccurately viewed as doing little more than presenting the pi-theorem, I want to emphasize that what has become known as the pi-theorem itself is not actually due to Buckingham. There were, in fact, many precursors who proved the same result, with varying levels of generality.

##### Vaschy and Bertrand

The *pi-theorem* is referred to in France as the Vaschy–Buckingham Pi-Theorem. In 1892, Vaschy (1857–1899) published *Sur les lois de similitude en physique* [18.66, 67], in which he stated the result about the number of parameters required to state a given relationship that is often attributed to Buckingham. However, unlike Buckingham, Vaschy did not mention dimensions or dimensional equations. He spoke of quantities and units, and did so as though they were the same sort of thing, though he did speak of some units

as fundamental and others as derived. More precisely, *Vaschy's* theorem is [18.67]:

“Let  $a_1, a_2, a_3, \dots, a_n$  be physical quantities, of which the first  $p$  are distinct fundamental units and the last  $(n-p)$  are derived from the  $p$  fundamental units (e.g.,  $a_1$  could be a length,  $a_2$  a mass,  $a_3$  a time, and the  $(n-3)$  other quantities would be forces, velocities, etc.; then  $p = 3$ ). If between these  $n$  quantities there exists a relation  $F(a_1, a_2, a_3, \dots, a_n) = 0$ , which remains the same whatever the arbitrary magnitudes of the fundamental units, this relationship can be transformed in another relationship between at most  $(n-p)$  parameters, that is  $f(x_1, x_2, x_3, \dots, x_{n-p}) = 0$ , the parameters  $x_1, x_2, x_3, \dots, x_{n-p}$  being monomial functions of  $a_1, a_2, a_3, \dots, a_n$ .”

The parameters  $x_1, x_2, x_3, \dots, x_{n-p}$  play the same role as the dimensionless  $\Pi$ 's in Buckingham's theorem. *Vaschy* then shows how to obtain reduced relations for the pendulum and for a telegraph cable. What is notable is that he produces a pair of ratios, not just one ratio, in each case, and he expresses the result as an unknown function of these parameters ( $x_i$ 's) set equal to zero. He does not use the terminology of systems, but he is interested in laws of similitude (in the sense of the similarity laws of Sect. 18.3.1) that can be derived from them, citing one by W. Thomson (Lord Kelvin) in the case of the telegraph line. The conditions of *Vaschy's* theorem are not exactly the same as in Buckingham's theorem, but *Vaschy* does emphasize that his reasoning does not assume any particular system of units, and he does derive the key move to the *Reduced Relation Equation of 1914*. The case is strong for crediting *Vaschy's* paper with containing the *pi-theorem*.

Some have also argued that Joseph Bertrand provided an even earlier, though less general, proof of the pi-theorem in 1878, in *Sur l'homogeneite dans les formules de physique* [18.66, p. 209]. This is the same Joseph Bertrand (1822–1900) cited above for the much earlier 1847 work drawing attention to the principle of similitude, in which he mentioned “an infinite number of possible systems, which may be regarded as similar to” a given system, and provided a new basis for Newton's theorem of similarity using a result by Cauchy involving the principle of virtual velocities.

These two works by *Bertrand* [18.28, 68] thirty years apart reflect an important late nineteenth century development that permitted using a logical principle about *the equations of physics*, that is, the homogeneity of equations of physics, rather than a principle of physics itself. This late nineteenth-century development was the idea of coherence as a constraint on a system of units; the idea, that is, of a coherent system of units. Coherence of a system of units, and its importance in connecting dimensional analysis and similarity, is discussed in [18.69].

### Riabouchinsky

Sometime after 1914, Buckingham became aware that *Dimitri Riabouchinsky* (1882–1962) had also proved a mathematical theorem about the number of dimensionless parameters needed to express a given physical relation, using the methods of dimensional analysis, in 1911 [18.65]. *Riabouchinsky* (spelled *Riabouchinski* in Buckingham's papers), was a scientist who had provided the private funding for the Aerodynamic Institute of Koutchino associated with the University of Moscow, which had a wind tunnel; hence, *Riabouchinsky* was, like Buckingham, faced with the problem of understanding how to interpret model experiments. After becoming aware of *Riabouchinsky's* proof, *Buckingham* credited him prominently for the proof in his writings. In a paper in 1921, discussing the desire that had arisen for a more systematic procedure for obtaining the results that Rayleigh and others had obtained using dimensional methods, he wrote [18.70, p. 696]:

“Such a routine procedure is provided by formulating the requirement of dimensional homogeneity as a general algebraic theorem, which was first published by *Riabouchinski* (sic), and which will be referred to as the  $\Pi$  theorem.”

*Buckingham* speculated that he might have seen a notice of *Riabouchinski's* result in one of the Annual Reports of the British Advisory Committee on Aeronautics [18.71], and that [18.70, p. 696n]:

“Guided [...] by the hint contained in this abstract, the present writer came upon substantially the same theorem. [...] The theorem does not differ materially from *Riabouchinski's*, except in that he confined his attention to mechanical quantities.”

## 18.5 Physically Similar Systems: The Path in Retrospect

We are now in a position to survey the path from Newton's theorem about similar systems of bodies in the seventeenth century to Buckingham's development of the notion of similar systems from what I have called the *Reduced Relation Equation of 1914*, in the early twentieth century. Painting what we can see in retrospect in broad brush strokes, the picture of this path is that there are several key ideas that made the twentieth century notion of physically similar systems possible. The first of these is the notion of a function developed in the eighteenth century, and the second is the notion of a coherent system of units developed in the late nineteenth century.

Brian Hepburn identifies Leonhard Euler as a key eighteenth century figure linking Newton's age and ours, and has argued that the concept of a function was crucial to the development of what we now know as Newtonian mechanics. Whereas Newton's mechanics "dictated how motions are generated in time by forces" and "would treat of the actual process of moving bodies," Hepburn says, for Euler, in contrast, "the central object of investigation in mechanics is the [mathematical] function" [18.12]. He points out that equilibrium relations are the most important among relations, and hence that "sets of quantities" characterized "states" – I would amend this to "states of a system." The notion of a function allowed the concept of a system to be expressed in terms of the interrelatedness of some quantities – if one quantity changed, any of the others in the system might be affected, too. The notation of a function set to 0, that is,  $f(x_1, x_2, \dots, x_n) = 0$  can be used to express this interrelatedness. The notion of equilibrium and an equation of state, which are expressible by the functional notation, are important in this newer notion of a system; what this new notion of system eventually replaced was the notion of a system as a configuration of particles and/or bodies. The notion of a similarity law likewise progressed from simply a single ratio to express an invariant relation, to a function with multiple arguments, each of which was a dimensionless ratio.

When Bertrand invoked the principle of virtual velocities in 1847 [18.25, p. 380] to derive the principle of mechanical similitude, he was using the notion of a function, but he was still using considerations and principles of mechanics. By 1878, he could take a much more general approach, using a principle that was a constraint on the equation expressing relations between the physical quantities, rather than the system of bodies and particles itself. Independently, many others could do so, too: Vaschy in France and Riabouchinsky in Russia, and they were not the only ones. In physical chem-

istry, van der Waals and Onnes, thinking of collections of molecules as systems, could apply these more formal notions of similar systems to come up with a way to predict the behavior of one substance based on only its critical points, along with observations about how another substance behaved. The amazing success of this approach in physical chemistry seems to have encouraged extending the approach of similar systems to electromagnetic theory and the kinetic theory of gases.

That the time was right in 1914 for deriving the pi-theorem and the Reduced Relation Equation of 1914 is clear from the fact that so many had already done it by then. That Buckingham was the one to write what has become the landmark paper articulating the notion of physically similar systems, which he developed from the *Reduced Relation Equation of 1914* in the  $\Pi$ -theorem, then, appears to be a matter of timing, at least in part: when he was suddenly asked to devote time to the question of the value of model experiments using wind tunnels, it was the early twentieth century, when the notion of a system was readily expressible by the notation for a function, when coherent systems of units in every part of physics was something that could be assumed, and someone with a doctorate in physics would have a facility with formal methods applied to equations.

Around the same time, or shortly thereafter, *D'Arcy Wentworth Thompson* wrote his classic work, *On Growth and Form* [18.72], on the mathematicization of biology. In that work, he carried the use of similitude in physics over into biology and he, too, explicitly cites Newton (for his use of similitude), as well as Galileo (for his discussion of scaling and similitude), Boltzmann, Helmholtz and numerous publications on aerial flight. A detailed discussion of D'Arcy Thompson on similitude may be found in Chap. 6 (*The Physics of Miniature Worlds*) of *Wittgenstein Flies a Kite* [18.11, pp. 117–130].

How do things stand today, in the early twenty-first century? Certainly, there are pockets in many disciplines – physics, hydrodynamics, aerodynamics, the geological and other sciences, hydrology, mechanics, biology, and more – where researchers recognize the value of thinking in terms of physically similar systems. However, it is not really a staple of the basic curriculum. Few philosophers of science understand the concept or why it is significant. This article is offered to help improve at least the latter situation.

**Acknowledgments.** Work on this paper was supported in part by a Visiting Fellowship at the Center for Philosophy of Science at the University of Pittsburgh in



2010, during which my project was the history of the concept of physically similar systems. This paper also incorporates some earlier work published in Chaps. 6 (*The Physics of Miniature Worlds*) and 7 (*Models of Wings and Models of the World*) of *Wittgenstein Flies*

*A Kite: A Story of Models of Wings and Models of the World*. Thanks also to Brian Hepburn and George Smith for conversations about Newton's use of similar systems, and to Jasmin Ozel for translating parts of Forchheimer's *Hydraulik*.

## References

- 18.1 E. Buckingham: Physically similar systems, *J. Wash. Acad. Sci.* **93**, 347–353 (1914)
- 18.2 E. Buckingham: On physically similar systems: Illustrations of the use of dimensional equations, *Phys. Rev.* **4**, 345–376 (1914)
- 18.3 I. Newton, A. Motte, F. Cajori, R.T. Crawford: *Sir Isaac Newton's Mathematical Principles of Natural Philosophy and His System of the World* (Univ. California Press, Berkeley 1946)
- 18.4 E. Mach: *The Science of Mechanics: A Critical and Historical Account of Its Development*, 6th edn. (Open Court Pub. Co, La Salle, Ill 1960), transl. by T. J. McCormack. New introduction by K. Menger with revisions through the ninth German edition
- 18.5 S.D. Zwart: Scale modelling in engineering: Froude's case. In: *Philosophy of Technology and Engineering Sciences*, Vol. 9, ed. by A.W.M. Meijers (North Holland/Elsevier, Amsterdam 2009) pp. 759–798
- 18.6 T.E. Stanton, J.R. Pannell: Similarity of motion in relation to the surface friction of fluids, *Philos. Trans. R. Soc. A* **214**, 199–224 (1914)
- 18.7 A. F. Zahm: *Theories of Flow Similitude*, Report No. 287 (National Advisory Committee for Aeronautics, Washington DC 1928)
- 18.8 M.L. Walter: *Science and Cultural Crisis: An intellectual biography of Percy Williams Bridgman (1882–1961)* (Stanford Univ. Press, Stanford 1990)
- 18.9 E.T. Layton: Escape from the Jail of Shape: Dimensionality and engineering science. In: *Technological Development and Science in the Industrial Age: New Perspectives on the Science-Technology Relationship*, ed. by P. Kroes, M. Bakker (Kluwer, Dordrecht, Boston 1992)
- 18.10 Galilei Galileo, S. Drake (Transl.): *Two New Sciences: Including Centers of Gravity and Force of Percussion*, 2nd edn. (Wall Emerson, Toronto 2000)
- 18.11 S.G. Sterrett: *Wittgenstein Flies a Kite: A Story of Models of Wings and Models of the World* (Pi Press/Penguin, New York 2006)
- 18.12 B.S. Hepburn: Equilibrium and Explanation in 18th Century Mechanics, Ph.D. Thesis (University of Pittsburgh, Pittsburgh 2007)
- 18.13 J. Thomson: Comparison of similar structures as to elasticity, strength, and stability, *Trans. Inst. Eng. Shipbuild. Scotl.* **54**, 361 (1875)
- 18.14 A. Barr: Comparisons of similar structures and machines, *Trans. Inst. Eng. Shipbuild. Scotl.* **42**, 322–360 (1899)
- 18.15 R. P. Torrance: Use of models in engineering design, *Engineering News*, 18 December (1913)
- 18.16 J. Thomson: Comparison of similar structures as to elasticity, strength and stability. In: *Collected Papers in Physics and Engineering*, ed. by J. Larmor, J. C. Thomson (Cambridge Univ. Press, Cambridge 1912) pp. 361–372
- 18.17 A.H. Gibson: *Hydraulics and its Applications* (D. Van Nostrand Company, New York 1908)
- 18.18 O. Darrigol: *Worlds of Flow: A History of Hydrodynamics from the Bernoullis to Prandtl* (Oxford Univ. Press, Oxford, New York 2005)
- 18.19 R.H.M. Robinson: Experimental model basin – I, *Sci. Am. Suppl.* **66**, 37–38 (1908)
- 18.20 G. Hagler: *Modeling Ships and Space Craft: The Science and Art of Mastering the Oceans and Sky* (Springer, New York 2013)
- 18.21 W. Froude: *On the Rolling of Ships* (Parker, Son, and Bourn, London 1862)
- 18.22 S. Schaffer: Fish and Ships: Models in the age of reason. In: *Models: The Third Dimension of Science*, ed. by S. de Chadarevian, N. Hopwood (Stanford Univ. Press, Stanford 2004)
- 18.23 W. Froude: On Experiments with HMS Greyhound, *Trans. R. Inst. Nav. Archit.* **15**, 36–73 (1874)
- 18.24 W. Denny: Mr Mansel's and the late Mr Froude's Methods of analysing the results of progressive speed trials, *Trans. Inst. Eng. Shipbuild. Scotl.* **28**, 1–8 (1885)
- 18.25 F. Cajori: *A History of Mathematics* (Macmillan, New York 1894)
- 18.26 R. Ettema: *Hydraulic Modeling Concepts and Practice* (American Society of Civil Engineers, Reston 2000)
- 18.27 F. Reech: *Cours de Mécanique d'Après la Nature Généralement Flexible et Élastique des Corps* (Carilian-Gœury et Vor. Dalmont, Paris 1852), in French
- 18.28 M.J. Bertrand: On the relative proportions of machinery, considered with regard to their powers of working, *Newton's Lond. J. Arts Sci.* **31**, 129–131 (1847)
- 18.29 H.A. Lorentz: The theory of radiation and the second law of thermodynamics, *KNAW Proceedings*, Vol. 3 (Huygens Institute, Royal Netherlands Academy of Arts and Sciences, Amsterdam 1901) pp. 436–450
- 18.30 The Nobel Prize in Physics 1910, [http://www.nobelprize.org/nobel\\_prizes/physics/laureates/1910/](http://www.nobelprize.org/nobel_prizes/physics/laureates/1910/)
- 18.31 The Nobel Prize in Physics 1913, [http://www.nobelprize.org/nobel\\_prizes/physics/laureates/1913/](http://www.nobelprize.org/nobel_prizes/physics/laureates/1913/)

- 18.32 H.K. Onnes: *Investigations into the Properties of Substances at Low Temperatures, Which have led, Amongst Other Things, to the Preparation of Liquid Helium, Nobel Lectures, Physics 1901–1921* (Elsevier Publishing Company, Amsterdam 1967)
- 18.33 J.M.H. Levelt Sengers: *How Fluids Unmix: Discoveries by the School of Van der Waals and Kamerlingh Onnes* (Koninklijke Nederlandse Akademie van Wetenschappen, Amsterdam 2002)
- 18.34 J. Wisniak: Heike Kamerlingh – The virial equation of state, *Indian J. Chem. Technol.* **10**, 564–572 (2003)
- 18.35 J. Mehra, H. Rechenberg: *The Historical Development of Quantum Theory*, Vol. 5 (Springer, New York 1987)
- 18.36 S.G. Brush: Ludwig Boltzmann and the Foundations of Natural Science. In: *Ludwig Boltzmann (1844–1906): Zum Hundertsten Todestag*, ed. by I.M. Fasel-Boltzmann, G.L. Fasel (Springer, Vienna 2006)
- 18.37 J.C. Maxwell: On Boltzmann's Theorem on the average distribution of energy in a system of material points. In: *The Scientific Papers of James Clerk Maxwell*, Vol. 2, ed. by W.D. Niven (Cambridge Univ. Press, Cambridge 1890)
- 18.38 L. Boltzmann: Model. In: *Encyclopedia Britannica*, Vol. 30, 10th edn., ed. by EDITOR ("The Times" Printing House, London 1902) pp. 788–791
- 18.39 L. Boltzmann: On aeronautics. In: *Wittgenstein Flies a Kite: A Story of Models of Wings and Models of the World*, ed. by S. G Sterrett (Pi Press/Penguin, New York 2005/6) Transl. I. Pollman, M. Mertens
- 18.40 C. Abbe: *Mechanics of the Earth's Atmosphere: A Collection of Translations*, Smithsonian Miscellaneous Collections Ser., Vol. 843 (The Smithsonian Institution, Washington DC 1891) original in German: H. von Helmholtz
- 18.41 H. von Helmholtz: On a theorem relative to movements that are geometrically similar in fluid bodies, together with an application to the problem of steering balloons. In: *Mechanics of the Earth's Atmosphere: A Collection of Translations*, by Cleveland Abbe. Smithsonian Miscellaneous Collections, Vol. 843, (The Smithsonian Institution, Washington DC 1891) pp. 67–77
- 18.42 H. von Helmholtz: Über ein Theorem, geometrisch ähnliche Bewegungen flüssiger Körper betreffend, nebst Anwendung auf das Problem, Luftballons zu lenken, Monatsber. Kgl. Preuß. Akad. Wiss. **1873**, 501–514 (1873)
- 18.43 H. von Helmholtz: On discontinuous motions in liquids. In: *Mechanics of the Earth's Atmosphere: A Collection of Translations*, Smithsonian Miscellaneous Collections, Vol. 843, ed. by C. Abbe (The Smithsonian Institution, Washington DC 1891) pp. 58–66
- 18.44 H. von Helmholtz: Über discontinuirliche Flüssigkeits-Bewegungen, Monatsber. Kgl. Preuß. Akad. Wiss. **1868**, 215–228 (1868)
- 18.45 G.G. Stokes: On the effect of the internal friction of fluids on the motion of pendulums, *Trans. Camb. Philos. Soc.* **9**, 8 (1850)
- 18.46 D. Cahan: Helmholtz and the British Scientific Elite: From force conservation to energy conservation, *Notes Rec. R. Soc.* **66**, 55–68 (2011)
- 18.47 D.M. McDowell, J.D. Jackson: *Osborne Reynolds and Engineering Science Today* (Manchester Univ. Press, Manchester 1970)
- 18.48 O. Reynolds: Letter to George Stokes, April 25, 1883. In: *Memoir and Scientific Correspondence of the Late Sir George Gabriel Stokes, Bart*, Vol. 1, (Cambridge Univ. Press, Cambridge 1907) p. 233, Selected and arranged by Joseph Larmor
- 18.49 G.G. Stokes, J. Larmor (Eds.): *Memoir and Scientific Correspondence of the Late Sir George Gabriel Stokes, Bart*, Vol. 1 (Cambridge Univ. Press, Cambridge 1907), Selected and arranged by Joseph Larmor
- 18.50 L. Prandtl: Motion of fluids with very little viscosity, english transl. In: *Early Developments of Modern Aerodynamics*, ed. by J.A.K. Ackroyd, B.P. Axcell, A.I. Ruban (Butterworth-Heinemann, Oxford 2001)
- 18.51 J. W. Strutt (Baron Rayleigh): Note as to the application of the principle of dynamical similarity. In: *Report of the Advisory Committee for Aeronautics 1909–1910* (London 1910)
- 18.52 *Report of the Advisory Committee for Aeronautics 1910–1911* (British Advisory Committee for Aeronautics, London 1911)
- 18.53 *Report of the Advisory Committee for Aeronautics 1911–1912* (British Advisory Committee for Aeronautics, London 1912)
- 18.54 *Report of the Advisory Committee for Aeronautics 1912–1913* (British Advisory Committee for Aeronautics, London 1913)
- 18.55 J.W. Strutt (Baron Rayleigh): Fluid motions, lecture delivered at the Royal Institute, March 20, *Engineering* **97**, 442–443 (1914)
- 18.56 O. Reynolds: An experimental investigation of the circumstances which determine whether the motion of water shall be direct or sinuous and the law of resistance in parallel channels, *Philos. Trans. R. Soc.* **174**, 935–982 (1883)
- 18.57 J.D. van der Waals: Coefficients of expansion and compression in corresponding states, *Amst. Ak. Vh.* **20**, 32–43 (1880)
- 18.58 R.C. Tolman: The principle of similitude, *Phys. Rev.* **3**, 244 (1914)
- 18.59 E. Buckingham: The interpretation of experiments on models, *J. Wash. Acad. Sci.* **93**, 336 (1914)
- 18.60 P. Forchheimer: *Hydraulik* (Teubner, Leipzig, Berlin 1914), in German
- 18.61 J.R. Chambers: *Cave of the Winds: The Remarkable History of the Langley Full-Scale Wind Tunnel* (NASA, Washington 2014)
- 18.62 E. Buckingham: Letter to Lord Rayleigh (John William Strutt) dated November 13, 1915, handwritten on official National Bureau of Standards stationery
- 18.63 E. Buckingham: *An Outline of a Theory of Thermodynamics* (Macmillan, New York 1900)
- 18.64 J.G. Kirkwood, O.R. Wulf, P.S. Epstein: *Richard Chace Tolman: Biographical Memoir* (National Academy of Sciences, Washington DC 1952)

- 18.65 D.P. Riabouchinsky: Methode des variables de dimension zero, et son application en aerodynamique, *L'Aerophile* **19**, 407–408 (1911), in French
- 18.66 J.J. Roche: *The mathematics of measurement: A critical history* (Anthione Press, London; Springer, New York 1998)
- 18.67 A. Vaschy: Sur les lois de similitude en physique, *Ann. Telegr.* **19**, 25–28 (1892) Transl. by A. C. Palmer in appendix to: *Dimensional Analysis and Intelligent Experimentation* (World Scientific Publishing Company, Hackensack, NJ and London 2008)
- 18.68 J. Bertrand: Sur l'homogénéité dans les formules de physique, *CR Acad. Sci. Paris* **86**, 916–920 (1878), in French
- 18.69 S.G. Sterrett: Similarity and Dimensional Analysis. In: *Philosophy of Technology and Engineering Sciences*, Vol. 9, ed. by A.W.M. Meijers (North Holland/Elsevier, Amsterdam 2009) pp. 799–823
- 18.70 E. Buckingham: Notes on the theory of dimensions, *Philos. Mag.* **42**, 696–719 (1921)
- 18.71 British Advisory Committee on Aeronautics: Abstract 134, Annual Report of the British Committee on Aeronautics for 1911–1912, p. 260 (1912)
- 18.72 D.W. Thompson: *On Growth and Form* (Cambridge Univ. Press, Cambridge 1917)

# Hypothetical

## 19. Hypothetical Models in Social Science

Alessandra Basso, Chiara Lisciandra, Caterina Marchionni

The chapter addresses the philosophical issues raised by the use of hypothetical modeling in the social sciences. Hypothetical modeling involves the construction and analysis of simple hypothetical systems to represent complex social phenomena for the purpose of understanding those social phenomena.

To highlight its main features hypothetical modeling is compared both to laboratory experimentation and to computer simulation. In analogy with laboratory experiments, hypothetical models can be conceived of as scientific representations that attempt to isolate, theoretically, the working of causal mechanisms or capacities from disturbing factors. However, unlike experiments, hypothetical models need to deal with the epistemic uncertainty due to the inevitable presence of unrealistic assumptions introduced for purposes of analytical tractability. Computer simulations have been claimed to be able to overcome some of the structures of analytical tractability. Still they differ from hypothetical models in how they derive conclusions and in the kind of understanding they provide.

The inevitable presence of unrealistic assumptions makes the legitimacy of the use of hypothetical modeling to learn about the world a particularly pressing problem in the social sciences. A review of the contemporary philosophical debate shows that there is still little agreement on what social scientific models are and what they are for. This suggests that there might not be a single answer to the question of what is the epistemic value of hypothetical models in the social sciences.

19.1	<b>Hypothetical Modeling as a Style of Reasoning</b> .....	413
19.2	<b>Models Versus Experiments: Representation, Isolation and Resemblance</b> .....	416
19.3	<b>Models and Simulations: Complexity, Tractability and Transparency</b> .....	420
19.4	<b>Epistemology of Models</b> .....	423
19.4.1	Instrumentalism and Predictive Ability .....	424
19.4.2	Isolation of Causal Mechanisms or Capacities .....	424
19.4.3	Learning About Possibilities .....	425
19.4.4	Inferential Aids .....	426
19.4.5	Models as Blueprints for the Design of Socio-Economic Mechanisms .....	427
19.4.6	Where Do We Go From Here? .....	428
19.5	<b>Conclusions</b> .....	428
19.A	<b>Appendix: J.H. von Thünen's Model of Agricultural Land Use in the Isolated State</b> .....	429
19.B	<b>Appendix: T. Schelling's Agent-Based Model of Segregation in Metropolitan Areas</b> .	430
	<b>References</b> .....	431

### 19.1 Hypothetical Modeling as a Style of Reasoning

In *Styles of scientific thinking in the European tradition*, the historian of science A.C. Crombie [19.1] lists six styles of thinking that characterize modern scientific thought:

1. The method of postulation
2. The use of experiments
3. The hypothetical construction of analogical models
4. The taxonomic method
5. The use of statistics and probability
6. Historical derivation.

*Ian Hacking* re-labels Crombie's classification as one of the *scientific styles of reasoning* and adds a sev-

enth: The laboratory style, which lies between methods (2) and (3) in that it relies on “built apparatus” to produce phenomena about which hypothetical models may be true or false [19.2]. Each style of reasoning introduces new kinds of objects, new evidence and new ways of being a candidate for truth. For our purposes the relevant features of a style of reasoning are its stability across disciplinary contexts and its autonomy, in other words, once a style of reasoning becomes established, it determines its own criteria of what counts as good reasoning [19.2].

Hypothetical modeling refers to that scientific strategy in which the known properties of an artifact are put to use in order to elucidate the unknown properties of a natural phenomenon [19.1, p. 1087]. *Mary Morgan* [19.3] deploys the concept of style of reasoning to characterize the practice of theoretical modeling in economics and traces the history of how modeling gradually became the prevalent style of reasoning in economics, achieving its present features around the second half of the last century. It is these features that Morgan’s work, and ours too, endeavors to describe. Here, however, we are concerned with hypothetical modeling as it takes place in the social sciences at large. By treating this method as one style, we seek to highlight its distinctive characteristics, which cut across disciplines.

Unlike in economics, in other social sciences such as sociology and political science, until a few decades ago the use of hypothetical models was limited to relatively narrow areas of inquiry. Well known, contentious attempts at introducing economic-style, rational-choice models in sociology and political science sparked accusations of economics imperialism [19.4]. More generally, formal modeling was often perceived in opposition to the qualitative leanings of many social scientists. Critics complained that mathematical models could not capture the complexity of social and economic phenomena, which are often hard to quantify and measure and do not obey the kinds of exceptionless laws that were believed to characterize the natural sciences. Disciplinary resistance to the method of hypothetical modeling, however, is not at odds with the stability characteristic of styles of reasoning: It is the deployment of a style in a new field and domain of inquiry that is contested, but its features, those that make it a style, might remain untouched. Moreover, the critical attitude toward the method of hypothetical modeling is now changing, at least in some social sciences. For instance, *Clarke and Primo* [19.5, p. 1] claim that, “[m]odels have come to be the dominant feature of modern political science and can be found in every corner of the field”. *Edling* [19.6, p. 197], writes that “since mathematical sociology was firmly established in the 1960s, it has grown tremendously”.

The transfer of models and modeling techniques across disciplinary boundaries is contributing to the establishment of shared modeling standards. Recent fields such as network theory and agent-based modeling are united by common modeling tools rather than by a set of principles or subject matter. These tools are then being modified and adapted *in house*, as it were, to satisfy the specific epistemic and nonepistemic needs of each field. For example, the use of network theory in sociology looks rather different from the use of network theory in economics and this is in part due to their being embedded in different disciplinary cultures [19.7, 8]. Thus, it is possible to talk of *field-specific modeling practices* to emphasize what is distinctive about a specific discipline and to talk of *style of reasoning* to underline the distinctiveness of model-based reasoning vis-à-vis other scientific styles of reasoning, such as the laboratory style. Which aspect is emphasized depends on the purposes of one’s enquiry.

Here we are interested in the commonalities: Treating hypothetical modeling as a style of reasoning encourages us to look at its characteristic features vis-à-vis other styles of reasoning employed in social science. Philosophers of science sometimes talk of model-based social science, a label that captures the same kind of scientific activity. Here is how *Peter Godfrey-Smith* characterizes it [19.9, p. 726]:

“What is most distinctive of model-based science is a strategy of indirect representation of the world [...] The modeler’s strategy is to gain understanding of a complex real-world system via an understanding of a simpler, hypothetical system that resembles it in relevant respects.”

The terms *hypothetical modeling* and *model-based science* both refer to the scientific activity of understanding phenomena by building hypothetical systems, which at once are much simpler than the phenomenon under investigation and hopefully resemble it in some respect. The modeler studies these simpler, hypothetical systems in order to gain insights into the more complex phenomena they represent. These hypothetical systems can be of different kinds: They can be concrete objects such as the scale models of engineers, or they can be the set of mathematical equations very familiar to both physicists and economists. In this chapter our main focus is on models that are abstract in the sense that they do not exist as physical objects to be manipulated by the modeler. They are theoretical rather than empirical models. Empirical models are built for testing and measuring relationships between variables and are based on empirical data; they do not describe hypothetical systems. In this sense they are better thought of as belonging to the statistical style of reasoning.

The attribute *theoretical* should not be taken to suggest that theoretical models are always instantiations of general theories. Theoretical principles might be only one of the several ingredients that go into the construction of theoretical models [19.10]. Ideally, theoretical (or hypothetical) modeling and empirical modeling are tightly connected. The results of theoretical models are translated into empirical models and thereby subject to testing. In many cases, however, the evaluation of theoretical models proceeds without the results being directly confronted with data. This is how the historian and methodologist of economics *Roger Backhouse* explains the relationship between theoretical and empirical models [19.11, p. 138]:

“Empirical work starts with a set of economic relationships, formulated in such a way that they can be confronted with data using formal, statistical techniques. These relationships may be the theoretical results [...], but typically they will be different. The reason for this is the requirement that they can be confronted with data: They must refer to variables on which statistical data exist, or for which proxies can be found; functional forms must be precisely specified and amenable to statistical implementation.”

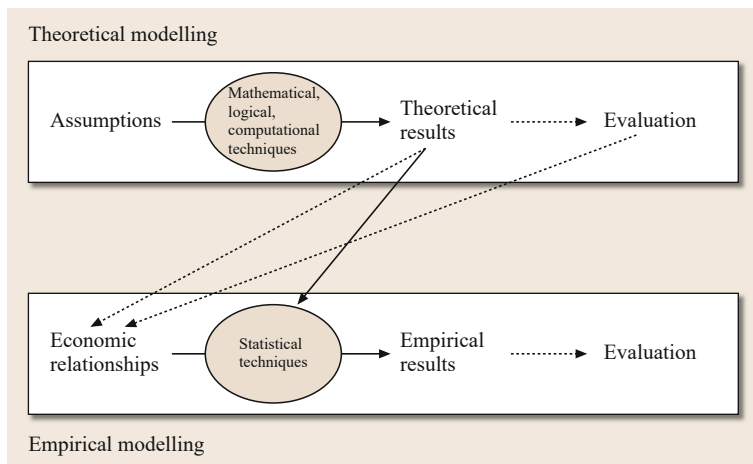
Backhouse’s account is about economics, but it can be generalized to other social scientific contexts in which theoretical results cannot be directly confronted with empirical data. In such cases, theoretical and empirical modeling might be only loosely connected with one another, for example, by way of the theoretical results informing empirical modeling and vice versa, as depicted in Fig. 19.1.

But which theoretical results are taken seriously enough to inform further empirical investigations? The

evaluation of hypothetical models often relies on other criteria such as credibility, insightfulness, explanatoriness or other modeling desiderata. This is one of the aspects that make hypothetical modeling partly autonomous from other styles of reasoning: The evaluation of hypothetical models and their results is based on criteria largely internal to the style, criteria that have developed together with the stabilization of the style.

In contemporary social science the diffusion of hypothetical modeling to tackle social scientific questions is taking place in parallel with an increasing reliance on computer simulations as well as on laboratory and field experimentation. Recall that Hacking takes the laboratory style to lie between the use of experiments and hypothetical models. The laboratory style differs from experimentation simpliciter in that it creates artificial environments about which the hypothetical models can be true or false. In the social sciences where laboratory experiments were believed to be virtually impossible, modeling was considered to be an attempt—to some fruitful, to others idle—to achieve theoretically what it is impossible to implement in the laboratory; even now, when both laboratory and field experimentation have become well-established practices in many social sciences, many questions of interest still cannot be studied experimentally. It is perhaps not surprising that models and experiments have been compared in order to understand their characteristic features as well as their common characteristics.

The discussion in Sect. 19.2 addresses the question of whether models are relevantly similar to experiments and, if not, where the deep differences lie. In Sect. 19.3, we examine the issue of whether computer simulations belong to the style of hypothetical modeling. In general terms, the question is whether or not computer simulation and analytical modeling are different ways of studying hypothetical systems. These comparisons



**Fig. 19.1** The relationship between theoretical and empirical modeling (after [19.11, p. 136])

allow us to highlight some of the main features of hypothetical modeling in the social sciences. Finally, we discuss the legitimacy of hypothetical modeling as a way of learning about social scientific phenomena. The debate reconstructed in Sect. 19.4 attempts to understand how hypothetical modeling deals with the specificities of the social sciences and to examine the conditions of its legitimacy. As will become clear, there is little agreement on the nature and function of theoretical models in social science. Perhaps this is not so much a sign of slow progress in understanding models as it is an indication that a substantive as well as a general account of the nature and function of models is not possible [19.12].

## 19.2 Models Versus Experiments: Representation, Isolation and Resemblance

The alleged impossibility of laboratory experimentation has for long been considered one of the features that sets the social sciences apart from the natural sciences. It was widely believed, among both scientists and philosophers, that experiments were the exclusive purview of the natural sciences: Many if not all questions of interest in the social sciences were thought not to be amenable to experimental investigation, owing to the difficulty of designing experimental settings capable of reproducing and confining the phenomena of interest. The broad scale of many social phenomena and the inevitable presence of disturbing factors of very different kinds (e.g., history, cultural background, value judgments, etc.) were seen as insurmountable obstacles to controlled experimentation. Obvious ethical issues also contributed to limit the range of feasible scientific experiments and continue to do so. Since the second half of the last century, however, the use of experimentation in social science has grown remarkably, both in the laboratory and in the field, thanks to technological and methodological developments, which allow control of many of the disturbing factors that were previously considered impossible to control for ([19.13], see also [19.14] for economics and [19.15] for political science).

The method of hypothetical modeling has been seen as an alternative to experimentation when experimentation is difficult or unfeasible [19.16, 17]. Both models and experiments are interpreted as devices for surrogate reasoning, which are examined to draw inferences about the target phenomena they aim to represent [19.18–20]. In the philosophical literature models and experiments have been compared in terms of the functions they play in scientific inquiry and in terms of their use for drawing inferences about the world. The

Most of the philosophical literature that we consider throughout the chapter is about economic modeling. This is because, in the social sciences, economic modeling is where hypothetical modeling has been most prevalent and hence has received the most philosophical attention. However, many of the insights from this literature apply across the social sciences. Not only because rational-choice models are widely employed in social sciences other than economics such as sociology and political science, but also because the indirect representation of phenomena through idealized models crosses disciplinary boundaries. If and when a given issue is peculiar to economics and cannot be generalized, it will be pointed out.

concepts of isolation, representation and manipulation are at the center of this comparison, which deals with functional, methodological and epistemic aspects of the two styles of reasoning.

Models and experiments can be seen as playing a similar function, i.e., that of isolating the phenomenon of interest from the interference of disturbing factors [19.16–19, 21, 22]. This analogy concerns the ideal model and the ideal experiment – or, as Cartwright calls it, the “Galilean experiment” – in which only the factor of interest is allowed to vary, leaving everything else constant. Ideally, experiments proceed by removing or controlling all potentially disturbing factors so as to allow scientists to create the conditions under which causal relations can be observed in isolation. An influential account of models holds that, like experiments, models aim at studying specific aspects of the phenomenon of interest (such as causal relations, capacities or mechanisms) in isolation from the interference of disturbing factors [19.3, 16, 17, 21, 23, 24]. On this account, the isolative function of the ideal model is akin to that of the ideal experiment. To emphasize this similarity, it has been suggested that models are *theoretical experiments* aimed at creating, theoretically, the kind of controlled conditions typical of the laboratory [19.18].

As an illustrative case, consider von Thünen’s well-known model of localization, which is described in the Appendix (Sect. 19.A). The model could be interpreted as the result of a process of isolation that zooms in on the relationship between spatial distance and land use. This is done by means of assumptions that neutralize the effect of other factors by assuming them to be constant, absent or negligible [19.23]. In the model the pattern of agricultural production around the city

depends exclusively on transportation costs, because all other factors that can influence land use, such as the presence of rivers or mountains, the different fertility of the soil and the presence of transport routes, are assumed away. Most of the assumptions listed in Sect. 19.A can be seen as having this function.

The investigations of this analogy by *Morgan* [19.19, 22] and *Mäki* [19.18] emphasize that the way in which models and experiments fulfill their isolating function is different. Experimenters control for disturbing factors by designing the experimental set-up so that disturbing factors are prevented from interfering with the mechanism of interest; this can be done both by physical interventions in the experimental environment and by choices regarding the procedure [19.22]. A social scientific experiment, for instance, might take place in a laboratory in which the subjects interact with the experimenter only via computer terminals, so as to minimize the possibility that the experimenter's expectations influence the subjects' behavior. Alternatively, in order to control for the effect on subjects' choices of the language employed to describe a number of alternatives, the experimenter might design the procedure so that the subjects, instead of all being presented with exactly the same description of the alternatives, are randomly assigned different descriptions, thereby ensuring that language will not have a systematic effect on aggregate choice behavior [19.15].

Hypothetical models try to achieve the same kind of control by means of theoretical assumptions, as in the case of von Thünen's model. The isolation achieved by assuming away all disturbing factors is only a theoretical exploration that does not provide a definitive answer to the question of what would happen in the real world in such a situation. It is open to debate what this difference amounts to and what its consequences are, in particular when it comes to using this style of reasoning for drawing inferences about the world. For *Morgan*, as we will see shortly, this dissimilarity is grounded in the different materials of which models and experiments are made, and this, in turn, has major epistemic consequences. In *Mäki's* account, however, the consequences of this dissimilarity are limited to the degree and strength of isolation that the two styles are able to provide and do not result in major differences concerning their use for making inferences about the world.

For *Mäki*, models are able to display a higher degree of control than experiments can do [19.18]. This is because experiments can isolate only to the extent to which it is practically feasible, and hence some interferences are left uncontrolled for or only weakly controlled for. Models, instead, can provide tighter isolation, because they are not subjected to these practical constraints and can simply assume all interferences

away. One implication of this is that the use of experiments to test models is bound to be imprecise.

In *Mäki's* account, however, the difference in the degree of control between models and experiments does not compromise the identification of further analogies. In particular, *Mäki* claims that "many theoretical models are (*thought*) experiments, and many ordinary experiments are (*material*) models", because both fall under the general concept of *scientific representation* [19.18, p. 303]. In *Mäki's* interpretation, something qualifies as a representation if (a) it is used as a representative of its target and (b) it resembles the target in some respects and to certain degrees. Hypothetical models, as we characterize them, are substitute systems that are examined in order to gather information about what they represent. Similarly, experiments are pieces of the world created in the laboratory that can be seen as material models, which again are not examined for their own sake, but rather for their value in gathering information about the world outside the laboratory. Models and experiments qua substitute representations face the question of whether it is legitimate to draw inferences from the representational tool to the real world. This can be thought of as an instance of the general problem of extrapolation, which concerns the generalization of the results obtained in a model or in an experimental situation to the world. For *Mäki*, in both cases the key to these problems rests on whether the world created in the laboratory or in the model resembles reality in the relevant respects and to a sufficient degree. In our example from von Thünen's work, the problem amounts to establishing whether the model is relevantly similar to reality so as to justify the claim that the distance from the market actually affects the distribution of economic activities as described in the model. In a laboratory experiment on individual choice behavior, the question arises as to whether the kind of task subjects are asked to perform and the artificial environment of their interaction are relevantly similar to situations that occur in the wild so as to warrant inferences from the experiment to the world. In *Mäki's* account, since models and experiments raise similar issues of resemblance, the ability to draw inferences about the world does not hinge on their respective features. Although models can sometimes be made closer to reality, the answer to the problem of extrapolation ultimately depends on what it means for a model to resemble a real situation, and on the ability to identify the relevant aspects and the sufficient degrees of resemblance [19.25]. Therefore, a margin of ambiguity in the notion of *sufficient relevant resemblance* remains.

*Morgan* takes a different stance on the problem of extrapolation [19.19, 22]. She emphasizes that models and experiments are similar in the way in which they are



used and manipulated, but argues that the distinct features that characterize these instruments provide them with different epistemic powers for investigating the world. Models and experiments are manipulated by introducing controlled variation of some of their aspects in order to check whether the results are affected by this change. These manipulations allow scientists to investigate the consequences of variations in the initial conditions on the result and/or the conditions under which a particular result of interest can be obtained. Von Thünen, for instance, after having illustrated his model as described in Sect. 19.A, continued his investigation by introducing the presence of a river and a smaller neighboring town, thereby altering two of the initial assumptions. Changing the model in this way enables investigation of how the newly introduced assumptions affect the pattern of land use in comparison to the initial scenario: The river facilitates the transportation of goods to the market and thus distorts the concentric pattern, whereas the presence of a smaller town generates its own concentric regions of land use. In *Morgan's* terms this can be described as a case of “experiment on models” [19.22]. The manipulation of models can be interpreted as a kind of experiment in which the scientists “interrogate” the model by modifying some of the initial assumptions. These questions can be motivated by theoretical issues, by the aim to explain or predict real-world situations or by the policy agenda the model is meant to guide. For instance, von Thünen’s manipulations of the basic model can be seen as being prompted by questions such as “what happens to the pattern of concentric rings if there is a river and/or a neighboring town?”

Yet when it comes to connecting the answers obtained in this process to the real world, models and experiments are substantially different for Morgan. She argues that experiments have greater epistemic power than models, because the experimental inferential leap is smaller than in the case of models. This dissimilarity is based on the different materials of which each is made: Experiments are concrete investigations, which deal directly with the world they are meant to study, whereas models are abstract and idealized. Economic experiments, for instance, deal directly with real people’s behavior, however constrained the behavior is by the experimental design. By contrast, models are abstract entities, which are made of different stuff than the reality they represent. For *Morgan*, the “materiality” of experiments can make inferences from experimental results to the world both easy and strong, because they are grounded on the material uniformity between the system on which the manipulation is conducted and the world about which the inference is made [19.19]. More precisely, Morgan maintains that, insofar as experiments share the same ontology with their target (which

is not obvious), we are more justified in claiming to learn something about the world from experiments than from models. Inference from the model to the world is much more difficult because the materials are not the same as the world’s.

Moreover, because of their materiality, experiments can create new phenomena that might be different or even contrary to theoretical expectations. When an unexpected experimental result is sufficiently stable across replications and manipulations of the experimental design, it qualifies as a *new* phenomenon, which requires theoretical explanation. For instance, behavioral regularities robustly observed in several economic experiments, such as co-operation in prisoner dilemma games, can be thought of as constituting new phenomena with respect to the expectations of rational choice theory. Such experimental phenomena have now become the target of sustained theoretical efforts to account for them. *Van Fraassen* makes a similar point, arguing that scientific instruments can be viewed not only as windows that allow us to see what happens in the world, but also as machines that create new genuine phenomena that would not occur in the wild and that theory needs to explain [19.26].

In *Morgan's* account, the creation of new experimental phenomena is only possible in the laboratory and does not belong to the style of hypothetical modeling [19.19]. Only real flesh and blood experimental subjects have the freedom to behave in other ways than expected. Experiments must allow a certain degree of freedom because, if the subjects’ behavior was fully determined, then the experiment would have no genuine potential to confirm or refute a theory. Therefore, according to Morgan, experiments have the potential to surprise and confound theory: They can illuminate unexpected or hidden consequences of a theory, but their results can also be in conflict with theoretical expectations. On the other hand, models cannot confound because the behavior of agents is pre-determined by the modeler’s assumptions [19.21]. In other words, the agents in the model lack the “potential for independent action”, which is what confers greater epistemic power to experiments [19.19].

Morgan’s ideas, however plausible, can be challenged. *Parker* rightly observes that it is not always the case that inferences are better justified when the representative tool and the target are made of uniform materials [19.27]. What is crucial is not the material, but the presence of relevant similarities, which can be material but also formal (in this respect, Parker’s point is similar to Mäki’s). The justification of inferences about the world depends on having good reasons to think that the relevant similarities are in place. Having experiment and target made of the same material does not guaran-

tee that all relevant similarities are in place, because it is possible that “same-stuff representations” fail to be relevantly similar to their target systems. Nevertheless, Parker agrees with Morgan that the material uniformity between experiments and the world can provide some epistemic advantage: Experimental and target systems that are made of the same stuff will often be similar in many relevant respects. In other words, being made of the same material does not guarantee that relevant similarities are in place, but it does make it more likely. As a consequence, inferences from experiments are more likely to be reliable than inferences from models.

A further challenge that models have to face arises from the fact that not all assumptions in a model can be thought to function as isolations. Assumptions introduced to make it possible, or easier, to handle the model analytically might impose constraints that are too tight to allow relevant similarities between the models and the represented aspects of the world. This is a point Cartwright makes specifically with regards to economic models: Although many economic models aim at mimicking Galilean experiments, they also include a number of assumptions that do not play the role of isolation, but are rather introduced for the purpose of mathematical tractability [19.21, 28]. If the models’ conclusions are due to assumptions that are completely unrelated to the world, then it is not clear what the model can tell us about the world.

*Derivational robustness analysis* has been proposed as a remedy for the problem of over-constraining assumptions [19.29, 30]. For Kuorikoski et al. derivational robustness analysis refers to the collective practice of building similar models of the same phenomenon that differ only in a few assumptions. The analysis of these groups of similar models can help to identify which assumptions are necessary for deriving a certain result: Results that are robust across a number of models are dependent on the shared, rather than on the differing, assumptions. Now if the over-constraining assumptions (or more generally the assumptions known to be unrealistic) are found to be unnecessary for deriving the result of interest, then there are reasons to think that this result is based primarily on the shared assumptions,

which are hoped to be realistic. Hence, according to Kuorikoski et al. even though robustness analysis is not a procedure of empirical confirmation, it can increase modelers’ confidence about their inferences from hypothetical models. Note that although Kuorikoski et al. are mainly concerned with analytical models, robustness analysis has also been claimed to be a crucial strategy in correcting for various sources of error that might affect the results obtained by computer simulations, as we will see in Sect. 19.3.

*Odenbaugh* and *Alexandrova* raise the valid objection that, although in principle robustness analysis might work, in practice it does not provide a defense of hypothetical models with over-constraining assumptions [19.31]. If, for example, in economic modeling, some of the core rational-choice axioms are never modified to check their effects on the modeling results and if these axioms are in fact wildly unrealistic, then robustness analysis turns out to be of limited use. Hence, although in principle hypothetical modeling might be aimed at isolating a mechanism of interest, there remains the problem that many assumptions do not have this isolating function. This situation can jeopardize the resemblance between the models and the represented target, which would warrant the inferences from the model to the world. We will return to these issues in Sect. 19.4.

In conclusion, both models and experiments can be considered as representations that are examined in order to draw inferences about what is represented. They differ in the kinds of representation involved: Models are abstract indirect representations, whereas experiments are concrete direct representations made of the same material as the target. It has been claimed that this has implications both for the kind and the degree of control that these devices are able to provide and for the way in which conclusions are drawn from them. On the one hand, theoretical models enable the investigation of phenomena that are difficult or impossible to reproduce in the laboratory; yet on the other hand, only experiments seem to have the genuine potential to bring to light new phenomena that require theoretical explanation (Table 19.1).

**Table 19.1** Models, experiments, simulations: A comparative perspective (after [19.22, p. 49])

	<b>Ideal model</b>	<b>Ideal laboratory experiment</b>	<b>Ideal simulation</b>
<i>Kind of representation</i>	Indirect: Different material	Direct: Same material	Indirect: Different material
<i>Isolation and control</i>	Assumed theoretical isolation	Experimental material isolation	Assumed theoretical isolation
<i>Advantages</i>	Theoretical exploration in which experiments are difficult or unfeasible	Discovery of phenomena for science to explain	Representation of complex and/or dynamic problems and other problems that are not solvable analytically
<i>Challenges</i>	Tractability	Material and ethical constraints	Transparency

### 19.3 Models and Simulations: Complexity, Tractability and Transparency

Together with theoretical models and experimental analysis, computer simulations are key instruments in the toolkit of the social scientist. Urban evolution, flows of goods, crowd effects, the stock exchange, virtual financial markets and regulatory policies are just a few examples of what can be analyzed by means of computer simulation. More concrete examples of some of the newest applications include a simulated model of the entire European Union economy, which describes the interaction of a massive number of financial components involved in several markets simultaneously [19.32]. In Stockholm, a data set has been created for the years 1990–2003 to map the entire metropolitan area and simulate segregation effects [19.33].

Some of the characteristics of computer simulation and the ways in which they are employed invite a comparison of these instruments with the style of analytical modeling. The use of computer simulation, in fact, can be thought of as either complementing or replacing analytical models. In the literature some authors have emphasized the elements of continuity between the two methods; others have highlighted the differences. For instance, according to *Guala* [19.20] and *Morgan* [19.22], models and simulations are akin to each other in the way they are used to learn about the world and for the functions they fulfill. Others have argued that computer simulations open up novel methodological questions that did not arise in dealing with analytical models [19.34]. Below, we will explore both the similarities and the differences between these methods, with special focus on the features of computer simulation that have been debated in relation to their adoption in economics and the social sciences.

To see how models and simulations connect, consider how computer simulation originally entered the field of the social sciences. One pioneer in the study of social phenomena with the aid of the computer has been the political scientist Robert Axelrod. In 1980, *Axelrod* launched a competition between experts in game theory from different fields [19.35, 36]. The challenge was to come up with a strategy for an iterated prisoner's dilemma game to be played in a computer tournament. Axelrod paired strategies – fifteen in all – and had the participants play for two hundred rounds in an all-play-all tournament. At the end of the tournament, the winning strategy turned out to be one of the simplest and most ancient strategies of human co-operation, *tit for tat*. The strategy is to co-operate in the first round of the game and then replicate the opponent's moves, i. e., to co-operate in case of co-operation and defect as soon as the other player defects. The strategy is successful insofar as it reaps the benefits of co-operation and does

not lose too much from retaliation. After the results of the tournament's first round were circulated, a second round was held. Once again, *tit for tat* won the competition out of sixty-two rival strategies.

Axelrod's tournament is one of the first examples of the combination of game theory with the computer to study co-operation. There are several reasons why the encounter between game theory and the computer was so fruitful. First and foremost, game theory has traditionally focused on strategic rationality, i. e., on the epistemic criteria for the solution of interaction problems. The discipline is usually silent on dynamic aspects, i. e., on what happens at the population level when different strategies encounter one another repeatedly. Evolutionary game theory focuses on the latter aspect of the problem to expand and complement the domain of inquiry of traditional game theory. Computer simulations are particularly useful for this purpose, as they enable scientists to focus on frequency aspects of strategic interactions rather than on the quality of the strategies per se.

This point can be illustrated by an example. Imagine for a moment that each of the procedures sent by each game theorist to Axelrod represents the way in which the theorist would have played the game in real life. The number of possible combinations quickly becomes unmanageable (in the first tournament, which was repeated 5 times, there was a total of 240 000 choices). By means of computer simulation, it is possible to study which strategies survive, which become extinct and which co-exist. Through computer simulation, agents can be represented more realistically than before, for example, as individuals with bounded rationality and with learning and memory constraints.

Precisely because simulations study how social phenomena emerge and evolve through the interactions of single individuals and their environment, it has been claimed that they represent an invaluable tool in social science [19.37]. This is because the way the simulations analyze the occurrences of social phenomena reproduces the dynamics in which such phenomena occur in the social world. They provide bottom-up, mechanistic explanations [19.38, 39]. Through computer simulation, the role of individual, structural and institutional variables can be represented in a particularly realistic fashion, which has been claimed to help capture the complexity of their interdependence. Note, however, that the enthusiasm with which simulations have been welcome in some fields, such as analytical sociology, has not been unanimously shared. In fact, most notably in economics, simulations have been viewed with suspicion and their adoption not always encouraged. In order

to understand the diverging attitudes of economists and other social scientists, in what follows we will address the following questions:

1. What are the features that characterize computer simulation?
2. Do such features relate to a cluster of properties that distinguish computer simulation from other styles of reasoning such as analytical models?
3. Is economists' preference for analytical modeling over computer simulation justified on epistemic grounds?

Firstly, consider that even if we talk about computer simulations as distinct from analytical models, computer simulations are ultimately based on models. The way in which the two methods differ is that computer simulations obtain their solutions by means of a program that runs on a computer, whereas the solutions of analytical models can be obtained without the aid of a computer. This is simply because simulated models take into account a higher number of variables and consider nonlinear relationships, which are easier to explore with the computer. Broadly speaking, computer simulation refers to the entire process of formulating a *model*, transforming it into an algorithm that runs on a computer, calculating the output and analyzing the results [19.34, 40–42]. Moreover, the contrast between analytical models and computer simulations should not convey the idea that computers do not proceed analytically. Obviously they do (at least if we narrow our focus to models in the social sciences that do not require numerical analysis); the difference is that when computer simulations analyze complex systems, they usually proceed by averaging over a sufficiently high sample of cases rather than in the manner of mathematical proofs (more on this below).

Secondly, note that *computer simulation* is a coarse-grained label, which generalizes different ways in which simulations can be developed. Different taxonomies have been proposed in the literature [19.34, 43]. A preliminary, common distinction is between agent-based models (ABM) and equation-based models. The former proceed by implementing local rules, such as a decision rule in a sociological model; the latter, by translating equation-based models, such as partial or ordinary differential equations in physics, into a computer program. The boundaries between disciplines, however, are not strict. Agent-based models are frequently used in areas other than sociology, including fields that were previously dominated by analytical approaches, such as population ecology and theoretical physics.

A less common but still well-known interpretation defines computer simulation as a subset of a more gen-

eral class of simulations that deploy computers for their ends [19.44]. In this view, a physical model of a target system – e.g., a scale model such as those used in structural engineering – counts as a simulation that adopts a specific means of representation, i.e., a physical model rather than a computer model. An extensive body of literature examines the similarities between computer simulations and experiments and highlights the fact that simulations are closer to the style of experimental analysis than to the style of analytical modeling. For instance, according to *Morrison*, in their relations to models, simulations are akin to experiments [19.45]. According to *Parker*, however, simulations lie between models and experiments in that they display features of both experimentation and modeling [19.27] (for a review of the literature on the experimental interpretation of computer simulations, see e.g., [19.34]).

Taxonomical differences aside, in the most basic sense defined above, computer simulations are simply a tool in the hands of the scientists. They help to achieve the model's results in a manner similar to the way in which a calculator helps perform difficult mathematical operations. Moreover, simulations enable the modeler to represent the target system with a greater level of detail than is usually found in analytical models and to spell out in a particularly precise way the assumptions behind the working hypothesis [19.34, 40, 41].

To illustrate, let us compare in more detail the differences between two mathematical approaches to a study of the same phenomenon. The Lotka–Volterra model is a model in population ecology that also has had applications in the social sciences for the study of organization–environment relations [19.46]. The analytical version of the Lotka–Volterra model is a highly abstract representation of the ecological (organizational) system under study, which omits features such as the environment in which the species live (the market), a realistic level of satiation (competition), lifetime (supply) and so on. When the same problem is addressed by means of an agent-based computer simulation model, a particularly detailed representation of the system of interest is provided. Hence, it is claimed, not only can computer simulation help to avoid common errors in intuition, it might also reveal a system's relevant aspects that had been underestimated or disregarded. Furthermore, computer simulations have heuristic functions: They trigger our intuitions and can be helpful in exploring new hypotheses; not least, they enable us to visualize the results of a problem in particularly efficacious ways. A more detailed and concrete example of how computer simulation proceeds is given in Sect. 19.B.

Given the features discussed above, it would be natural to expect computer simulation to be called upon to

complement analytical models. In 1982 *Richard Feynman* was already justifying the adoption of simulations in physics on the following grounds [19.47, p. 468]:

“The present theory of physics [. . .] allows space to go down into infinitesimal distances, wavelengths to get infinitely great, terms to be summed in infinite order, and so forth. With computers we might change the idea that space is continuous to the idea that space perhaps is a simple lattice and everything is discrete (so that we can put it into a finite number of digits) and that time jumps discontinuously.”

Physical theories trade formal rigor for unrealistic assumptions, such as continuous space and infinite wavelengths. As Feynman suggests, computer simulations can help physicists reduce the constraints of mathematical tractability in favor of descriptive accuracy. Although similar considerations also apply to the social sciences at large, in economics the endorsement of computer simulation has been slower than in physics and other areas of science – with a few exceptions, such as the one we saw above from evolutionary game theory.

Why then is there such an uneven reception of simulation in the social sciences? Why have economists, who have the most established tradition of modeling, been reluctant to embrace simulation? *Lethinen* and *Kuorikoski* investigate the reason for economists’ preference for analytical models [19.42]. The authors claim that this tendency might even slow down progress in the subject and lead economists to dismiss results that would be reasonable to accept. According to *Lehtinen* and *Kuorikoski*, it is because simulations do not provide the kind of *understanding* that is perceived as legitimate in the economic community that this methodology is more often considered as a secondary option at best. Two key assets of economic theory, namely rationality and equilibrium analysis, have a marginal role in agent-based modeling, and these are aspects that economists do not seem willing to dispense with. This also partly explains why in other social sciences, where there is no strong commitment to a unified theoretical corpus, agent-based models are increasingly used. As we have seen, agent-based models allow a greater degree of flexibility in the behavioral rules ascribed to the agents. For many social scientists, this represents an asset rather than a liability. In economics, however, this flexibility is often considered a problem in that the choice of behavioral assumptions is ad hoc rather than guided and constrained by a unified theory (see [19.7] for a discussion of this issue).

In the paper *Robust simulations*, *Ryan Muldoon* addresses another source of concern related to the adoption of computer simulation in science, mainly in-

volving the problem of verifying results [19.48]. To adopt a term used in the literature, simulations are said not to be *transparent*. While it is possible to check each step in the derivation of an analytical model, the same does not apply to simulation. Errors might be concealed within the particular machine used to run the simulation or within the particular programming language or within the algorithm itself. According to *Muldoon*, the best strategy for increasing confidence in the results is to show that simulations provide robust results, i. e., results that are invariant to changes in the hardware, the programming language and the algorithm. Depending on the degree of confirmation a scientist needs to achieve, a robustness test investigates the source of possible mistakes similar to the way in which experimental scientists test their experimental results.

Probably for a combination of the reasons given above, recourse to computer simulation in economics has been legitimized mainly when models become too complex to be analytically solvable or when the volume of data collected is such that only high-powered computers can process them. But what precisely does it mean for a problem to be intractable, and how do computer simulations deal with that? An example that illustrates this issue with particular clarity is *Schelling’s* model of racial segregation (see Sect. 19.B). *Schelling’s* model explains the emergence of ethnic clusters in different metropolitan areas as a consequence of the preference of individuals for having a few neighbors of their own ethnic group. Agents have different information about their neighborhood, and, at any point in time, they can decide to move to another neighborhood that better suits their preferences. Agents move randomly in space, and when they move, they tend to generate further movements of those individuals whose neighborhood has now changed. The chain of possible effects triggered by each agent’s decision makes segregation processes particularly difficult to formalize analytically.

Note that the issue of tractability does not concern only the probabilistic nature of the problem. Analytical methods can in fact be used to calculate the development of a probabilistic system without the need to resort to simulation. In the case of *Schelling’s* model it is because agents have different information and because neighborhoods overlap with one another that analytical treatments are usually excluded. One way to proceed analytically would be, for instance, to assume that the entire city is a unique neighborhood. At that point, all agents would share the same information and the problem of overlapping neighborhoods would be solved. However, no one would move anywhere simply because there would be no neighborhood to go to. In this situation computer simulations can remedy the lack of analytical solutions by providing an approximation of

the process under investigation. At each step of the simulation the state of the system probabilistically depends on its configuration in the previous round. When the results aggregate we can observe whether an underlying dynamic emerges. Furthermore, every time we rewind the tape and run the simulation again, we can observe whether the macrophenomenon is stable despite the contingencies that characterize each particular stack of simulations. Finally, since the simulation environment is significantly flexible, we can consider a variety of factors and their impact, such as agents with different utility functions, or cities with different network structures [19.49].

Notice, however, that there is no reason why a way could not eventually be found to develop analytical solutions for Schelling's model. In fact, in evolutionary game theory, progress has been made in developing analytical solutions with Schelling's model, which rely on stochastic processes [19.50]. More generally, it is often the case that a certain problem does not have an analytical solution until a scientist finds one. When we look at the conditions that make a problem mathematically tractable or intractable we find that there are no neat boundaries between the two. Something that has been mathematically intractable up until today might become tractable tomorrow, thanks to progress in the discipline. But there are no neat criteria that define the tractability/intractability of a problem. Hence, the economists' claim that simulations should be limited to intractable problems, and that scientists should not leap to simulations when an alternative is possible, appears unjustified

after all. According to a less narrow perspective, the adoption of computer simulation might be welcomed even in cases in which an analytical solution is possible, but which is particularly demanding to find, time consuming and expensive with respect to research costs.

To conclude, this section opened with a number of questions on the nature of computer simulations and their relation to analytical models. As we have seen, the answer to the question of whether computer simulations constitute a different style of reasoning from that of analytical models depends on the level of analysis we consider. The differences between analytical and simulated models appear clearer when we look more closely at the two methods: Computer simulations are particularly apt to deal with complex systems, even though they do so at the cost of dispensing with analytical solutions. At a more general level, however, the two practices can be seen to be similar: They both concern the formulation of models and their manipulation for the achievement of results (Table 19.1). Computer simulation should not be taken as a remedy for the problems that affect analytical models, such as whether and under what conditions we are justified in transferring their results to real-world phenomena [19.51]. In these respects, computer simulations deal with issues similar to those dealt with using analytical models, if not with more complicated ones. This, however, does not constitute a reason to avoid their adoption. Rather it indicates that scientists' efforts are needed to meet the challenges that this new methodological tool offers to actual scientific practice.

## 19.4 Epistemology of Models

Scientific models, both analytical and simulation-based, are used for a variety of purposes: Some are used for heuristic or pedagogical reasons; others for prediction or explanation of socio-economic phenomena; still others are built and used with some practical application in mind. The use of models for these purposes, however, has been considered far from unproblematic. Clearly, the source of the problem depends to some extent on the specific purpose to which models are put. For example, von Thünen's model can be thought of as yielding the prediction that, other things being equal, the production of goods that are cheap to transport will take place farther away from the central market. If this prediction is not fulfilled, it might be because the real situation we are considering is not as described in the model. This is something to be expected, if we consider that von Thünen's model abstracts away from many of the factors that in reality have an impact on the localization of

economic activities. The same model can also be seen as providing an explanation of the formation of a pattern of concentric rings around the market, if such is observed, by pointing to the transportation cost of the goods as the cause of the phenomenon. However, even if such a pattern is observed, the question remains as to whether the cause of its emergence is in fact the one theorized in von Thünen's model. Similar issues arise also with respect to Schelling's model, which assumes that the only factor affecting an agent's decision about where to live is the color of other people in a stylized neighborhood. Obviously, in the real world many other factors, such as housing prices and commuting distance, influence individuals' location choices.

Skeptical positions regarding the possibility of acquiring knowledge about the social world by means of hypothetical models are often based on the observation that models typically contain a number of false assump-

tions, and it is unclear how accurate predictions or true explanations can be derived from models that are partly false. No doubt, false assumptions are also employed in the natural sciences, and the role of scientific idealizations is likewise central to the philosophical debate about modeling in the natural sciences. The presence of false assumptions, however, has been regarded as being a particularly acute problem in the social sciences. It has been argued that whereas in the physical sciences it is possible to test idealized models by recreating the same conditions in the laboratory, in the social sciences this can rarely be done. Moreover, unlike the natural sciences, the social sciences typically lack general theoretical principles (or laws) that indicate how deviations from the model's assumptions will affect the result in the real world. For instance, suppose that our model of a falling object assumes that there is no air resistance. The effect of air resistance on the acceleration of a feather falling on the floor can be calculated with the appropriate formula of classical mechanics. In social science, there are only few, if any, general principles of this kind [19.21].

Nevertheless, hypothetical modeling is regarded as a legitimate style of reasoning in many quarters of the social sciences. Social scientists do distinguish between good and bad models in ways that do not necessarily have anything to do with an attempt at direct empirical testing. Both von Thünen's and Schelling's models enjoy a high standing in the social sciences, and their results are generally believed to be relevant, even if not fully or straightforwardly applicable to the real world. How then are such judgments made? And are they legitimate? The questions related to whether and when it is legitimate to use models for epistemic and practical purposes has been at the center of the philosophical debate. However, there is very little agreement among philosophers as to how to address these questions. Below, we reconstruct the discussion of how hypothetical modeling is and should be used in economics and the other social sciences by organizing the different perspectives around the functions of models that are taken as primary:

1. To make qualitative predictions
2. To isolate mechanisms or capacities
3. To learn about possibilities
4. To help with inferences
5. To design socio-economic mechanisms.

#### 19.4.1 Instrumentalism and Predictive Ability

According to an instrumentalist interpretation, if a model yields accurate predictions, then the truth or

falsity of its assumptions does not matter. This position is well exemplified by *Friedman's* position advocated in his famous essay *The methodology of positive economics* [19.52, p. 14]:

"In so far as a theory can be said to have *assumptions* at all, and in so far as their *realism* can be judged independently of the validity of predictions, the relation between the significance of a theory and the *realism* of its *assumptions* is almost the opposite of that suggested by the view under criticism. Truly important and significant hypotheses will be found to have *assumptions* that are wildly inaccurate descriptive representations of reality, and, in general, the more significant the theory, the more unrealistic the assumptions (in this sense)."

The Friedmanian version of instrumentalism has been very popular among economists; the result has been that the impression that abstract modeling was somehow at odds with a commitment to realism has been fostered. Independently of general philosophical arguments for or against instrumentalism about models, this version of instrumentalism seems unsuitable as a defence of social scientific modeling, especially when it comes to theoretical models. As discussed above, in the social sciences theoretical models can seldom be confronted with data to test their qualitative predictions and, when they are, their predictive record has not been spectacular. Thus, even if some models are used and defended because of the correctness of their qualitative predictions, instrumentalism about social scientific models is not widely entertained by philosophers, who have sought to offer alternative accounts of the epistemology of models.

#### 19.4.2 Isolation of Causal Mechanisms or Capacities

Another strategy for dealing with the problem of unrealistic assumptions is to argue that not every assumption has to be true, or descriptively accurate, in order for the model to tell something about the world. Even if there are differences between their philosophical commitments, influential advocates of this position include *Mäki* and *Cartwright* [19.16, 17, 23, 25, 28, 53–56]. As mentioned above, according to *Mäki*, unrealistic components of models serve to isolate a mechanism of interest from disturbing factors. It is precisely thanks to these falsehoods that the operation of a mechanism can be studied in isolation from interfering factors. In *Mäki's* interpretation, therefore, models with false components can deliver truths if and when the isolated mechanism resembles the real-world target in relevant respects and degrees [19.57].

Figure 19.2 depicts the way in which resemblance between model and target is supposed to explain why the model is a successful model of a given phenomenon. Success here is not only of a predictive kind. Mäki's account is meant to encompass explanation as well: His claim is that, if the model successfully explains a phenomenon, this is because it resembles the target in relevant respects and degrees. How this is established in practice constitutes an altogether different challenge, which as we will see partly induces some authors to question the extent to which models can in fact provide explanations. As an illustration of Mäki's position, consider again his interpretation of von Thünen's model (Sect. 19.A): The localization of agricultural activities is explained, at least in principle, in terms of the mechanism that relates distance and revenue; it is this mechanism which is to be judged either true or false. In other words, what matters is that the localization of economic activities depends on the distance from the market as described in the model, even if, due to the interference of other factors, we do not observe a concentric pattern. Mäki's account can be seen as a direct response to some criticisms of unrealistic models in social science, because it points out that the mere presence of false assumptions does not in itself prevent the possibility that the model is true about important aspects of the target system. Hence, unrealistic models should not be dismissed out of hand, but evaluated on a case-by-case basis.

According to Cartwright, models can be seen as isolating causal capacities. Many false assumptions are introduced with the purpose of building a hypothetical situation in which those capacities would act undisturbed from the effects of disturbing factors. Unlike Mäki, however, she maintains that abstract models, that is, models in which the operation of a capacity is examined by abstracting away from concrete situations or system-specific details, cannot be meaningfully interpreted as providing explanations of real-world phenomena. For such models to be used to understand real-world phenomena, they have to undergo a process of concretization: The factors that can potentially affect the operation of the isolated capacity in the concrete situation of interest should be reintroduced in the model.

Cartwright has been rather skeptical that many models in economics and in the social sciences more generally have the right features to be used for understanding the social world – for two main reasons [19.21]. First, socio-economic phenomena are often brought about by many causes, which do not combine vectorially making it hard if not impossible to predict their net effect when they interact. Moreover, these causes often do not have capacities stable

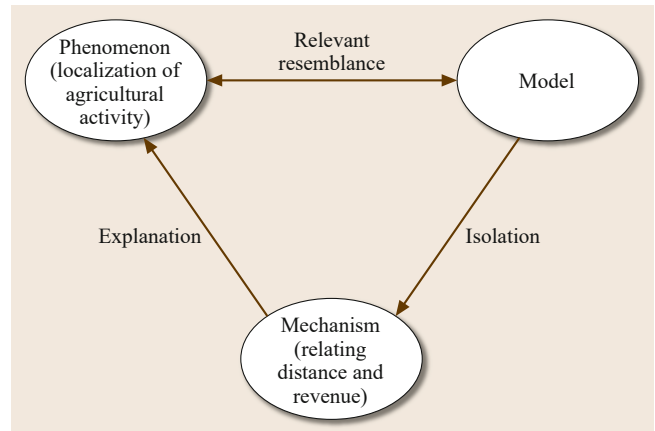


Fig. 19.2 Relevant resemblance between model and phenomenon as the basis for successful inference

enough to support generalizing from their behavior in the isolated model to their behavior in the world. Finally, compared to physics, the social sciences have few theoretical and empirical principles on which to rely for the derivation of conclusions about phenomena of interest. Therefore, deriving conclusions from these few theoretical and empirical principles requires a wide range of assumptions that do not serve the purpose of isolating capacities, but are instead needed to lend structure to the models while at the same time allowing the models to be (mathematically) tractable. Cartwright's concern is that the models' conclusions might be artifacts of these assumptions rather than genuine effects of the capacity in isolation. According to Cartwright, these assumptions are at risk of overly constraining the models, whose results, as a consequence, would often be artifacts of such over-constraining assumptions rather than genuine effects of the capacity in isolation. The scarcity of well-established theoretical principles is a problem that economics shares with the other social sciences. Nevertheless, one of the characteristics peculiar to economics is its strong commitment to a small set of axioms. Cartwright's worry might be less of a problem in fields that are ready to use a wide range of behavioral assumptions as well as to rely on agent-based modeling. This flexibility, however, might come at a cost: One often-heard criticism is that there is a flavor of the ad hoc in the way agent-based simulations are used in social science [19.7, 12].

### 19.4.3 Learning About Possibilities

Thus far, we have been taking for granted that models have specific real-world targets that they are taken to represent. Some social scientific models, however, do not seem to represent any specific target, and thus



they lack a representational link to the real world. Such target-less models deserve a discussion about whether they offer opportunities for learning about the world and, if they do, what kind of learning [19.44, 58–61]. Let us go back to Schelling’s model, which, according to *Grüne-Yanoff*, does not try to represent any particular city or any type of city, thereby making the issue of model-target similarity seem meaningless [19.59]. The only bit of the model that is informed by the real world is the assumption that people have preferences for not being in a minority. According to Grüne-Yanoff, evaluating the model as a representation by inquiring about its similarity to a city, or to cities in general, would force us to conclude that it is defective. The model, however, still offers opportunities to learn about the world, because it teaches us that, contrary to prior belief, residential segregation need not be brought about by racist preferences. By describing how it is possible for the phenomenon of segregation to come about rather than how the phenomenon has actually occurred, the model gives us a *how-possibly* explanation as opposed to a *how-actually* explanation [19.62]. For Grüne-Yanoff, learning from a model amounts to a justifiable change in one’s confidence in one or more necessity or impossibility hypotheses: We learn from Schelling’s model insofar as it justifiably changes “one’s confidence in hypotheses about racist preferences being a necessary cause of segregation” [19.60, p. 7]. Grüne-Yanoff’s approach makes sense of why social scientists often talk about models as indirectly providing insights about the world rather than offering specific hypotheses about real systems, and of why in some cases little effort is spent in applying models to those systems. However, the concept of learning, which is supposed to replace that of understanding or explanation, seems to be a heuristic rather than an epistemic one. Unless criteria are laid

out to specify how the model justifies one’s change in confidence, learning becomes a rather subjective affair [19.64, 65].

#### 19.4.4 Inferential Aids

According to an inferentialist approach to models, the question of how models can teach us about a target system is ill posed, because models are tools to help scientists’ inferential processes rather than autonomous entities capable of delivering information, learning or explanation [19.66–68]. Models are not abstract entities which social scientists manipulate to learn about something else; in fact, the very idea of manipulating an abstract entity sounds rather suspect. According to the inferentialists, models are tools that aid inferences from a set of assumptions to a conclusion; they help “to derive some conclusions about the empirical system, starting from information extracted from *this same system*” [19.67, p. 103; emphasis in the original]. As depicted in Fig. 19.3, the modeling activity (on the right-hand side) aids inferences from premises to conclusions about the phenomenon (on the left-hand side). Denotation, demonstration and interpretation refer to three stages of the modeling activity: First, aspects of the phenomenon are denoted by specific elements of the model, results are then derived within the model, and finally the results are interpreted again in terms of the phenomenon [19.63]. The special features of models as scientific tools are those that make them useful for inference-making by expanding our limited cognitive abilities. According to the inferentialist approach, the relationship between (good) representation and (correct) inference is inverted: It is its usefulness for drawing correct inferences that makes a model a good representation, not vice versa. In other words, while for

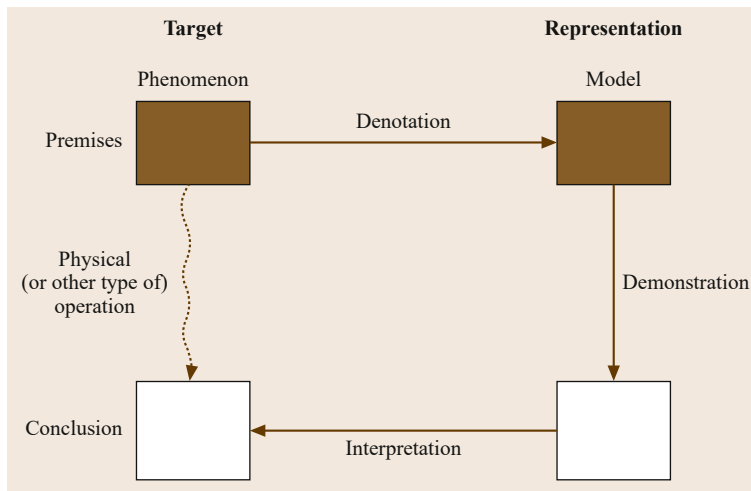


Fig. 19.3 The inferentialist view of models (after [19.63, p. 328])

representationalism a model helps draw correct inferences about its target only if it is a good representation, for inferentialism, if by using a model we are able to draw correct inferences about a target, then we can say that the model represents its target well enough. Saying this does not add much since there is nothing of substance in the relation of representation itself.

The difference between the two accounts can be further appreciated by comparing Fig. 19.2 and Fig. 19.3. Whereas in the inference described in Fig. 19.2 it is the symmetric relation of resemblance that makes a model a representation of the target phenomenon, in the case described in Fig. 19.3, it is first the activity of denotation and then that of interpretation that translate the inferences reached through the model into conclusions about a phenomenon. The goodness of a model, in turn, depends on the number and variety of successful inferences that it enables and on the ease with which it allows the user to draw those inferences. The inferentialist approach, however, raises the questions of how to establish whether the inferences are correct, and, if they are, what explains their success. Different answers are possible. De Donato and Zamora Bonilla maintain that what matters is success in prediction and intervention, but this answer casts doubt on a large portion of social scientific modeling: As mentioned above, few such models yield successful predictions and they rarely function as blueprints for interventions (however, see Sect. 19.4.5 for a counter-argument). In contrast, Kuorikoski and Ylikoski claim that explanatory success is itself explained by the model having captured some or another portion of the causal structure responsible for the phenomenon, bringing them close to the isolationist position.

#### 19.4.5 Models as Blueprints for the Design of Socio-Economic Mechanisms

Philosophers of social science have not only been interested in the fit between hypothetical models and the phenomena they seek to represent, but also in the inverse relation, namely, in the role of hypothetical models in the design of socio-economic mechanisms. An example of successful institutional design guided by theoretical models – and of the broader phenomenon of economic engineering – is the auction mechanism, which a group of economists was asked to draw up for the efficient distribution of radio-electronic frequencies by the federal communications commission (FCC). *Guala* regards the FCC example as an instance in which, rather than the theoretical model trying to represent a real-world phenomenon, the real world is molded so as to resemble the model as closely as possible [19.69, p. 456; emphasis in the original]:

“According to a widely shared view of scientific knowledge, the main task of the theorist is to explain spontaneously occurring and experimental processes, by designing an appropriate model [...] The FCC case belongs to an altogether different kind of scientific activity, proceeding in the opposite direction, *from models to mechanisms and processes.*”

Obviously, the process from the model to the design of the mechanism is not a straightforward one. In the FCC auction mechanism, for example, there was no simple way of implementing existing and highly abstract auction models because the real-world situation had specificities that needed to be taken into account. This required tinkering with models as well as probing them experimentally in a back-and-forth process that led to the design of the mechanism that was eventually implemented. Interestingly, it is precisely the case of the FCC auction mechanism that *Alexandrova* uses to argue that models do not provide explanations that are directly applicable to real-world situations [19.70].

According to *Alexandrova*, the theoretical models were only indirectly implicated in this successful case of economic engineering; rather it is the experimental efforts that should be credited with the achievement. By choosing this example, she seems to suggest that there is nothing relevantly different between using models to *explain* the working of existing institutions and using them to design new institutional arrangements; in both cases, the models merely provide templates for the formulation of causal hypotheses. In particular, *Alexandrova* proposes the view that models are “open formulae” taking the following form: “In a situation of type  $x$  with some characteristics that may include  $C_1 \dots C_n$ , a certain feature  $F$  causes a certain behavior  $B$ ”, where  $x$  is a variable that needs to be specified,  $F$  and  $B$  refer to cause and effect and the  $C_i$  refers to the conditions under which  $F$  causes  $B$ . It is only when  $x$  is specified that the open formula becomes a causal claim [19.70, 71].

The reason is that often it is either impossible to know whether the modeling assumptions are satisfied in a particular context of application, or it is known that they cannot be satisfied at all. To illustrate, let us return to von Thünen’s model. While the open formula would be something like *in a situation such as type  $x$ , transportation costs  $T$  cause the location of economic activities according to pattern  $L$* , a causal hypothesis would say that *in a situation in which the transportation costs depend only on the kind of good to be transported and on the distance from the market ( $c_i = f(d)$ ), the costs cause a concentric distribution of economic activities around the market (plus other conditions)*. This

hypothesis needs to be subjected to empirical or experimental testing. The characterization of the kind of situation in which  $F$  causes  $B$  need not correspond to the assumptions of the model, however. So the causal hypothesis will not specify that the town has no spatial dimension (assumption (5) in Sect. 19.A), as we already know that this assumption cannot be satisfied by any real-world system.

Whether or not it is in fact the experimental efforts that ultimately identified the successful auction mechanism, the FCC case points to the possibility that highly abstract models, which are not representations of any real-world phenomenon to begin with, can nevertheless be used as guides to the design of successful institutional arrangements. As *Guala* suggests, such uses of models distinguish the engineering ambitions of the social sciences – an aspect of model-based social science connected with wider debates about the allegedly unique capacity of the social sciences to influence their object of study [19.69].

#### 19.4.6 Where Do We Go From Here?

The question of when it is legitimate to use theoretical models for epistemic and practical purposes is not yet settled. It appears to be rather uncontroversial that different *kinds* of models are suitable for aiding different *kinds* of inferences: Some models can be used to make inferences about specific targets (the pattern of con-

centric rings around a particular city) or about generic targets (the localization of agricultural activity), while other models do not have a target at all. Whereas in the latter case the inferences might be of the how-possibly kind, in the former cases, the model might be said to be explanatory or to help with explanatory inferences, provided other conditions are also met. The challenge is to identify such conditions. Although, according to a very general principle, for purposes of explanation and possibly reliable prediction and control, a model should somehow capture the relevant features of the phenomenon it targets, different bits of the model need to be empirically confirmed, depending on the kind of inferences at stake. This is where model manipulation in the form of robustness analysis becomes important: The manipulation of modeling assumptions helps to identify which components of the model are crucial for obtaining the result we are interested in. It is these components that need to be empirically valid, not the whole. This seems to hold also when the model's crucial assumptions are satisfied, not because they accurately describe relevant features of their target, but because the conditions for those assumptions to hold true have been created by design. Although the real world, however engineered, will rarely, if ever, approximate the model in every detail, at least in some successful instances it can be molded so that the relevant features – those that drive the modeling result – are in place.

### 19.5 Conclusions

The social sciences are now undergoing significant methodological change. Experimentation both in the laboratory and in the field has become an important addition to social scientists' toolkit, not only for theory testing, but also for theory formation and policy design. There are also important new developments related to the availability of large databases, as well as means to analyze them that were previously unavailable. In all this, not only has modeling become more widespread, but also new modeling techniques such

as agent-based simulation are making headway. It has even been suggested that computational science may entail a new reorganization of the sciences around computational templates that cut across the natural and social sciences [19.72]. The place of modeling within the arrays of styles that characterize the social sciences is significantly changing, and it remains to be seen how the different styles will interact to produce scientific knowledge about social and economic phenomena.

## 19.A Appendix: J.H. von Thünen's Model of Agricultural Land Use in the Isolated State

Von Thünen's model of agricultural land use describes how the distance from a market affects the distribution of agricultural productions around a city [19.73]. This model is considered to be one of the first examples of modern economic modeling, and it is still a classic model in geography and urban economics from which an entire tradition of models of land use in urban spaces has originated. Von Thünen's model has also received some attention in the philosophy of economics, and thanks to its analytical simplicity, it is particularly suitable for illustrating some of the ideas discussed in this chapter [19.3, 23, 74].

Von Thünen's localization model is based on a set of assumptions that describes a homogeneous and isolated agricultural space in which a single town is located:

1. The area is a plain, i. e., there are no mountains or valleys
2. There are no streets or navigable rivers
3. The plain is completely cut off from the outside world
4. Climate and fertility are uniform across space
5. The town is located centrally and has no spatial dimension
6. All markets and industrial activities take place in the town
7. Production costs are constant across space
8. Transportation costs are directly proportional to the distance, the weight and the perishability of the goods
9. Selling prices are fixed and the demand is unlimited
10. Farmers have complete information and they act to maximize their revenue.

Under these assumptions, a pattern of concentric rings around the town emerges. Dairying and intensive farming (vegetables and fruit) occupy the ring closest to the town, because these products are perishable and incur the highest transportation costs. Timber and firewood are located in the second ring, because wood is heavy and hence difficult and costly to transport. The third ring consists of extensive farming of crops, such as grain for bread, that are more durable than fruit and less heavy than wood. On the outermost ring stock farming and cattle ranching take place, because animals can walk to the city to be sold at the market and thus have low transportation costs.

This result can also be described in analytical terms by determining which production is most profitable at different distances from the town.

The revenue  $r$  of each agricultural production consists in its selling price  $p$  minus its production and transportation costs. Since the selling price and the unitary production and transportation costs are fixed, the revenue depends only on the distance from the city

$$r^i(d) = (p - c)x - tdx \quad i = \{A, B, C\},$$

where  $x$  is the quantity of the good,  $c$  the production cost per unit,  $t$  the transportation cost per unit and  $d$  the distance from the market. The apex  $i$  indicates the kind of agricultural production: Dairying and intensive farming  $A$ , timber and firewood  $B$ , and extensive farming  $C$ .

The slope of each revenue curve depends on transportation cost and distance  $-td$ .

The descending curves in Fig. 19.4 represent the revenue of each production depending on its distance from the town; e.g., at distance  $a$  it becomes more profitable to produce product  $B$ .

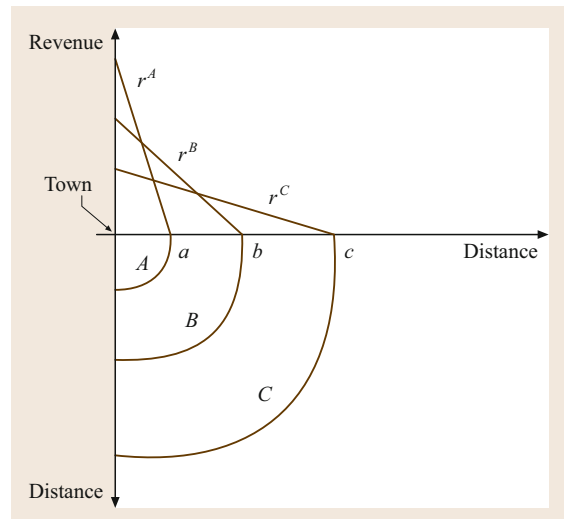


Fig. 19.4 The production revenue and the land use in von Thünen's model (after [19.75, p. 76])

## 19.B Appendix: T. Schelling's Agent-Based Model of Segregation in Metropolitan Areas

Thomas Schelling's work on racial segregation paved the way for the use of simulations in the study of social phenomena. In his seminal work, *Schelling* studied how macro-phenomena, such as segregation, can emerge as an unintended effect of the combination of many interrelated decisions [19.76]. Racial sorting is a case in point. Segregation has been proven to occur as a side effect of the preference of individuals for having a few neighbors of the same ethnic group, rather than as the consequence of a preference for segregation itself.

Schelling represented the segregation process by means of a checkerboard and dimes and pennies, standing respectively for a certain metropolitan area and for the individuals of two different groups (Fig. 19.5). The model is based on a set of assumptions that describe an

idealized metropolitan area and its inhabitants. Examples of such assumptions are:

1. There are only two kinds of agents, Blacks and Whites
2. Agents' decisions only depend on preferences regarding their neighbors
3. The city is uniform, i. e., there are no architectural or topological boundaries that constrain individual choices
4. Agents move randomly in space
5. There are no costs of moving from one point to another.

On the checkerboard it is possible to track the movements of the agents and to observe how the configuration of the neighborhood changes over time.

The resulting dynamics reflect the individual decisions to move to areas whose composition meets the agents' preferences. Rather than obtaining analytical solutions, Schelling's model explores the conditions under which segregation emerges by means of local rules. What it shows is that, regardless of the initial position of the agents and the spatial configuration, given a certain range of people's preferences, clusters of neighbors of the same color eventually emerge.

Even though agent-based models do not need to be implemented on a computer, nowadays they are of-

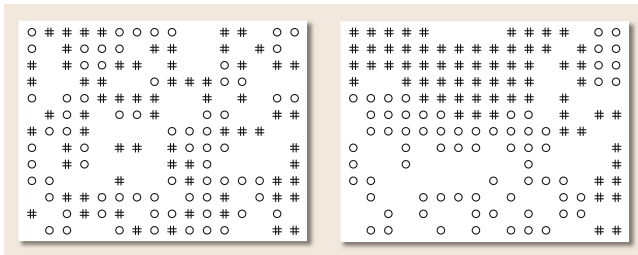


Fig. 19.5 Schelling's checkerboard: Initial and final configuration (after [19.76, p. 155–157])

```
globals [
  percent-similar ;; on the average, what percent of a turtle's neighbours
                  ;; are the same color as that turtle?
  percent-unhappy ;; what percent of the turtles are unhappy?
]

turtles-own [
  happy? ;; for each turtle, indicates whether at least %-similar-wanted percent of
          ;; that turtles' neighbours are the same color as the turtle
  similar-nearby ;; how many neighbouring patches have a turtle with my color?
  other-nearby ;; how many have a turtle of another color?
  total-nearby ;; sum of previous two variables
]

to setup
  clear-all
  if number > count patches
  [ user-message (word "This pond only has room for " count patches " turtles.")
    stop ]

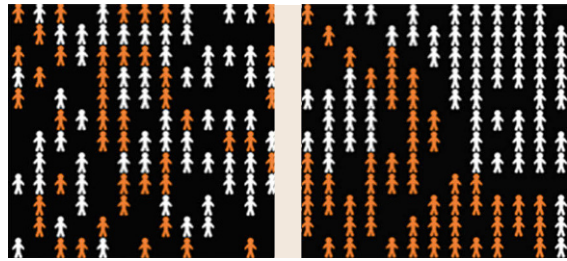
  ;; create turtles on random patches.
  ask n-of number patches
  [ sprout 1
    [ set color red ] ]
  ;; turn half the turtles green
  ask n-of (number/2) turtles
  [ set color green ]
  update-variables
  reset-ticks
end
```

Fig. 19.6 Netlogo code of Schelling segregation model (after [19.77])

ten used together. The premise is to build a model that captures the relevant variables of the agents' decisions, such as personal preferences and responses to other agents' behavior and to the context. Next, a way has to be found to implement the model and the other components that characterize the system – such as the network structure – in a computer code. Call the set of relevant factors that are external to the model its *environment*. Together, the model and the environment constitute the algorithm that runs on the computer.

Figure 19.6 shows an extract of the algorithm of the segregation model implemented on NetLogo. Each run of the program corresponds to a step in the simulation, which in turn represents a change in the system. The evolution of the system can be represented graphically by means of software that transforms the numerical analysis into visual representations (Fig. 19.7).

**Acknowledgments.** Alessandra Basso was mainly responsible for writing Sects. 19.2 and 19.B, Chiara Lisciandra, Sects. 19.3 and 19.B, Caterina Marchionni,



**Fig. 19.7** Visual representation of Schelling's segregation model (after [19.77])

Sect. 19.1 and Sect. 19.4. Section 19.4 draws extensively on A. Basso, C. Marchionni: *I Modelli in Economia*, APhEx (2015).

We thank our colleagues at TINT for helpful comments on an earlier draft of the chapter. In particular, we thank Aki Lehtinen, Miles MacLeod, Jaakko Kuorikoski and Till Grüne-Yanoff. Special thanks goes to Juho Pääkkönen for his invaluable assistance. All remaining mistakes are obviously ours.

## References

- 19.1 A.C. Crombie: *Styles of Scientific Thinking in the European Tradition* (Gerald Duckworth, London 1995)
- 19.2 I. Hacking: The disunities of the sciences. In: *The Disunity of Science: Boundaries, Context and Power*, ed. by P. Galison, D. Stump (Stanford Univ. Press, Palo Alto 1996) pp. 37–74
- 19.3 M.S. Morgan: *The World in the Model: How Economists Work and Think* (Cambridge Univ. Press, Cambridge 2012)
- 19.4 U. Mäki: Symposium on explanations and social ontology 2: Explanatory ecumenism and economics imperialism, *Economics Phil.* **18**, 237–259 (2002)
- 19.5 K.A. Clarke, D.M. Primo: *A Model Discipline. Political Science and the Logic of Representation* (Oxford Univ. Press, Oxford 2012)
- 19.6 C.R. Edling: Mathematics in sociology, *Annu. Rev. Sociol.* **28**, 197–220 (2002)
- 19.7 C. Marchionni: Playing with networks: How economists explain, *Eur. J. Philos. Sci.* **3**(3), 331–352 (2013)
- 19.8 J. Kuorikoski, C. Marchionni: Unification and mechanistic detail as drivers of model construction: Models of networks in economics and sociology, *Stud. Hist. Philos. Sci.* **48**, 97–104 (2014)
- 19.9 P. Godfrey-Smith: The strategy of model-based science, *Biol. Philos.* **21**, 725–740 (2006)
- 19.10 M. Boumans: *How Economists Model the World into Numbers* (Routledge, London 2005)
- 19.11 R. Backhouse: Representation in economics. In: *Measurement in Economics: A Hand Book*, ed. by M. Boumans (Elsevier, Amsterdam 2007) pp. 135–152
- 19.12 J. Kuorikoski, C. Marchionni: Broadening the perspective: Epistemic, social, and historical aspects of scientific modeling, *Perspect. Sci.* (2015), doi:10.1162/POSC\_e\_00179
- 19.13 U. Webster, J. Sell (Eds.): *Laboratory Experiments in the Social Sciences* (Elsevier, Amsterdam 2007)
- 19.14 F. Guala: *The Methodology of Experimental Economics* (Cambridge Univ. Press, Cambridge 2005)
- 19.15 R.B. Morton, K.C. Williams: *Experiment in Political Science and the Study of Causality: From Nature to the Lab* (Cambridge Univ. Press, Cambridge 2010)
- 19.16 U. Mäki: On the method of isolation in economics. In: *Idealization IV: Intelligibility in Science*, ed. by C. Dilworth (Rodopi, Amsterdam 1992) pp. 317–351
- 19.17 N. Cartwright: *Nature's Capacities and Their Measurement* (Clarendon, New York 1989)
- 19.18 U. Mäki: Models are experiments, experiments are models, *J. Econ. Methodol.* **12**, 303–315 (2005)
- 19.19 M.S. Morgan: Experiments versus models: New phenomena, inference and surprise, *J. Econ. Methodol.* **12**, 317–329 (2005)
- 19.20 F. Guala: Models, simulations, and experiments. In: *Model-Based Reasoning: Science, Technology, Values*, ed. by L. Magnani, N. Nersessian (Kluwer Academic/Plenum, New York 2002) pp. 59–74
- 19.21 N. Cartwright: The vanity of rigour in economics: Theoretical models and Galileian experiments. In: *The Experiment in the History of Economics*, ed. by P. Fontaine, R. Leonard (Routledge, London 1999) pp. 135–153
- 19.22 M.S. Morgan: Model experiments and models in experiments. In: *Model-Based Reasoning: Science, Technology, Values*, ed. by L. Magnani, N. Nersessian (Kluwer Academic/Plenum, New York 2002) pp. 41–58

- 19.23 U. Mäki: Models and the locus of their truth, *Synthese* **180**(1), 47–63 (2011)
- 19.24 M.S. Morgan: Models, stories and the economic world, *J. Econ. Methodol.* **8**(3), 361–384 (2001)
- 19.25 U. Mäki: MISSING the world. Models as isolations and credible surrogate systems, *Erkenntnis* **70**(1), 29–43 (2009)
- 19.26 B.C. Van Fraassen: *Scientific Representation: Paradoxes of Perspective* (Oxford Univ. Press, Oxford 2008)
- 19.27 W.S. Parker: Does matter really matter? Computer simulations, experiments, and materiality, *Synthese* **169**, 483–496 (2009)
- 19.28 N. Cartwright: *Hunting Causes and Using Them* (Cambridge Univ. Press, Cambridge 2007)
- 19.29 J. Kuorikoski, A. Lehtinen, C. Marchionni: Economic modelling as robustness analysis, *Br. J. Philos. Sci.* **61**(3), 541–567 (2010)
- 19.30 J. Kuorikoski, A. Lehtinen, C. Marchionni: Robustness analysis disclaimer: Please read the manual before use!, *Biol. Philos.* **27**(6), 891–902 (2012)
- 19.31 J. Odenbaugh, A. Alexandrova: Buyer beware: Robustness analyses in economics and biology, *Biol. Philos.* **26**(5), 757–771 (2011)
- 19.32 C. Deissenberg, S. Hoog, H. Dawid: EURACE: A massively parallel agent-based model of the European economy, *Appl. Math. Comput.* **204**, 541–552 (2008)
- 19.33 V. Spaizer, P. Hedstrom, S. Ranganathan, K. Jansson, M. Nordvik, and D. Sumpter: Identifying Complex Dynamics in Social Systems: A New Methodological Approach Applied to Study School Segregation, *Sociological Methods and Research*, Art.No. 0049124116626174 (2016)
- 19.34 E. Winsberg: Computer simulations in science. In: *The Stanford Encyclopedia of Philosophy*, ed. by E. Zalta <http://plato.stanford.edu/archives/sum2013/entries/simulationscience> (Summer 2013 Edition)
- 19.35 R. Axelrod: More effective choice in the prisoner's dilemma, *J. Confl. Resolut.* **24**(3), 379–403 (1980)
- 19.36 R. Axelrod, W.D. Hamilton: The evolution of cooperation, *Science* **211**(4489), 1390–1396 (1981)
- 19.37 J.M. Epstein: *Generative Social Science: Studies in Agent-Based Computational Modeling* (Princeton Univ. Press, New Jersey 2006)
- 19.38 P. Hedström: *Dissecting the Social: On the Principles of Analytical Sociology* (Cambridge Univ. Press, Cambridge 2005)
- 19.39 C. Marchionni, P.K. Ylikoski: Generative explanation and individualism in agent-based simulation, *Philos. Soc. Sci.* **43**(3), 323–340 (2013)
- 19.40 S. Hartmann: The world as a process: Simulations in the natural and social sciences. In: *Modelling and Simulation in the Social Sciences from the Philosophy of Science Point of View*, ed. by R. Hegselmann, U. Mueller, K. Troitzsch (Kluwer, Dordrecht 1996) pp. 77–100
- 19.41 P. Humphreys: Computer simulation. In: *PSA 1990*, Vol. 2, ed. by A. Fine, M. Forbes, L. Wessels (The Philosophy of Science Association, East Lansing 1990) pp. 497–506
- 19.42 A. Lehtinen, J. Kuorikoski: Computing the perfect model – Why do economists shun simulation?, *Philos. Sci.* **74**, 304–329 (2007)
- 19.43 N. Gilbert, K.G. Troitzsch: *Simulation for the Social Scientist* (Open Univ. Press, Buckingham 1999)
- 19.44 M. Weisberg: *Simulation and Similarity: Using Models to Understand the World* (Oxford Univ. Press, Oxford 2013)
- 19.45 M. Morrison: Models, measurement and computer simulation: The changing face of experimentation, *Philos. Stud.* **143**(1), 33–57 (2009)
- 19.46 M.T. Hannan, J. Freeman: The population ecology of organizations, *Am. J. Sociol.* **82**(5), 929–964 (1977)
- 19.47 R. Feynman: Simulating physics with computers, *Int. J. Theor. Phys.* **21**, 467–488 (1982)
- 19.48 R. Muldoon: Robust simulations, *Philos. Sci.* **74**(5), 873–883 (2007)
- 19.49 E. Bruch, R. Mare: Neighborhood choice and neighborhood change, *Am. J. Sociol.* **112**, 667–709 (2006)
- 19.50 J. Zhang: A dynamic model of residential segregation, *J. Math. Sociol.* **28**, 147–170 (2004)
- 19.51 R. Frigg, J. Reiss: The philosophy of simulation: Hot new issues or same old stew, *Synthese* **169**, 593–613 (2009)
- 19.52 M. Friedman: The methodology of positive economics. In: *Essays in Positive Economics*, ed. by M. Friedman (Chicago Univ. Press, Chicago 1953) pp. 3–43
- 19.53 U. Mäki: Isolation, idealization and truth in economics, *Poznan Stud. Philos. Sci. Humanit.* **38**, 147–168 (1994)
- 19.54 U. Mäki: Realistic realism about unrealistic models. In: *The Oxford Hand Book of Philosophy of Economics*, ed. by D. Ross, H. Kincaid (Oxford Univ. Press, Oxford 2009) pp. 68–98
- 19.55 N. Cartwright: Capacities. In: *The Hand Book of Economic Methodology*, ed. by J. Davis, W. Hands, U. Mäki (Elgar, Northampton 1999) pp. 45–48
- 19.56 N. Cartwright: If no capacities then no credible worlds. But can models reveal capacities?, *Erkenntnis* **70**, 45–58 (2009)
- 19.57 R.N. Giere: *Explaining Science: A Cognitive Approach* (Univ. Chicago Press, Chicago 1988)
- 19.58 T. Grüne-Yanoff: Learning from minimal economic models, *Erkenntnis* **70**, 81–99 (2009)
- 19.59 T. Grüne-Yanoff: Appraising models non-representationally, *Philos. Sci.* **80**(5), 850–861 (2013)
- 19.60 N.E. Aydinonat, P. Ylikoski: Understanding with theoretical models, *J. Econ. Methodol.* **21**(1), 19–36 (2014)
- 19.61 L. Casini: Not-so-minimal models: Between isolation and imagination, *Philos. Soc. Sci.* **44**(5), 646–672 (2014)
- 19.62 N.E. Aydinonat: Models, conjectures and explanation: An analysis of Schelling's checkerboard model of residential segregation, *J. Econ. Methodol.* **14**, 429–454 (2007)
- 19.63 R.I.G. Hughes: Models and representation, *Proc. Philos. Sci.*, Vol. 64 (1997) pp. S325–S336
- 19.64 R. Northcott, A. Alexandrova: It's just a feeling: Why economic models do not explain, *J. Econ.*

- Methodol. **20**(3), 262–267 (2013)
- 19.65 J. Reiss: The explanation paradox, *J. Econ. Methodol.* **19**(1), 43–62 (2012)
- 19.66 J. Kuorikoski, A. Lehtinen: Incredible worlds, credible results, *Erkenntnis* **70**, 119–131 (2009)
- 19.67 X. de Donato Rodríguez, J.Z. Bonilla: Credibility, idealisation, and model building: An inferential approach, *Erkenntnis* **70**(1), 101–118 (2009)
- 19.68 J. Kuorikoski, P. Ylikoski: External representations and scientific understanding, *Synthese* **192**(12), 3817–3837 (2015)
- 19.69 F. Guala: Building economic machines: The FCC auctions, *Stud. Hist. Philos. Sci.* **32**, 453–477 (2001)
- 19.70 A. Alexandrova: Connecting economic models to the real world: Game theory and the FCC spectrum auctions, *Philos. Soc. Sci.* **36**(2), 173–192 (2006)
- 19.71 A. Alexandrova: Making models count, *Philos. Sci.* **75**(3), 383–404 (2008)
- 19.72 P. Humphreys: *Extending Ourselves: Computational Science, Empiricism, and Scientific Method* (Oxford Univ. Press, New York 2004)
- 19.73 J.H. von Thünen: *Der isolierte Staat in Beziehung auf Landwirtschaft und Nationalökonomie* (Phuthes, Hamburg 1826)
- 19.74 U. Mäki: Realism and the nature of theory: A lesson from J H von Thünen for economists and geographers, *Environ. Plan. A* **36**(10), 1719–1736 (2004)
- 19.75 R. Capello: *Economia Regionale* (Il Mulino, Bologna 2004)
- 19.76 T. Schelling: Dynamic models of segregation, *J. Math. Sociol.* **1**, 143–186 (1971)
- 19.77 U. Wilensky: NetLogo Segregation model (Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston 1997), <http://ccl.northwestern.edu/netlogo/models/Segregation>



# Model-Based

## 20. Model-Based Diagnosis

Antoni Ligęza, Bartłomiej Górny

Diagnostic reasoning is an activity aimed at finding the causes of incorrect behavior of various technological systems. In order to perform diagnosis, a typical diagnostic system should be equipped with the expert knowledge of the domain and statistical evidence of former failures. More advanced solution combines model-based reasoning (MBR) and abduction. It is assumed that a *model* of the system under investigation is specified. Such a model allows us to simulate the normal behavior of the system. It can also be used to detect incorrect behavior and perform sophisticated reasoning in order to identify *potential causes* of the observed failure. Such potential causes form a set of possible diagnoses. In this chapter, formal bases for the so-called model-based diagnostic reasoning paradigm are presented and application examples are discussed in detail. A method of modeling system behavior with the use of causal graphs is put forward. Then, a systematic method for discovering all the so-called *conflict sets* (disjunctive conceptual faults) is described. Such conflict sets describe sets of elements in such a manner that in order to explain the observed misbehavior at least one of them must be faulty. By selecting and removing such elements from all conflicts sets – for each conflict set one such element – the proper candidate diagnoses are generated. An example of the application of the proposed methods to the three-tank dynamic system is presented and some bases for on-line generation of diagnoses for dynamic systems are outlined, together with some theorems. The chapter introduces an easy and self-contained material being an introduction to modern model-based diagnosis, covering static and dynamic systems.

20.1	<b>A Basic Model for Diagnosis</b> .....	437
20.2	<b>A Review and Taxonomy of Knowledge Engineering Methods for Diagnosis</b> .....	438
20.2.1	Knowledge Engineering.....	438
20.2.2	Expert Methods.....	439
20.2.3	Model-Based Methods.....	439
20.3	<b>Model-Based Diagnostic Reasoning</b> ..	440
20.4	<b>A Motivation Example</b> .....	440
20.5	<b>Theory of Model-Based Diagnosis</b> .....	442
20.6	<b>Causal Graphs</b> .....	444
20.7	<b>Potential Conflict Structures</b> .....	446
20.8	<b>Example Revisited. A Complete Diagnostic Procedure</b> .....	448
20.9	<b>Refinement: Qualitative Diagnoses</b> ...	450
20.9.1	Qualitative Evaluation of Faults.....	450
20.9.2	Elimination of Spurious Diagnoses.....	450
20.9.3	Deduction for Enhanced Diagnoses.....	452
20.9.4	Analysis of Diagnoses.....	452
20.10	<b>Dynamic Systems Diagnosis: The Three-Tank Case</b> .....	454
20.11	<b>Incremental Diagnosis</b> .....	456
20.12	<b>Practical Example and Tools</b> .....	458
20.13	<b>Concluding Remarks</b> .....	459
	<b>References</b> .....	460

*Diagnostic reasoning* can be considered as a *cover name* for man and machine inference activities aimed at discovering what is wrong when some systems do not work as expected [20.1–3]. In fact, it constitutes a set of mutually complementary paradigms of inference with the ultimate goal to produce a set of rational explanations of the observed misbehavior of some systems under consideration. Diagnostic reasoning often combines *causal reasoning* with domain experience including statistical and *expert knowledge*. Hence, diagnostic reasoning makes use of the so-called *shallow knowledge* and in the case of the *model-based diagnosis* the *deep knowledge*.

*Shallow knowledge* refers to empirical knowledge based on numerous observations and experience, typical for practitioners, usually encoded with rules and taking the form of the so-called *expert system* [20.4]. The diagnosed system is usually considered as a *black box* where only inputs and outputs are known, and the expert system covers the operational knowledge about its behavior, properties, and possible failures in the form of if–then rules. The development of the rules is based on experience documented with accumulated records.

On the other hand, *deep knowledge* refers to the contents of the box (the term *white box* is sometimes used) and mathematical models of system behavior. Hence, deep knowledge is the knowledge of the internal structure, components, and their interactions. It allows us to simulate the behavior of the system for any admissible input conditions.

Diagnosis is usually carried out by domain experts, and expert knowledge hardly undergoes any smart formalization. Nevertheless, some most successful approaches mimic diagnostic reasoning by the combination of *abduction/deduction* including causal inference, *model-based reasoning* (MBR), and *case-based reasoning* (CBR). Let us briefly explain the meaning of these terms in practice.

*Abduction* [20.5, 6] consists in looking for valid explanations for observed effects. In general, it is not a valid inference rule. The result of abductive inference is a set of hypotheses explaining what is observed; the hypotheses must be consistent with the knowledge at hand.

*Deduction* [20.6–8] consists in looking for consequences of some assumptions and initial knowledge at hand. In general, it is a valid inference rule. The result of deductive inference is a set of facts true under the assumptions and background knowledge.

*Model-based reasoning* [20.1, 2] refers to reasoning about system behavior and properties on the basis of its mathematical model. Note that no practical experience or observations are necessary. Complete information on the system can be gathered during a single session of model investigation.

In contrast to deep knowledge and MBR, *CBR* [20.9–11] is in opposite to using deep knowledge and MBR. It consists in gathering a number of *cases* (in the case of diagnosis – failure descriptions and fault identifications) to be stored and used as patterns for solving new problems. Case-based reasoning looks for an identical case which occurred in the past, or a *similar* one; in the latter case, reasoning by analogy can be used.

The process of fault diagnosis of technical systems typically requires the use of different methods of knowledge representation and inference paradigms. The most common scenario of such a process consists of the detection of the faulty behavior of the system, classification of this behavior, search for and determination of potential causes of the observed misbehavior, that is, generation of potential diagnoses, verification of those hypothetical diagnoses, and selection of the correct one, and finally a repair phase.

There exist a number of approaches and diagnostic procedures having their origin in very different branches of science, such as mechanical engineering, electrical engineering, electronics, automatic control, or computer science. In the diagnosis of complex dynamical systems, approaches from automatic control play an important role ([20.12–14]; a good state of the art can be found in the handbook [20.3]). The point of view of computer science, and especially artificial intelligence (AI) is presented, for example, in [20.1, 2, 15]. A good comparative analysis of some selected approaches was presented in [20.16] and in [20.17]. A recent, comprehensive in-depth study aimed at comparison of approaches emerging from AI and from classical automatic control is presented in [20.18].

This chapter is devoted to the presentation of some selected approaches originating from AI, located in the area of *model-based reasoning* (or MBR) and based on *consistency-based reasoning* [20.7, 19–22]. The presentation is based on the authors' experience and some former publications including [20.23–29]. Many concepts and results were prepared during the work on [20.30].

## 20.1 A Basic Model for Diagnosis

Consider the mathematical point of view on the diagnostic process. Taking such a viewpoint, the problem of building a diagnostic system, including issues of representation and acquisition of diagnostic knowledge, as well as the implementation of diagnostic reasoning engine, can be considered as one way of searching an inverse function or relation.

In fact, the problem of searching for diagnoses can be considered as an inverse task to extended simulation task; some specific features are as follows: (i) one observes a faulty behavior of the analyzed system (and thus, apart from the knowledge about correct behavior, also the one about faulty behavior should be accessible), and then (ii) taking into account the observed state (output), the main goal is not the reconstruction of the input (control) but rather the causes of the failure manifestations. Such causes can be considered as faulty components of the system, or, at some more detailed level, as wrong (faulty) parameter values produced by these components.

For simplicity, assume that one of  $n$  system components can become faulty, where each elementary fault is of binary character – such a component can just work or do not work correctly. Let  $D$  denote the set of potential elementary causes to be considered and let  $D = \{d_1, d_2, \dots, d_n\}$ . Faulty behavior of the system can be stated (detected) through the observation of one or more symptoms of failure. Assume that there are  $m$  such symptoms to be considered and their evaluation is also binary. Let  $M$  denote the set of such symptoms, where  $M = \{m_1, m_2, \dots, m_m\}$ . The detection of a failure consists in the detection of the occurrence of at least one symptom  $m_i \in M$ . In general, some subset  $M^+ \subseteq M$  of the symptoms can be observed to be true in the case of a failure. The goal of the diagnostic process is the generation of a diagnosis being any set  $D^+ \subseteq D$ , such that taking into account the *domain expert knowledge* and the *system model* it explains the observed misbehavior.

Let  $KB$  denote the *knowledge base* – in our case, the model of the system behavior. Furthermore, let  $D^-$  denote the components that are assumed to work correctly,  $D^+ \cup D^- = D$ , and let  $M^-$  denote the failure manifestations that are absent,  $M^+ \cup M^- = M$ . More formally, any valid candidate diagnosis  $D^+$  must satisfy the following conditions

$$D^+ \cup D^- \cup KB \models M^+ \cup M^-, \quad (20.1)$$

and

$$D^+ \cup D^- \cup KB \cup M^+ \cup M^- \not\models \perp. \quad (20.2)$$

Condition (20.1) means that the diagnosis must explain the observed misbehavior in view of the accessible knowledge about the system. Condition (20.2) means that the diagnosis must be consistent with the accessible knowledge about the system and the currently observed manifestations.

In the general case, the result of the diagnostic process can consist of one or more potential diagnoses; these diagnoses – subsets of the set  $D$  – can be single-element sets (i. e., the so-called *elementary diagnoses*) or multielement ones. For simplicity, in a number of practical approaches, only single-element diagnoses are taken into account. In the case of complex, multielement diagnoses, the discussion is frequently restricted to the so-called *minimal diagnoses*, that is, subsets of  $D$  which explain the observed misbehavior in a satisfactory way and simultaneously such that all elements of them are necessary for the justification of the diagnosis.

For the sake of general consideration, it can be assumed that there exists causal dependency between elementary faults represented by the elements of  $D$  and failure symptoms represented by the elements of  $M$ . Hence, there exists some relation  $R_C$  (i. e., a *causal relation*), such that

$$R_C \subseteq 2^D \times 2^M, \quad (20.3)$$

that is, any failure being defined as a subset of  $D$  is assigned one or more sets of possible failure symptoms; in certain particular cases, the failure – although occurred – may also be unobservable.

Such approach, however, is indeterministic: a single failure may be assigned *several different* sets of symptoms of the observed misbehavior. Therefore, it is frequently assumed that the causal dependency  $R_C$  is of functional nature, that is,  $R_C$  is, in fact, the following function

$$R_C : 2^D \rightarrow 2^M. \quad (20.4)$$

In this approach, any failure causes some unique and well-defined set of symptoms to occur. In this case, the task of building a diagnostic system consists in finding the inverse function, that is, the so-called *diagnostic function*  $f$ , where  $f = R_C^{-1}$ .

Unfortunately, in the case of many realistic systems, the function  $R_C$  is not a one-to-one mapping, so there does not exist the inverse mapping in the form of a function.

Consider a simple example of such a system; let it be the set of  $n$  bulbs connected serially (e.g., a set

for a Christmas tree). An elementary diagnosis  $d_i$  is equivalent to the  $i$ th bulb being blown. However, the set of manifestations of the failure  $M = \{m_1\}$  is a single-element set, where  $m_1$  indicates that the bulbs are not switched on. Even in the case, when the analysis is restricted to considering single-element elementary diagnoses, there exists  $n$  potentially equivalent diagnoses; each of them causes the same result with  $m_1$  being true. If multielement diagnoses are admitted, then there exist  $(2^n - 1)$  potential diagnoses.

In practice, the development of a diagnostic system consists in finding the inverse relation  $R_C^{-1}$ , and more precisely it searches for this inverse relation during the diagnostic process. In many practical diagnostic systems, the diagnostic process is interactive, and additional tests and measurements can be undertaken in

order to restrict the area of search. In the case of the serially connected bulbs, such an approach may consist in the examination of certain bulbs or rather groups of them (an optimal strategy is that of dividing the circuits into two equal parts).

Note that complex diagnostic systems use a variety of technologies to deal with complexity; these include hierarchical strategies for the identification of faulty subsystem, interactive diagnostic procedures with the use of supplementary tests, and observations aimed at restricting the search area, accessible statistical data in order to establish the most probable diagnoses, and apply heuristic methods in diagnosis. One of the basic and frequently applied heuristics is considering only elementary diagnoses. Another, more advanced one consists in considering only minimal diagnoses.

## 20.2 A Review and Taxonomy of Knowledge Engineering Methods for Diagnosis

### 20.2.1 Knowledge Engineering

Knowledge engineering (KE) methods occupy an important position both in technological system diagnosis and in medical diagnosis [20.4]. They originate from the research in the domain of AI, and, in particular, from those concerning knowledge representation methods and automated inference. These methods are good examples of practical applications of AI techniques. They are mostly based on the algebraic, logical, graphical, and rule-based knowledge representation and automated inference methods [20.2, 28].

A characteristic feature of KE methods is that they use mostly the *symbolic* representation of the domain and expert knowledge as well as *automated inference* paradigms for knowledge processing. They can also make use of numerical data and models (if accessible) as well as uncertain, incomplete, fuzzy, or qualitative knowledge. A common denominator and core for all the methods is constituted by mathematical logic.

The key issue of KE is knowledge representation and knowledge processing; some other typical activities include knowledge acquisition, coding and decoding, analysis, and verification [20.4, 6]. Because of a specific character of KE methods originating mostly from symbolic methods for knowledge manipulation in AI, the taxonomy of such methods is different than those of the methods developed in the automatic control area [20.3] and it constitutes some extension and complement. This extension is oriented toward taking into account specific aspects of KE methods, while the taxonomy takes into account both the applied tools and the *philoso-*

*phy* of specific approaches. In particular, in the case of KE methods, some essential issues of diagnostic approaches are the following ones:

- Source and the way of specification of diagnostic knowledge (model-based approaches vs experience, statistics, and expert systems)
- Applied knowledge representation methods (algebraic, numerical, logical, graphical)
- Applied inference methods (abduction, deduction, search)
- Inference control mechanism (systematic search, heuristic search).

The diagnostic knowledge can, in fact, be of two different origins. First, it can be the so-called *shallow knowledge*, having the source in input–output observations and experience. Such kind of knowledge is also called *expert knowledge* in case it is appropriately significant, and frequently its acquisition consists in interviewing some domain experts. In this case, the knowledge of the system model (the structure, principles of work, mathematical models) is not required. The specification of such kind of knowledge may take the external form of a set of observations of faults and assigned to them diagnoses, learning sequence of examples, or ready-to-use rules coming from an expert. The approaches based on the use of shallow knowledge are generally classified as *expert systems methods*.

Secondly, knowledge may be originating from the analysis of mathematical models (the structure, equations, constraints) of the diagnosed system; this knowl-

edge is referred to as the so-called *deep knowledge*. In case such knowledge is accessible, the diagnostic process can be performed with the use of the model of the system being analyzed, that is, the so-called *model-based diagnosis* is performed. The deep knowledge takes the form of a specific mathematical model (adapted for diagnostic purposes) and perhaps some heuristic or statistical characteristics useful to direct diagnostic reasoning.

Most frequently, the specification of deep knowledge includes definition of the internal structure and dependencies valid for the analyzed system in connection with the set of elements, the faults of which are subject to diagnostic activities, as well as the specification of the current state of the system (observations). The approaches based on the use of deep knowledge are classified as *model-based approaches*. Obviously, a wide spectrum of intermediate cases comprising both of the above approaches in appropriate proportions are also possible.

Knowledge representation methods include mostly symbolic ones, such as facts and inference rules, logic-based methods, trees and graphs, semantic networks, frames, scenarios, and hybrid methods [20.3, 28]. Numerical data (if present) may be represented with the use of vectors, sequences, tables, etc. Mathematical models (e.g., in the form of functional equations, differential equations, constraints) may also be used in modeling and failure diagnosis.

Reasoning methods applied in diagnosis include logical inference (deduction, abduction, consistency-based reasoning, nonmonotonic reasoning, and induction) as well as originating in logical methods for knowledge processing in rule-based systems (forward chaining, backward chaining, bidirectional inference), pattern matching algorithms, search methods, case-based reasoning, and other. In the case of numerical data, various methods of learning systems, both parametric and structural ones, are also applied.

The control of diagnostic inference is mainly aimed at enhancing efficiency, so that all the diagnoses (in the case of complete search) or only the most probable ones (in the case of incomplete search) are generated as fast as possible, so that the obtained diagnoses are ordered from the most likely ones to the most un-

likely ones. The applied methods include blind search, ordered search, heuristic search, use of statistical information and methods, use of qualitative probabilities, use of supplementary tests in order to confirm or reject search alternatives as well as hierarchical strategies.

The following taxonomy of diagnostic approaches is based on the KE point of view, and it takes into account mainly the type and way of diagnostic knowledge specification and the applied methods of knowledge representation.

### 20.2.2 Expert Methods

1. Methods based on the use of numerical data:
  - Pattern recognition methods in feature space
  - Classifiers using the technology of artificial neural networks
  - Simple rule-based classifiers, including fuzzy rule-based systems
  - Hybrid systems.
2. Methods using symbolic data and knowledge (classical knowledge engineering methods, simple algebraic formalisms, graphic- and logic-based methods, and ones based on domain expert knowledge):
  - Diagnostic tests
  - Fault dictionaries
  - Decision trees
  - Decision tables
  - Logic-based methods including rule-based systems and expert systems
  - Case-based systems.

### 20.2.3 Model-Based Methods

1. Consistency-based methods:
  - Consistency-based reasoning using purely logical models (*Reiter's theory* [20.22])
  - Consistency-based reasoning using mathematical, causal models, and qualitative models.
2. Causal methods:
  - Diagnostic graphs and relations
  - Fault trees
  - Causal graphs (CGs)
  - Logical abductive reasoning
  - Logical CGs.

## 20.3 Model-Based Diagnostic Reasoning

*Model-based diagnostic reasoning* appears as a relatively new diagnostic inference paradigm which is based on a formal theory presented in the paper by Reiter [20.22]. The main idea of this paradigm consists in the comparison of the observed system behavior and the one which can be predicted with use of the knowledge about system model. On one hand, such kind of reasoning does not require an expert knowledge, long-term data acquisition or experience, or a training stage of the diagnostic system. On the other hand, what is required is the knowledge about the *system model* allowing for the prediction of its normal correct behavior. More precisely, what is called for is the *model of correct behavior* of the system, that is, a model which can be used to simulate the normal work of the system in the case of lack of any faults.

An idea of such a diagnostic approach is presented in Fig. 20.1.

Both the real system and its model process the same input signals In. The output of the system Out is compared to the expected output Exp generated with the use of the model. The difference of these signals, the so-called residuum  $R$ , is directed to the diagnostic system DIAG. The residuum being equal to zero (with some predefined accuracy) means that the currently observed behavior does not differ from the expected one, that is, the one obtained with the use of the model; if this is the case, it may be assumed that the system works correctly.

In case some significant difference of the current behavior of the system from the one predicted with the use of the model can be observed, then it must be stated that the observed behavior is *inconsistent* with the model.

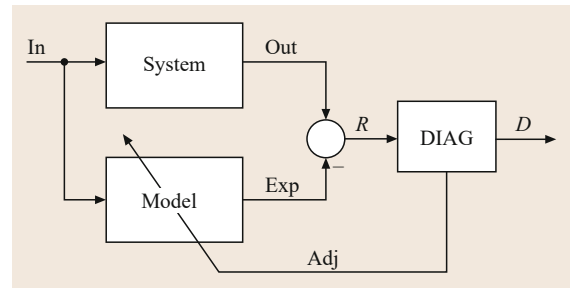


Fig. 20.1 A presentation of diagnostic activity

Detection of such behavior (or *misbehavior*, in fact) implies that a fault occurs (under the assumption that the model is correct and appropriately accurate), that is, *fault detection* takes place.

In order to determine potential diagnoses, an appropriate reasoning allowing for some modifications of the assumptions about the model must be carried out (in the figure, it is shown with the use of the arrow meaning a somewhat specific *tuning* of the model); in case, it is possible to relax the assumption about correct work of the system components in such a way that the predicted behavior would be consistent with the observed one, then the modified model defines which of its elements may have become faulty. In this way, a potential diagnosis  $D^+$  can be obtained (or a set of alternative diagnoses). The diagnoses are represented with sets of system components which (potentially) are faulty, and such that assuming them faulty explains in a satisfactory way the observed misbehavior by regaining the consistency of the observed output with the output of the modified model.

## 20.4 A Motivation Example

In this section, a simple example of a system is introduced. The work of the system is analyzed. It is used to show:

1. The basic concepts, such as system inputs, system outputs, internal variables, system components.
2. System structure and model.
3. How can the model-based diagnostic reasoning be performed?

Let us consider the multiplier–adder system, as introduced in [20.22]. The scheme of the system is presented in Fig. 20.2. The presented system is, in fact, a nontrivial demonstration and benchmark system be-

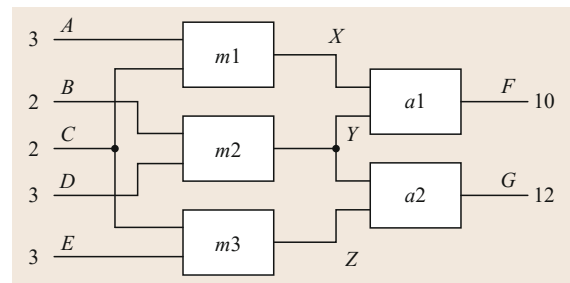


Fig. 20.2 A simple multiplier–adder system

ing a combined multiplier–adder; it is widely explored in the domain literature [20.2, 28]; it is also used for the

illustration of the FDI and DX procedures along with the [20.16, 17] papers.

The system is composed of two layers. The first one contains three independent multipliers  $m1$ ,  $m2$ , and  $m3$ . It receives the input signals  $A$ ,  $B$ ,  $C$ ,  $D$ , and  $E$  and produces values of *internal variables*, namely  $X$ ,  $Y$ , and  $Z$ . The second layer is composed of two adders, namely  $a1$  and  $a2$ . It receives the values of the internal variables and produces the output values, namely  $F$  and  $G$ . Only inputs (of the first layer) and outputs of the system (of the second layer) are directly observable (i. e., they can be measured). The intermediate variables are hidden and cannot be measured.

Note that the system has five *internal components*; this can be written as

$$\text{COMP} = \{m1, m2, m3, a1, a2\}. \quad (20.5)$$

Note that any of the components can be a cause of single failure; hence, the set of potential diagnoses  $D$  considered in Sect 20.1 can be considered as  $D = \text{COMP}$ . The components are interconnected, as shown in Fig. 20.2. The system has also five *inputs*, three *internal variables*, and finally, two *output variables*.

In the following, we shall refer mostly to the classical diagnostic problem as follows. The current state of the system is that the inputs are:  $A = 3$ ,  $B = 2$ ,  $C = 2$ ,  $D = 3$ , and  $E = 3$ . It is easy to check that – if the system works correctly – the outputs should be  $F = 12$  and  $G = 12$ . Since the current value of  $F$  is incorrect, namely  $F = 10$ , the system is faulty. At least one of its components must be faulty.

Now, let us ask, which component (or components) is/are faulty. Note that, the fault of a component (or multiple faults of several components) affects the value of variable  $F$  (it is smaller than expected), but simultaneously does not affect the value of variable  $G$  (12 is the correct value for the observed inputs).

The simplest approach may be based on the observation that the fault must be caused by an error of a component responsible for producing the value of  $F$ . In fact, one can expect a *causal influence* of the following type: faulty component leads to the faulty value. In order to perform the search, a simple *CG* presented in Fig. 20.3 may be useful.

The bottom nodes of the graph correspond to faults of components. The three top-level nodes correspond to the observed values of output variables; in particular, the rightmost node marked with  $F - G$  corresponds to the mutual relationship of signals  $F$  and  $G$ ; in fact, for the observed inputs, the values should be equal, and, since they are not, one can also expect a double

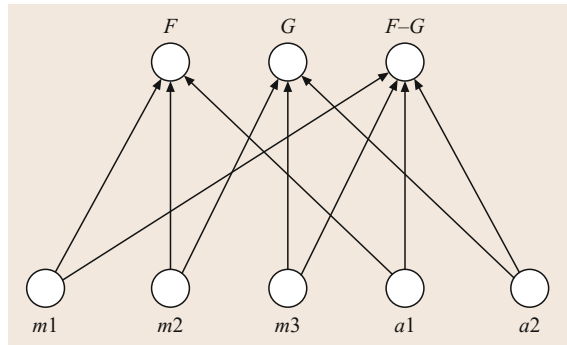


Fig. 20.3 A simple CG for the multiplier–adder system

fault, where one component influences  $F$  (observed to take value lower than expected) and  $G$ , and the other fault improves the value of  $G$  so that it is correct. Such a phenomenon is referred to as the effect of *compensation*.

Let us perform a diagnostic analysis at the intuitive level first. When considering the signal flow inside the system, it can be observed that components of disjunctive conceptual fault (DCF)<sub>1</sub> = { $m1, m2, a1$ } (here, DCF stands for *disjunctive conceptual fault* – a set defining a *conflict*; the definition of these terms will be provided later on) are the only ones influencing signal  $F$ . Similarly, components DCF<sub>2</sub> = { $m2, m3, a2$ } are the only ones influencing signal  $G$ . Finally, components of DCF<sub>3</sub> = { $m1, a1, a2, m3$ } are the ones responsible for the symmetry of signals  $F$  and  $G$ . This is also visualized with the CG presented in Fig. 20.3.

Now, we are at the core of model-based diagnostic reasoning. Since the value of  $F$  is incorrect, at least one of the components of DCF<sub>1</sub> must be faulty. In other words, assumption that all the elements of DCF<sub>1</sub> are correct is inconsistent with the observations. The set is called a *conflict set* [20.22] or a *disjunctive conceptual fault* (DCF) [20.29]. If considering single-element potential diagnoses, all the three elements of DCF<sub>1</sub> are candidate diagnoses. (In fact,  $m2$  is not a valid candidate for a single-element diagnosis; it also influences the value of  $G$ , while no deviation of that value is observed.)

But there is another problem to be explained. The observed values of  $F$  and  $G$  considered together are inconsistent with the model. Note that if all the components were correct, then  $Z = C \times E$  calculated by  $m3$  must be equal to 6, and since  $G$  is observed to be 12,  $Y$  (calculated backward and under the assumption that  $a2$  is correct) must also equal 6; hence, if  $m1$  is correct, then  $X$  must be 6 as well, and if  $a1$  is correct,  $F$  would be equal to 12. Since it is not the case, at least one of the used components must be faulty, that is, DCF<sub>3</sub> is

also a conflict set (a disjunctive conceptual fault) – at least one of its components must be faulty.

Now, in order to *regain consistency* of the observations with the system model, one must remove some of the assumptions that all the components are correct. We look only for *minimal* such explanations, since typically we prefer the simplest possible diagnoses. And there are four such explanations:

- $\{a1\}$  is a single-element candidate diagnosis; it repairs both  $DCF_1$  and  $DCF_3$
- $\{m1\}$  is another single-element candidate diagnosis; it repairs both  $DCF_1$  and  $DCF_3$
- $\{a2, m2\}$  is a two-element candidate diagnosis;  $m2$  repairs  $DCF_1$ , while  $a2$  repairs  $DCF_3$
- $\{m2, m3\}$  is another two-element candidate diagnosis;  $m2$  repairs  $DCF_1$ , while  $a2$  repairs  $DCF_3$ .

## 20.5 Theory of Model-Based Diagnosis

The theory of diagnostic reasoning using the system model and based on the analysis of inconsistency of the observed system behavior with the one predicted with the use of system model was described by Reiter [20.22]. The basic ideas of this theory are presented in brief in the following.

A basic definition in this theory is the definition of a *system*.

### Definition 20.1

A *system* is a pair (SD, COMP) where:

1. SD is a set of first-order predicate calculus formulas defining the system, that is, the *system description*
2. COMP is a set of constants representing distinguished elements of the system (its components).

An example of such a system – the multiplier–adder – has already been presented in this chapter and shown in Fig. 20.2.

The model of the system SD describes its correct behavior. The distinguished elements appear in the model (SD) and they are the only elements which are considered to become faulty – the potential diagnoses will be built from these elements only. In order to do this, the system model is completed with formulas of the form  $\neg AB(c_i)$  which means that component  $c$  works correctly (AB stands for *abnormal behavior*).

The current behavior of the system is assumed to be observed and values of certain variables can be mea-

It is straightforward to observe that in the case of the two-element diagnoses, the *compensation* phenomenon is observed.

The diagnoses – in principle – are obtained as the so-called *hitting sets* for all the conflict sets at hand. Such a hitting set takes one element from each conflict. Hence, the conflicts are removed (repaired), and the collection of taken in such a way components form a candidate diagnosis. In the case of the first two – it was the same element taken from both the conflict sets, so that single-element diagnoses are yielded.

In the following sections, an attempt to present the formal theory underlying *model-based diagnosis* is presented, and an approach to systematic generation of all conflict sets on the base of CG is put forward. Then the multiplier–adder example is revisited, and the formal analysis of it is presented.

The representation of these observations can be formalized in the form of a set of first-order logic formulas; let us denote this set as OBS (OBS stands for *observations*).

Note that an assumption of the form

$$\{\neg AB(c_1), \dots, \neg AB(c_n)\}$$

means, in fact, that each component of the analyzed system works correctly. Hence, a set of the form

$$SD \cup \{\neg AB(c_1), \dots, \neg AB(c_n)\}$$

represents correct behavior of the system, that is, the behavior which can be observed under the assumption that all the components are not faulty. In the case at least one of the components  $c_i \in COMP$  becomes faulty, the set of formulas of the form

$$SD \cup \{\neg AB(c_1), \dots, \neg AB(c_n)\} \cup OBS$$

will become *inconsistent*. The diagnostic process consists in searching for components which may have become faulty, and as such, explain the misbehavior of the system.

For intuition, a diagnosis in Reiter's theory is a hypothesis stating that some set of system elements being a subset of COMP became faulty. Making such an assumption must lead to regaining consistency of the observed system behavior with the one predicted with use of the model. For simplicity, only *minimal diagnoses* will be considered explicitly.



### Definition 20.2

A diagnosis for the system with observations specified by (SD, COMP, OBS) is any minimal set  $\Delta \subseteq \text{COMP}$ , such that the set

$$\text{SD} \cup \text{OBS} \cup \{\text{AB}(c) \mid c \in \Delta\} \cup \{\neg \text{AB}(c) \mid c \in (\text{COMP} - \Delta)\}$$

is consistent.

Roughly speaking, one may say that a diagnosis for some system failure which results with observed misbehavior is any minimal set composed of system components, such that assuming all them to be faulty, and assuming that all the other elements work correctly, is satisfactory for regaining the consistency of the observed behavior with a system model.

Direct search for diagnoses in the form of minimal sets of components sufficient for regaining consistency between the observed and predicted behavior based on the analysis of the set of formulas given by (20.6) would be a tedious task from the computational point of view. In the case of multiple faults, one would have to search for all single-element subsets, then two-element subsets of COMP, etc., and each time such a subset should be verified if it constitutes a diagnosis; some simplifications may consist in the elimination of any superset of a diagnosis found earlier. In Reiter's theory, further improvements are proposed.

The idea of a *conflict set* (or just *conflict* for simplicity; recall also the name *disjunctive conceptual fault* introduced in Sect. 20.4) is of key importance for the theory of consistency-based diagnostic reasoning with use of the system model. A conflict set is any subset of the distinguished system elements, that is, COMP, such that all items belonging to such a set cannot be claimed to work correctly (i. e., at least one of them must be faulty) – it is just the assumption about their correct work which leads to inconsistency.

Assume that we consider a system specified as a pair (SD, COMP), where SD is the theory describing the work of the system (i. e., *system description*) and where  $\text{COMP} = \{c_1, c_2, \dots, c_n\}$  is the set of distinguished system elements. Any of these elements can become faulty, and the output of diagnostic procedure is restricted to be a subset of the elements of COMP.

In diagnostic reasoning, it is assumed that the correct behavior of the system is fully described by the theory of SD. Assumptions about the correct work of system components take the form

$$\neg \text{AB}(c_1) \wedge \neg \text{AB}(c_2) \wedge \dots \wedge \neg \text{AB}(c_n).$$

Hence, assuming that the observed behavior is described with the formulas of the set OBS and in the case

of lack of any faults, the set given by

$$\text{SD} \cup \{\neg \text{AB}(c_1), \dots, \neg \text{AB}(c_n)\} \cup \text{OBS} \quad (20.6)$$

should be consistent. In the case of failure, however, the set of formulas (20.6) turns out to be inconsistent. In order to regain consistency, one should withdraw some of the assumptions about the correct work of system components of the form  $\neg \text{AB}(c_i)$ . Such an approach leads to one or several sets of the form  $\{c^1, c^2, \dots, c^k\} \subseteq \text{COMP}$  of components such that at least one of them must have become faulty. From a logical point of view, the assumptions about such a conflict set are equivalent to stating that the formula

$$\text{AB}(c^1) \vee \text{AB}(c^2) \vee \dots \vee \text{AB}(c^k) \quad (20.7)$$

is true. Obviously, formula (20.7) is true if  $\text{AB}(c^i)$  holds for at least one  $i \in \{1, 2, \dots, k\}$ .

After Reiter [20.22], let us introduce a formal definition of a *conflict set* (*conflict*).

### Definition 20.3

A *conflict set* for (SD, COMP, OBS) is any set  $\{c^1, \dots, c^k\} \subseteq \text{COMP}$ , such that

$$\text{SD} \cup \text{OBS} \cup \{\neg \text{AB}(c^1), \dots, \neg \text{AB}(c^k)\}$$

is inconsistent.

For intuition, a conflict set (under the given observations and system model) is a set of components, such that at least one of its elements must be faulty. Any conflict set represents, in fact, a disjunction of potential faults. A conflict set is *minimal* if any of its proper subsets is not a conflict set. Note that if the analysis is restricted to minimal conflicts, then removing a single element from such a conflict set makes this set become no longer conflict. In other words, the system regains consistency.

Now, let us define an important concept, that is, a *hitting set*.

### Definition 20.4

Let  $C$  be any family of sets. A *hitting set* for  $C$  is any set  $H \subseteq \bigcup_{S \in C} S$ , such that  $H \cap S \neq \emptyset$  for any set  $S \in C$ .

A hitting set is minimal if and only if any of its proper subsets is not a hitting set for  $C$ .

For intuition, a hitting set is any set having a nonempty intersection with any conflict set; it is minimal if removing from it any single element violates the requirement of nonempty intersection with at least one conflict set.

Having defined the idea of a conflict set and a hitting set, we can present the basic theorem of Reiter's theory [20.22]:

**Theorem 20.1**

$\Delta \subseteq \text{COMP}$  is a diagnosis for (SD, COMP, OBS) if and only if  $\Delta$  is a minimal hitting set for the family of conflict sets for (SD, COMP, OBS).

Since any superset of a conflict set for (SD, COMP, OBS) is also a conflict set, it can be shown that  $H$  is a minimal hitting set for (SD, COMP, OBS) if and only if  $H$  is a minimal hitting set for all minimal conflict sets defined for (SD, COMP, OBS). This observation (proved in [20.30]) leads to the following theorem being a fundamental result of Reiter's theory:

**Corollary 20.1**

$\Delta \subseteq \text{COMP}$  is a diagnosis for (SD, COMP, OBS) iff  $\Delta$  is a minimal hitting set for the collection of minimal conflict sets for (SD, COMP, OBS).

To summarize, the role of conflict sets in Reiter's theory is that they provide specifications of components, such that for each conflict set, at least one element must be faulty. By restricting the analysis to

minimal conflicts, one makes sure that by removing any single element from such a set leads to the elimination of conflict. Hence, the union of all such elements (i. e., a hitting set) allows for regaining global consistency; it then constitutes a (potential) diagnosis. Of course, for the considered system failure described with (SD, COMP, OBS), there can exist many diagnoses explaining the observed misbehavior.

In the case of the multiplier–adder system presented in Fig. 20.2, the following conflicts were found

$$\{a1, m1, m2\}, \quad \{a1, a2, m1, m3\}.$$

On the basis of them, all the potential diagnoses can easily be found, that is,

$$D_1 = \{a1\}, \quad D_2 = \{m1\}, \\ D_3 = \{a2, m2\}, \quad D_4 = \{m2, m3\}.$$

Let us notice that only minimal diagnoses are found (i. e., if some set is a diagnosis, then any of its supersets will not be generated as a diagnosis) and that Reiter's theory allows for the generation of both single-element diagnoses (single faults) and multielement ones (multiple faults).

## 20.6 Causal Graphs

In *model-based diagnosis*, modeling of causal relationships plays a very important role. It makes possible pointing to the dependencies of potential faults of the elements of the system under consideration on its observed behavior, which is the result of the faults. The knowledge about causal dependencies allows for efficient diagnostic reasoning based on the direct use of a causal model or some rules generated on the basis of that model.

Let  $d$  denote any fault of some element of the diagnosed system. In the most simple case, it is assumed that the fault can occur or not; from a logical point of view,  $d$  may be considered to denote an *atomic formula* of *propositional logic*. For the purpose of diagnosis,  $d$  will be referred to as an *elementary diagnosis*, and as a logical formula, it will be assigned logical value *true* (if the fault occurs) or *false* (in case the fault is not observed).

Analogously, let  $m$  denote a visible result of some fault  $d$ ;  $m$  can be observed in a direct way or can be detected with the use of appropriate tests or measurements. In the most simple case, manifestation  $m$  may be just observed or not, so as before from a logical point of

view,  $m$  can be considered to denote an atomic formula of the propositional logic which can be assigned a logical value: *true* or *false*. For the purpose of diagnosis,  $m$  will be referred to as a *diagnostic signal* or a *manifestation* or just a symptom of a failure.

If there exists a causal relationship between  $d$  and  $m$  it means that  $d$ , is a cause of  $m$  and  $m$  is an effect of  $d$ . Let  $t_p$  denote the time instant when some symptom  $p$  occurred. For the existence of a causal relationship between symptoms  $d$  and  $m$ , it is necessary that the following conditions are valid:

- $d \models m$ , that is,  $m$  is a logical consequence of  $d$
- $t_d < t_m$ , that is, a cause precedes its result in time
- There exists a flow of a physical signal from symptom  $d$  to symptom  $m$ .

The first condition – the one of logical consequence – means that whenever  $d$  takes the logical value of *true*,  $m$  must also take the logical value *true*. So, this condition means that the existence of causal relationship also requires the existence of logical consequence. (Existence of logical consequence of the form  $d \models m$

does not mean that there exists also causal relationship, for example, the occurrence of  $d$  and  $m$  may be observed as some independent results of some other, external common cause.) This allows for the application of logical inference models for the simulation of systems behavior as well as for reasoning about possible causes of failure.

The second condition, that is, the one of precedence in time means that the cause must occur *before* its result, and that the result occurs *after* the occurrence of its cause. This implies some obvious consequences for modeling of the behavior of dynamic systems in the case of a failure.

The last condition means that there must exist a way for transferring the dependency (a signal channel enabling the flow of the physical signal); lack of such connection indicates that two symptoms are independent, that is, there is no cause–effect relationship among them. A more detailed analysis of theoretical foundations of the causal relationship phenomenon from the point of view of diagnosis can be found in [20.23, 25] or [20.3].

The above presented model of the causal relationship is, in fact, a simplest kind of the so-called *strong causal relationship*; by relaxing the condition of logical implication *potential* relationship can be obtained (in such a case occurrence of  $d$  may, but need not necessarily, mean the occurrence of  $m$ ), including a causal relationship of probabilistic nature (characterized by some quantitative or qualitative probability). For simplicity, such extensions are not considered here. Another extension may consist in the causal relationship between several cause symptoms and several result symptoms described with some functional dependencies; as this case is important for technical diagnosis, it will be considered in brief.

Let  $V$  denote some set of symptoms

$$V = \{v_1, v_2, \dots, v_k\}.$$

The discussion here is restricted to logical symptoms taking the value of *true* or *false*. In some cases, it may be observed that there exists a causal relationship between the symptoms of  $V$  constituting a common cause for some symptom  $v$  and this symptom. In particular, the following two cases are of special interest

$$v_1 \vee v_2 \vee \dots \vee v_k \models v, \quad (20.8)$$

and

$$v_1 \wedge v_2 \wedge \dots \wedge v_k \models v. \quad (20.9)$$

In the first case, occurrence of at least one symptom from  $V$  causes the occurrence of  $v$ ; it is said that there

exists a disjunctive relationship and the symptom  $v$  is of *OR* type. By using an arrow to represent the causal relationship, a dependency of disjunctive type will be denoted as  $v_1 | v_2 | \dots | v_k \longrightarrow v$ .

In the latter case, it is said that the relationship is a conjunctive one – for the occurrence of  $v$ , it is necessary that all the symptoms of  $V$  must occur; it is said that the symptom  $v$  is of *AND* type. The conjunctive relationship is denoted as  $[v_1, v_2, \dots, v_k] \longrightarrow v$ .

Furthermore in some cases, it may happen that the occurrence of some symptom causes another symptom to disappear and vice versa; in such a case, it is said that the relationship is of *NOT* type, that is

$$u \models \bar{v} \quad \text{and} \quad \bar{u} \models v. \quad (20.10)$$

In such a case, the causal relationship is denoted as  $u \succ v$ .

The above-presented causal relationship applies to symptoms having the character of propositional logic variables, that is, to formulas taking the value *true* or *false*. Such symptoms, apart from denoting the occurrence of a discrete event (e.g., tank overflow, signal is on, etc.) may also denote that certain continuous variables take some predefined values or achieve certain levels, that is, de facto they can encode some formulas of the form  $X = w$  or  $X \in W$ , where  $X$  is some process variable and  $w$  is its value, and  $W$  is some set (interval) of values. In such a case, qualitative reasoning and qualitative modeling of the causal relationship at the level of propositional logic only may turn out to be insufficient.

A more general notation for the representation of the causal relationship in case when values of certain variables influence the values taken by other variable may take the following form

$$v_1, v_2, \dots, v_k \longrightarrow_{\psi} v, \quad (20.11)$$

or in the form of an equation

$$\psi(v_1, v_2, \dots, v_k) = v. \quad (20.12)$$

Note that, in this case, it is important that variables  $v_1, v_2, \dots, v_k$  influence variable  $v$ , and the quantitative (or qualitative) characteristics of this influence are expressed with the appropriate equation. In practice, such characteristics can be expressed with a look-up table specifying the values of  $v$  for different combinations of values of the input variables.

Now, let us pass to more general case, that is, modeling the causal relationship among variables taking discrete, continuous, qualitative, or even symbolic values. This will be done with the use of *causal graphs* (CGs).

Consider two system variables, say  $X$  and  $Y$ . Then, if  $X$  influences  $Y$ , we speak about *causal dependency*.

Note that causality assumes at least the following three phenomena:

- Logical implication: a certain change of  $X$  implies a certain change of  $Y$ .
- Directed flow of a physical signal; if  $X$  influences  $Y$ , then there must be some way a physical signal flows from  $X$  to  $Y$ .
- Temporal precedence – the change of  $X$  must happen prior to the change of  $Y$ .

In the following, the definition of a CG is formally introduced.

**Definition 20.5**

A CG is a set of nodes representing system variables and a set of edges or links describing mutual influences among these variables. The edges of the graph are assigned equations describing the influences in a quantitative way, and the variables are assigned some domains.

Let us introduce the following notation:

- $A, B, C, \dots$  – measurable variables of the system
- $[U], [V], [W]$  – immeasurable variables (*internal* or *hidden* ones)
- $X^*$  – conflicting variable, that is, one taking the value *inconsistent* with the *model-based prediction*.

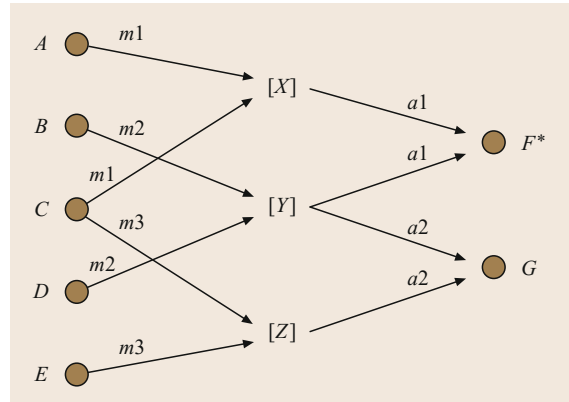
And let ( $\longrightarrow$ ) denote the existence of causal influence between two variables. Any such influence is assigned an expression of the form

$$i = ([X_1, X_2, \dots, X_k], f, Y, [c_1, c_2, \dots, c_k, c_Y]) \tag{20.13}$$

where  $X_1, X_2, \dots, X_k$  are the input variables,  $f$  is a function defining the dependency in quantitative terms,  $Y$  is an output variable and  $c_1, c_2, \dots, c_k$  are the system components responsible for the correct work of the subsystems generating the output values;  $c_Y$  is the component responsible for the value of the output variable  $Y$ .

### 20.7 Potential Conflict Structures

In general case, search for conflicts is not an easy task. In the original work of Reiter [20.22] no efficient method for conflict generation was given. To make things worse, in the general case it is necessary to use an automated theorem prover for proving inconsistency of the set  $SD \cup OBS \cup \{\neg AB(c) \mid c \in COMP\}$  in order to find all the refutations of it;



**Fig. 20.4** A complete CG for the multiplier–adder system

For example, a CG for the previously analyzed arithmetical unit has the structure, as shown in Fig. 20.4. Recall that variable  $F$  took an incorrect value; according to the accepted notation, in Fig. 20.4, it is marked with an asterisk.

The core idea of using the CG for search of conflicts is based on the following observations and assumptions:

- Existence of all the conflicts is indicated by misbehavior of some variables (behavior different from the predicted one).
- In order to state that a conflict exists in the current value of it (observed or measured) must be different from the one predicted with the use of the model.
- The conflict set will be composed of the components responsible for the correct value of the misbehaving variable.

So, in case the CG for the analyzed system is defined, the detection of all the conflicts requires the detection of all the misbehaving variables and next – search of the graph in order to find all the sets of components responsible for the observed misbehavior.

It seems helpful for the discussion to introduce the idea of a *potential conflict structure* (PCS) [20.24, 30].

for any such case the instances of predicates  $AB(\cdot)$  used in refutation should be collected because they form the conflict sets. Of course, the applied theorem prover should be consistent and complete. In conclusion, in the general case the task of finding all minimal conflicts is hard to accomplish and computationally complex.

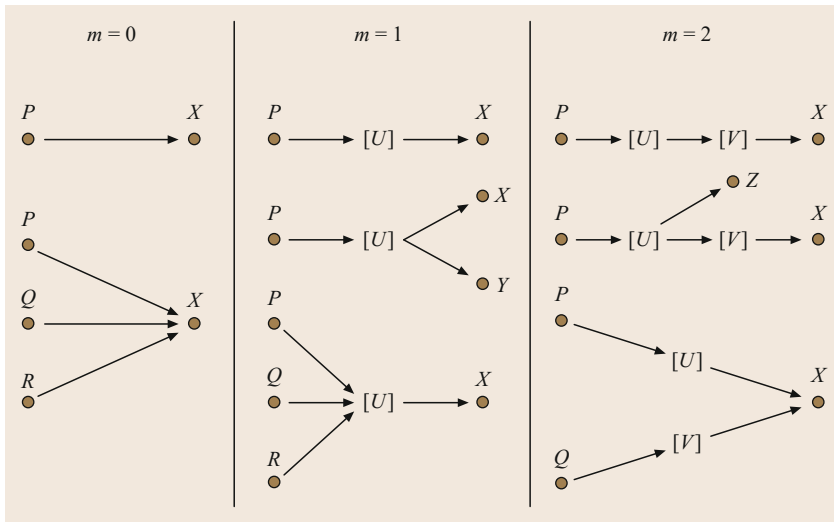


Fig. 20.5 Examples of a simple PCS (after [20.24])

A *potential conflict structure* is a subgraph of the CG sufficient for conflict generation; this idea was first introduced in [20.24]; then it was discussed and developed in [20.26, 27, 30] and [20.29]; an attempt at defining an approach to the automated search of conflict sets for a wide class of dynamic systems which is based on the use of CG representing the flow of signals in the analyzed system was undertaken. The use of such a graph simplifies the procedure of conflict generation and allows for relatively efficient search for all potential minimal conflicts.

Similar concept named *possible conflict* has also been described then in [20.31, 32].

Next, as a result of computational verification of potential conflicts, those which are not real ones are eliminated.

**Definition 20.6**

A PCS structure defined for variable  $X$  on  $m$  hidden variables is any subgraph of the CG, such that:

- It contains exactly  $m$  hidden variables (including  $X$ ).
- The values of all incorporated variables are measured or calculated (they are well defined).
- The value of variable  $X$  is double-defined (e.g., measured and calculated with the use of values of the other variables).
- In the considered PCS, all the values of the  $m$  variables are necessary for  $X$  in order to be double-defined.

A structure just defined allows for potential conflict generation. Some examples of PCS for  $m$  hidden variables are shown in Fig. 20.5.

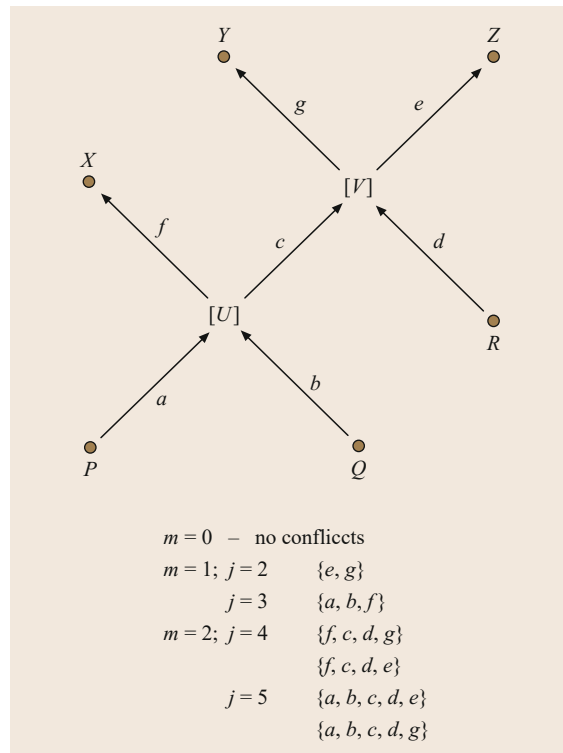
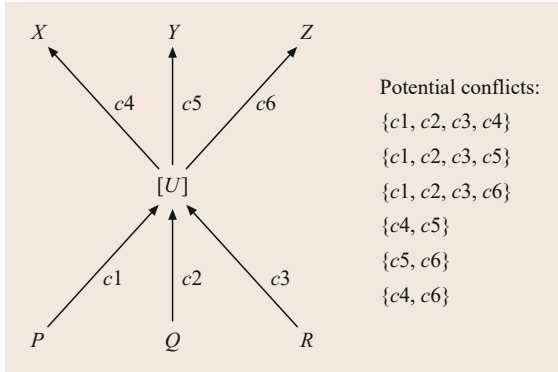


Fig. 20.6 Example conflict structures;  $j$  is the number of links used (after [20.24])

In Fig. 20.6, we show how the number of conflicts and their structure changes with  $m = 0, 1, 2$  for a simple CG of two hidden variables.

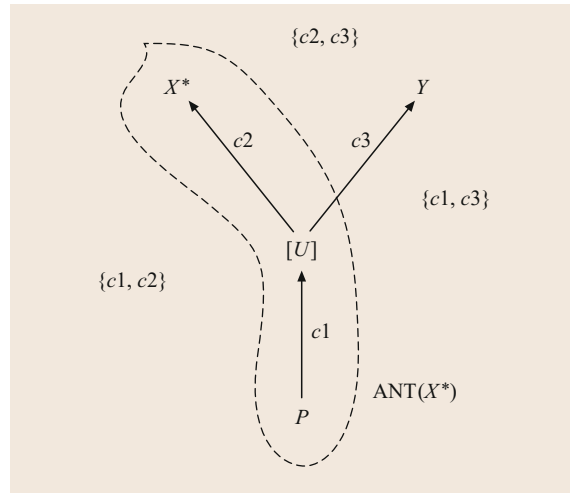
In Fig. 20.7, some further examples for a simple CG with a single immeasurable variable are shown.



**Fig. 20.7** A number of PCSs for a simple CG (after [20.24])

Finally, consider the graph presented in Fig. 20.8.

Since variable  $X$  takes a wrong value, certainly there exists a conflict of the form  $\{c_1, c_2\}$ . But if  $c_1$  were faulty, then  $Y$  should show an incorrect value, and it is not the case. But there is another conflict set  $\{c_2, c_3\}$  – since  $Y$  takes the correct value,  $U$  must be correct, and  $c_2$  must be wrong. Hence, the most obvious diagnosis is  $\{c_2\}$ ; it explains the misbehavior of  $X$  and is consistent with the observation of the correct value at  $Y$ .



**Fig. 20.8** An illustration to the compensation phenomenon (after [20.24])

On the other hand, a *legal* explanation is also diagnosis  $\{c_1, c_3\}$ . Here,  $c_1$  explains the wrong behavior of  $X$  (with  $c_2$  being OK), and faulty component  $c_3$  compensates the error caused by  $c_1$ , so that  $Y$  is observed to have the correct value.

## 20.8 Example Revisited. A Complete Diagnostic Procedure

Let us come back to the multiplier–adder example presented in Sect. 20.4. A detailed analysis of the observed failure is presented in the following. Moreover, an analysis of all potentially observed faults is provided.

Consider once again the multiplier–adder system as presented in Fig. 20.2. Assume that, as before  $F = 10$  and  $G = 12$ , that is, an incorrect output is observed at  $F$ . The current state of the system is that the inputs are:  $A = 3, B = 2, C = 2, D = 3$ , and  $E = 3$ . Having in mind the *model* of the system, it is easy to check that – if the system works correctly – the outputs should be  $F = 12$  and  $G = 12$ . Since the current value of  $F$  is incorrect, namely  $F = 10$ , a fault has been detected. At least one of its components must be faulty.

For simplifying further analysis, consider the CG for the multiplier–adder presented in Fig. 20.2. The graph itself is presented in Fig. 20.4. The CG represents a simplified model of the original system, at the level of detail satisfactory for automated diagnosis.

In order to perform diagnostic reasoning, let us start with *abduction*. We shall try to build a rule specifying hypothetical faulty elements. Note that the value of  $F$  is influenced by the inputs (observed) and the work of elements  $m_1, m_2$ , and  $a_1$ . If all the three elements work

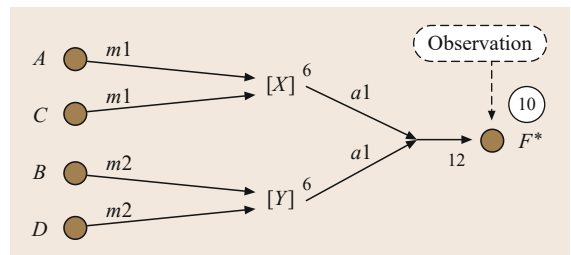
correctly, then the output would be correct. Since it is not, we can conclude that a conflict (or *disjunctive conceptual fault* [20.29])  $DCF_1$  is observed: at least one of the elements  $\{m_1, m_2, a_1\}$  must be faulty. Hence, a rule of the form

$$\text{rule}_{1\_or} : m_1 \vee m_2 \vee a_1 \longrightarrow DCF_1 \quad (20.14)$$

can be stated.

The situation is illustrated in Fig. 20.9.

A further analysis leads to the detection of conflict  $DCF_2$ : under the assumed manifestations, one of the elements  $\{m_1, a_1, a_2, m_3\}$  must also be faulty. This is so



**Fig. 20.9** First conflict detected for the multiplier–adder example

since if all of them were correct, then  $Z = C \times E$  calculated by  $m3$  must be equal to 6, and since  $G$  is observed to be 12,  $Y$  (calculated backward and under the assumption that  $a2$  works correct) must also equal 6; hence, if  $m1$  is correct, then  $X$  must be 6 as well, and if  $a1$  is correct,  $F$  would be equal to 12. Since it is not the case, at least one of the used components must be faulty. So we have the following rule

$$\text{rule}_{2\_or} : m1 \vee m3 \vee a1 \vee a2 \longrightarrow \text{DCF}_2 . \quad (20.15)$$

The situation is illustrated in Fig. 20.10.

Note that if  $F$  would be correct and  $G$  would be faulty, for example,  $F = 12$  and  $G = 10$ , then another observed conflict would be  $\text{DCF}_3 = \{m2, m3, a2\}$  and so we would have a third OR rule of the form

$$\text{rule}_{3\_or} : m2 \vee m3 \vee a2 \longrightarrow \text{DCF}_3 . \quad (20.16)$$

Moreover,  $\text{DCF}_2$  equivalent to a fault in  $\{m1, m3, a1, a2\}$  would occur as well.

If both the outputs were incorrect (e.g.,  $F = 10$  and  $G = 14$ ), then, in general case, one can observe  $\text{DCF}_1$ ,  $\text{DCF}_2$ , and  $\text{DCF}_3$ . Note, however, that whether  $\text{DCF}_2$  is a valid conflict may depend on the observed outputs. For example, if  $F = 10$  and  $G = 10$  (both outputs are incorrect but equal), then the structure and equations describing the work of the system do not lead to a conceptual fault [20.16, 17].

Note that any DCF is modeled with some PCS. Depending on the current manifestations, a DCF can be observed (be active), that is, a real conflict exists or it may be a potential conflict only (be inactive). For effective diagnosis, one needs only the specification of active DCFs.

The diagnoses are calculated as reduced elements of the Cartesian product of the conflict sets associated with the active DCFs. The reduction consists in the elimination of duplicates.

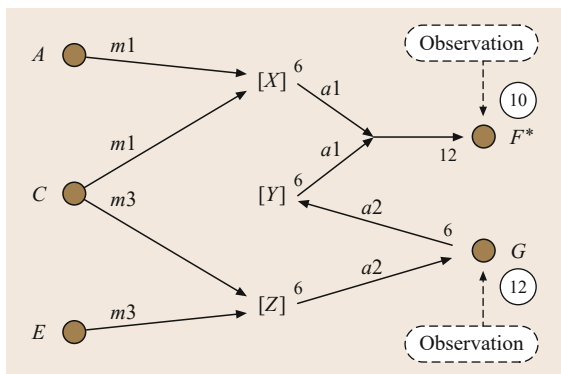


Fig. 20.10 Second conflict detected for the multiplier-adder example

The OR matrix for the diagnosed system is presented in Table 20.1.

The AND matrix defining the relationship between the DCFs (active in the case of  $F$  being incorrect and  $G$  correct) and the manifestations is presented in Table 20.2.

In Table 20.2,  $F^*$ ,  $G^*$ , etc., mean that the output is incorrect, while  $F$ ,  $G$ , etc. denote the correct output observed at the variable.

In the analyzed case, that is  $F$  being faulty and  $G$  correct, the final diagnoses for the considered case are calculated as reduced elements of the Cartesian product of  $\text{DCF}_1 = \{m1, m2, a1\}$  and  $\text{DCF}_2 = \{m1, m3, a1, a2\}$ . There are the following potential diagnoses:  $D_1 = \{m1\}$ ,  $D_2 = \{a1\}$ ,  $D_3 = \{a2, m2\}$ , and  $D_4 = \{m2, m3\}$ . They all are shown in Fig. 20.11.

The potentially possible final diagnoses in general case are presented in Table 20.3.

The calculation of diagnoses can be easily interpreted by using AND/OR CGs [20.25, 28]. An appropriate AND/OR graph is presented in Fig. 20.12. The active links are represented with continuous lines, while the potential ones are represented with dashed lines. Active DCFs are marked with bold circles and the current diagnostic problem (manifestations) are also represented with a bold-line circle.

The final diagnoses are calculated as the minimal sets of the lowest level elements which are necessary to satisfy the currently observed set of manifestations. The intermediate nodes representing the DCFs are OR

Table 20.1 An OR binary diagnostic matrix for the adder system (the lower level)

DCF	m1	m2	m3	a1	a2
DCF <sub>1</sub>	1	1		1	
DCF <sub>2</sub>	1		1	1	1
DCF <sub>3</sub>		1	1		1

Table 20.2 An AND binary diagnostic matrix for the adder system (the upper level)

M	DCF <sub>1</sub>	DCF <sub>2</sub>	DCF <sub>3</sub>
$F^*, G, (F - G)^*$	1	1	
$F, G^*, (F - G)^*$		1	1
$F^*, G^*, (F - G)$	1		1
$F^*, G^*, (F - G)^*$	1	1	1

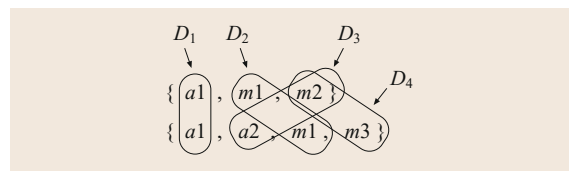


Fig. 20.11 Generation of potential diagnoses

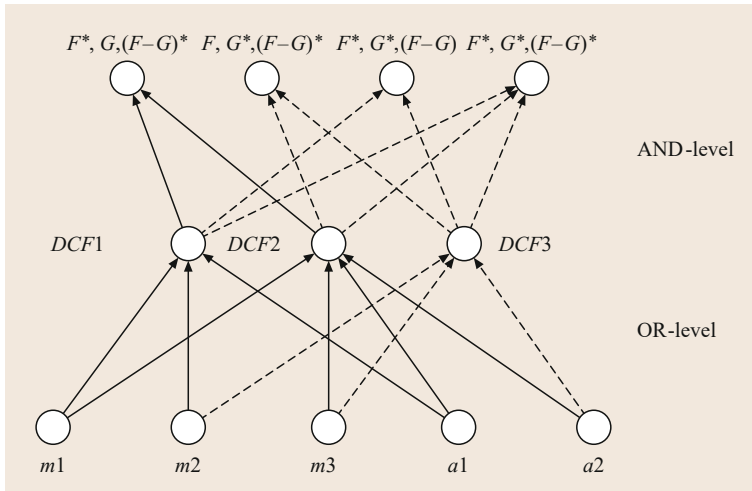


Fig. 20.12 An AND/OR CG for the example multiplier-adder system

Table 20.3 Final possible diagnoses

Observations	Diagnoses
$F^*, G, (F-G)^*$	$\{a1\}, \{m1\},$ $\{a2, m2\}, \{m2, m3\}$
$F, G^*, (F-G)^*$	$\{a2\}, \{m3\},$ $\{a3, m2\}, \{m1, m2\},$
$F^*, G^*, (F-G)$	$\{m2\},$ $\{a1, a2\}, \{a1, m3\},$ $\{a2, m1\}, \{m1, m3\}$
$F^*, G^*, (F-G)^*$	$\{a1, a2\}, \{a1, m2\},$ $\{a1, m3\}, \{a2, m1\},$ $\{a2, m2\}, \{m1, m2\},$ $\{m2, m3\}, \{m1, m3\}$

nodes, while the top-level nodes representing current manifestations are AND nodes.

The presented graphical interpretation can be considered as the effect of *knowledge compilation* for building an efficient diagnostic procedure. In fact, the graph covers all the possible potential failures. In order to build an automated diagnostic system, it would be enough to apply a simple, three-stage procedure:

- Decide which of the top-level nodes describe the current diagnostic situation.
- Find all the real conflicts corresponding to the unique top-level node.
- Generate all the minimal hitting sets.

The resulting diagnoses will be formed by minimal sets of the lowest level nodes, such that a combination of them *supports* the identified conflicts.

## 20.9 Refinement: Qualitative Diagnoses

### 20.9.1 Qualitative Evaluation of Faults

A most popular classification of faults is the binary one. An element can be just faulty ( $f = 1$ ) or not ( $f = 0$ ). This kind of classification is prevailing in technological systems, sometimes extended to several degrees or a fuzzy fault description.

Note, however, that in some particular cases, the fault can be interpreted as a significant deviation from some expected status or value, and the deviation has not only an amplitude but also a direction or sign as well. In this case, the fault can be said to be *negative* or *positive* one, and a classification described with three values  $\{-, 0, +\}$  can be established. This kind of knowl-

edge can be used for further refinement of diagnoses without taking extraordinary measurements, tests, or observations. Furthermore, the same classification can also be assigned to manifestations, that is, values of certain variables can be normal (0), below the norm (-) or above it (+). The presentation in the following is mostly based on [20.29].

### 20.9.2 Elimination of Spurious Diagnoses

The idea is that in numerous cases the influence of faults on the manifestations can be analyzed in a qualitative way using the three-valued approach. Two key observations may be useful: (1) that certain faults can be only



negative or only positive and (2) that the defined sign of deviation of a fault defines also the sign of deviation of the influenced manifestation.

For example, the voltage of a battery can only be normal (0, no fault) or low (−, below normal). The level of liquid in a tank can be normal (0), low (−), or high (+). The clock can be exact, but when faulty it can slow down (−) or advance (+).

The influence of a fault on manifestation can be denoted using the sign. For example, low battery (*battery\_fault*) causes low light (*light\_fault*), that is,  $\text{battery\_fault}(-) \longrightarrow \text{light\_fault}(-)$ .

Let  $V$  denote a set of variables. For diagnostic purposes, we shall assume that  $V = O \cup H \cup C$ , and these sets are pairwise disjoint.  $O$  is the set of observable (measurable) system variables,  $H$  is the set of hidden variables, and  $C$  is the set of diagnostic variables aimed at describing different faulty modes of diagnosed components.

In the case of the multiplier–adder example system, we have

$$\begin{aligned} O &= \{A, B, C, D, E, F, G\}, \\ H &= \{X, Y, Z\}, \text{ and} \\ C &= \{m1, m2, m3, a1, a2\}. \end{aligned}$$

Variables of  $O$  and  $H$  take the values of the real numbers (or integers) restricted to some reasonable intervals, while variables of  $C$  are restricted to some of the possible modes of misbehavior. In our case, the values are restricted to  $\{-, 0, +\}$  with the obvious meaning of lowering the output value, producing correct output and producing the output higher than expected.

For enhancing diagnostic reasoning, however, we shall assume that all the variables can take three qualitative logical values  $\{-, 0, +\}$ . For any variable  $V \in V$ , we can interpret these values as follows:

- $V(0)$  – the proposition that *the value of  $V$  is correct* holds
- $V(+)$  – the proposition that *the value of  $V$  is incorrect; deviation is positive* holds
- $V(-)$  – the proposition that *the value of  $V$  is incorrect; deviation is negative* holds.

In other words, the first statement can be interpreted as a *kind of true*, while the last two statements can be interpreted as some *two different types of negation*.

We shall extend the knowledge about the models of the system over incorrect behavior. In order to do that we shall define some qualitative inference rules. Let  $R$  be a set of rules defining all the accessible knowledge about the behavior of faulty components depending on the faulty mode. Note that, in fact, there are three generic forms of such rules, that is those:

- describing the faulty behavior of elements in the case of normal inputs
- describing the normal behavior in the case of deviated inputs
- describing faulty behavior in the case of deviated inputs.

Let  $c$  denote a single component and  $X$  a variable. By  $c(v)$ , where  $v \in \{-, 0, +\}$ , we shall denote the type of failure; if undefined, we shall write  $c(?)$ . For partial definition, we can use a set of values. The same applies to variables. Now, more detailed characteristics of the diagnoses can be found. Let us introduce a definition of *qualitative diagnosis* taking into account the deviation sign of a fault.

#### Definition 20.7

A *qualitative diagnosis*

$$D = \{d_1(\#), d_2(\#), \dots, d_k(\#)\}$$

is a diagnosis fully explaining the observed misbehavior and covering the knowledge of the deviation sign for any fault (if accessible). Here,  $\#$  is  $+$  if the deviation sign is positive,  $-$  if the deviation sign is negative, and  $?$  if the deviation sign is unknown (any, undetermined).

In the following, three types of causal rules describing the qualitative behavior are discussed in detail.

A generic form of the first type of rules is as follows

$$c(v) \longrightarrow \text{Out}(w)$$

where  $c$  denotes one of the five components of the system, and  $v$  is one of the logical values defining the operating mode,  $v \in \{-, 0, +\}$  and  $\text{Out}$  is its output,  $w \in \{-, 0, +\}$ . For example, a faulty  $m1$  lowering its output signal is described with the rule  $c(-) \longrightarrow \text{Out}(-)$  with the obvious meaning. In the case of the example system, we have as much as 10 such rules (two for each of the five components) defining particular lowering or increasing of the output values when in the faulty state.

A generic form of the second type of rules is as follows

$$\text{In}_1(v_1) \wedge \text{In}_2(v_2) \longrightarrow \text{Out}(w)$$

where  $\text{In}_1$  and  $\text{In}_2$  denote the inputs of a component and  $\text{Out}$  is its output,  $v_1, v_2, w \in \{-, 0, +\}$ .

A generic form of the third type of rules is as follows

$$\text{In}_1(v_1) \wedge \text{In}_2(v_2) \wedge c(v) \longrightarrow \text{Out}(w)$$

where  $In_1$  and  $In_2$  denote the inputs of a component  $c$  and Out is its output,  $v, v_1, v_2, w \in \{-, 0, +\}$ .

The rules of the second type (normal behavior, deviated inputs) are summarized in Table 20.4.

For example, a faulty component  $a1$  showing a lower value of its output signal (but assumed to be correct) and taking one input signal lower and one normal can produce a lower output value. Such a behavior is described with the rule

$$X(-) \wedge Y(0) \longrightarrow F(-)$$

The rules of the third type (abnormal behavior, deviated inputs) are summarized in decision Table 20.5.

For all other 10 combinations of input signals and component mode, the output is undefined. For example, a faulty component  $a2$  increasing its output signal and taking one input signal higher than normal and one normal produces a higher output value. Such a behavior is described with the rule

$$Y(0) \wedge Z(+) \wedge a2(+) \longrightarrow G(+)$$

### 20.9.3 Deduction for Enhanced Diagnoses

The analysis of potential qualitative diagnoses is performed by the propagation of values over the CGs forward, that is, deduction with the rules defined above. This is performed for all qualitative diagnoses. Whenever an inconsistency of the expected and observed values is encountered, the candidate diagnosis is eliminated.

**Table 20.4** Behavior of the correct component with deviated inputs

Inputs	-	0	+
-	-	-	?
0	-	0	+
+	?	+	+

**Table 20.5** Behavior of the incorrect component with deviated inputs

Input1	Input2	Component mode	Output
-	-	-	-
-	0	-	-
0	-	-	-
0	0	-	-
+	+	+	+
+	0	+	+
0	+	+	+
0	0	+	+

### 20.9.4 Analysis of Diagnoses

Let us analyze, in turn, all the four potential diagnoses and their potential qualitative forms. The analysis is aimed at finding all *admissible qualitative diagnoses*.

#### Case of $m1$

There are two false values for  $m1$ , that is,  $m1(-)$  and  $m1(+)$ .

Consider  $m1(-)$  first. Using an appropriate deduction rule of the form

$$m1(-) \longrightarrow X(-),$$

we have  $X(-)$ . Since  $a1$  is correct, but one of its inputs is false (lower), we can use another rule of the form

$$X(-) \wedge Y(0) \longrightarrow F(-)$$

and so we have  $F(-)$ . This is consistent with observations since  $F = 10$ , and the reference value was 12.

Now, consider  $m1(+)$ . Using an appropriate deduction rule of the form

$$m1(+) \longrightarrow X(+),$$

we have  $X(+)$ . Since  $a1$  is correct, but one of its inputs is false (upper), we can use another rule of the form

$$X(+) \wedge Y(0) \longrightarrow F(+)$$

and so we have  $F(+)$ . This is inconsistent with observations since  $F = 10$ , and the reference value is 12. Hence, the diagnosis  $m1(+)$  is inadmissible.

#### Case of $a1$

There are two modes of faulty operation of  $a1$ , that is,  $a1(-)$  and  $a1(+)$ .

Consider  $a1(-)$  first. The appropriate rule is of the form

$$a1(-) \longrightarrow F(-)$$

and so we have  $F(-)$ . This is consistent with observations since  $F = 10$ , which is lower than 12. Finally, diagnosis  $a1(-)$  is admissible.

On the other hand, consider  $a1(+)$ . The appropriate rule is of the form

$$a1(+) \longrightarrow F(+)$$

and so we have  $F(+)$ . This, however, is inconsistent with observations since  $F = 10$ , and the reference value is 12 and should be even higher. Hence, the diagnosis  $a1(+)$  is inadmissible.

**Case of  $\{a2, m2\}$** 

There are four combined potential faulty modes for diagnosis  $\{a2, m2\}$ , that is,  $\{a2(-), m2(-)\}$ ,  $\{a2(-), m2(+)\}$ ,  $\{a2(+), m2(-)\}$ ,  $\{a2(+), m2(+)\}$ . Let us analyze them in turn. For simplicity, we shall only show the applied rules and conclusions.

Case:  $\{a2(-), m2(-)\}$ . Rule to be used

$$m2(-) \longrightarrow Y(-)$$

Conclusion:  $Y(-)$ . Next rule to be applied

$$Y(-) \wedge Z(0) \wedge a2(-) \longrightarrow G(-)$$

Conclusion:  $G(-)$  is inconsistent with observations. Finally, diagnosis  $\{a2(-), m2(-)\}$  is eliminated as inconsistent with observations.

Case:  $\{a2(-), m2(+)\}$ . Rule to be used

$$m2(+) \longrightarrow Y(+)$$

Conclusion:  $Y(+)$ . Next rule to be applied

$$X(0) \wedge Y(+) \longrightarrow F(+)$$

Conclusion:  $F(+)$  is inconsistent with observations. Finally, diagnosis  $\{a2(-), m2(+)\}$  is eliminated as inconsistent with observations.

Case:  $\{a2(+), m2(-)\}$ . Rule to be used

$$m2(-) \longrightarrow Y(-)$$

Conclusion:  $Y(-)$ . Next rule to be applied

$$X(0) \wedge Y(-) \longrightarrow F(-)$$

Conclusion:  $F(-)$  is consistent with observations. We can proceed. Next rule to be applied

$$Y(-) \wedge Z(0) \wedge a2(+) \longrightarrow G(?)$$

Conclusion:  $G(?)$  may be consistent with observations. Finally, diagnosis  $\{a2(+), m2(-)\}$  is a potentially admissible one.

Case:  $\{a2(+), m2(+)\}$ . Rule to be used

$$m2(+) \longrightarrow Y(+)$$

Conclusion:  $Y(+)$ . Next rule to be applied

$$X(0) \wedge Y(+) \longrightarrow F(+)$$

Conclusion:  $F(+)$  is inconsistent with observations. Finally, diagnosis  $\{a2(+), m2(+)\}$  must be rejected.

**Case of  $\{m2, m3\}$** 

There are four combined potential faulty modes for diagnosis  $\{m2, m3\}$ , that is,  $\{m2(-), m3(-)\}$ ,  $\{m2(-), m3(+)\}$ ,  $\{m2(+), m3(-)\}$ ,  $\{m2(+), m3(+)\}$ . Let us analyze them in turn.

Case:  $\{m2(-), m3(-)\}$ . Rule to be used

$$m2(-) \longrightarrow Y(-)$$

Conclusion:  $Y(-)$ . Next rule to be applied

$$m3(-) \longrightarrow Z(-)$$

Conclusion:  $Z(-)$ . Next rule to be applied

$$Y(-) \wedge Z(-) \longrightarrow G(-)$$

Conclusion:  $G(-)$  is inconsistent with observations. Finally, diagnosis  $\{m2(-), m3(-)\}$  is eliminated as inconsistent with observations.

Case:  $\{m2(-), m3(+)\}$ . Rule to be used

$$m2(-) \longrightarrow Y(-)$$

Conclusion:  $Y(-)$ . Next rule to be applied

$$X(0) \wedge Y(-) \longrightarrow F(-)$$

Conclusion:  $F(-)$  is consistent with observations. We can proceed. Next rule to be applied

$$m3(+) \longrightarrow Z(+)$$

Conclusion:  $Z(+)$ . Next rule to be applied

$$Y(-) \wedge Z(+) \longrightarrow G(?)$$

Conclusion:  $G(?)$  may be consistent with observations. Finally, diagnosis  $\{m2(-), m3(+)\}$  may be considered admissible.

Case:  $\{m2(+), m3(-)\}$ . Rule to be used

$$m2(+) \longrightarrow Y(+)$$

Conclusion:  $Y(+)$ . Next rule to be applied

$$X(0) \wedge Y(+) \longrightarrow F(+)$$

Conclusion:  $F(+)$  is inconsistent with observations. Diagnosis  $\{m2(+), m3(-)\}$  must be eliminated.

Case:  $\{m2(+), m3(+)\}$ . Rule to be used

$$m2(+) \longrightarrow Y(+)$$

Conclusion:  $Y(+)$ . Next rule to be applied

$$X(0) \wedge Y(+) \longrightarrow F(+)$$

Conclusion:  $F(+)$  is inconsistent with observations. Diagnosis  $\{m2(+), m3(+)\}$  must be eliminated.

### 20.10 Dynamic Systems Diagnosis: The Three-Tank Case

As another example, let us consider the model of a widely used dynamic system composed of three interconnected tanks [20.3, 27, 30]. The schematic diagram of the system is presented in Fig. 20.13.

The components of the system selected for diagnostic purposes are specified by COMP = {k1, k12, k23, k3, z1, z2, z3}; they are the channels (responsible for flow) and tanks (responsible for the volume of the liquid). The signals to be observed are  $L_1, L_2,$  and  $L_3,$  and they describe the level of the liquid in the consecutive tanks and the signals controlling the input valve  $U$ . The model of the system is specified with the following set of differential equations

$$f(U) = F \tag{20.17}$$

$$A_1 \frac{dL_1}{dt} = F - F_{12} \tag{20.18}$$

$$A_2 \frac{dL_2}{dt} = F_{12} - F_{23} \tag{20.19}$$

$$A_3 \frac{dL_3}{dt} = F_{23} - F_3 \tag{20.20}$$

where  $F_{ij} = \alpha_{ij} C_{ij} \sqrt{2g(L_i - L_j)}, F_3 = \alpha_3 C_3 \sqrt{2gL_3}, A_i$  denote the cross-sectional areas of the tanks for  $i = 1, 2, 3,$  and  $C_{ij}, C_3$  denote the cross-sectional areas of the channels connecting the tanks for  $ij = 12, 23.$  Note that in the case of this system even a superfluous anal-

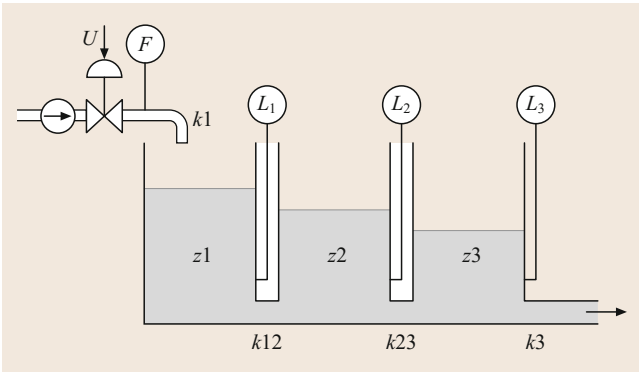


Fig. 20.13 Three-tank system

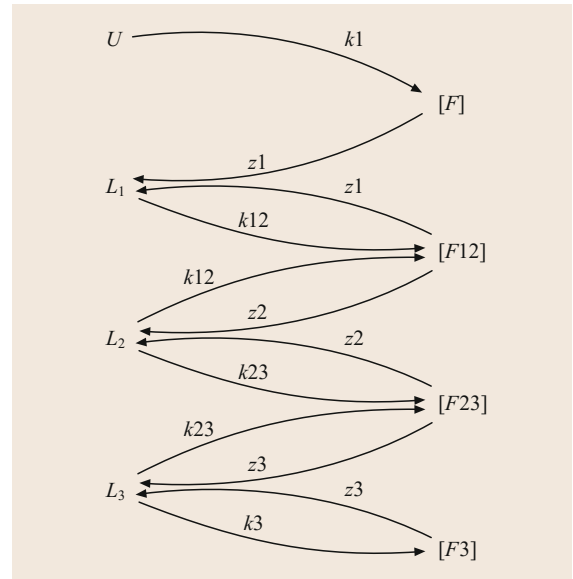


Fig. 20.15 The CG for the three-tank system

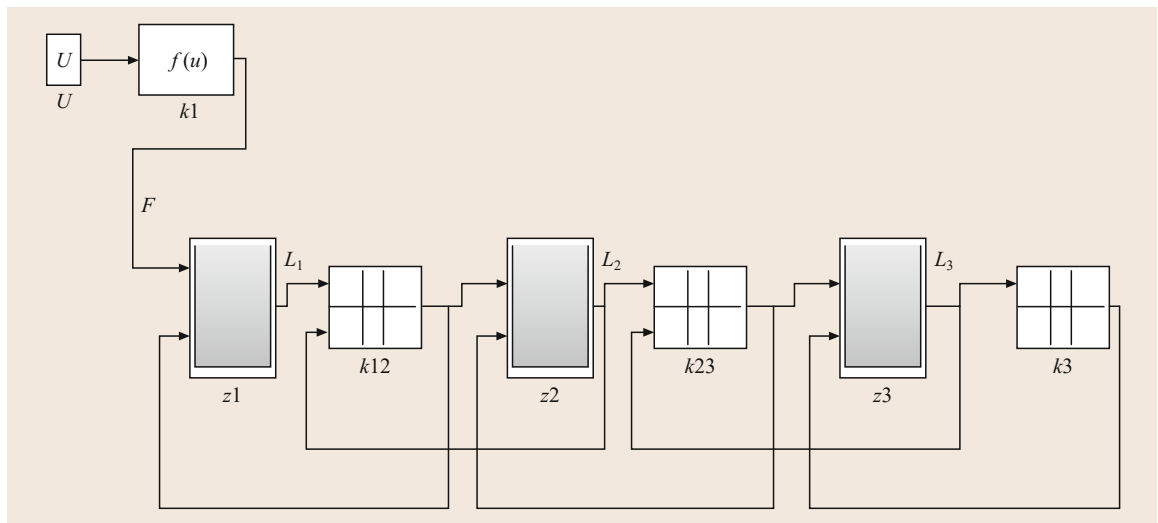


Fig. 20.14 Matlab/Simulink model of the three-tank system

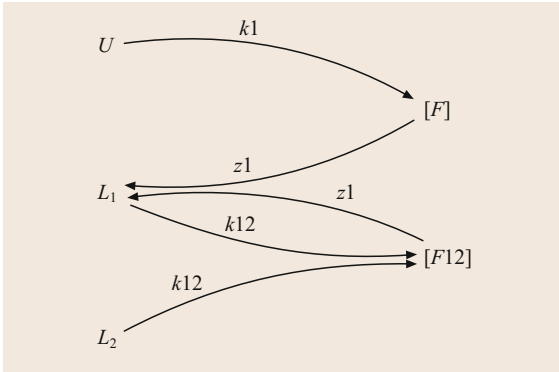


Fig. 20.16 Conflict set  $\{z1, k1, k12\}$

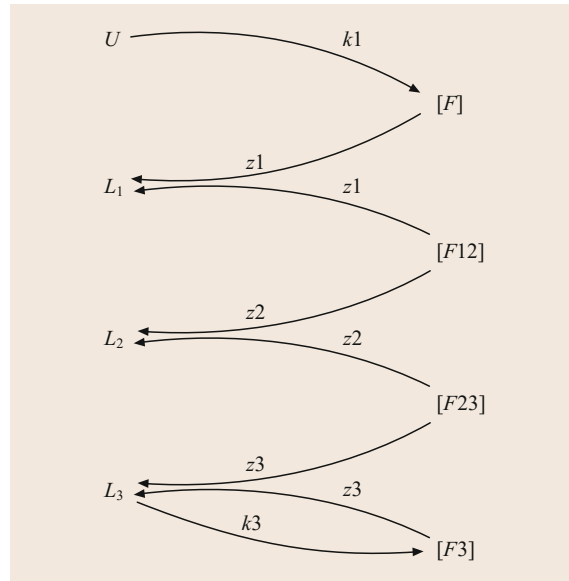


Fig. 20.18 Conflict set  $\{z1, k1, z2, z3, k3\}$

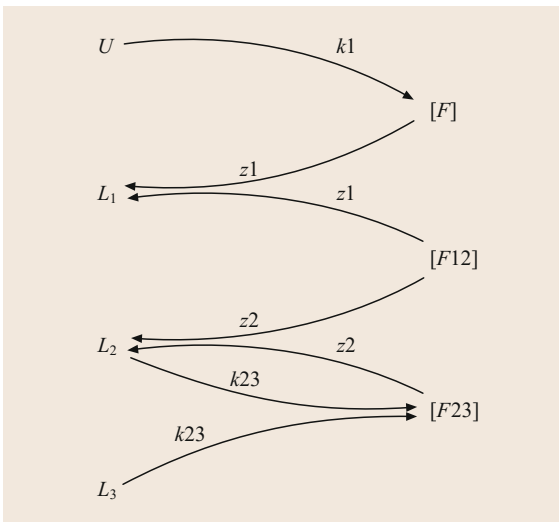


Fig. 20.17 Conflict set  $\{z1, k1, z2, k23\}$

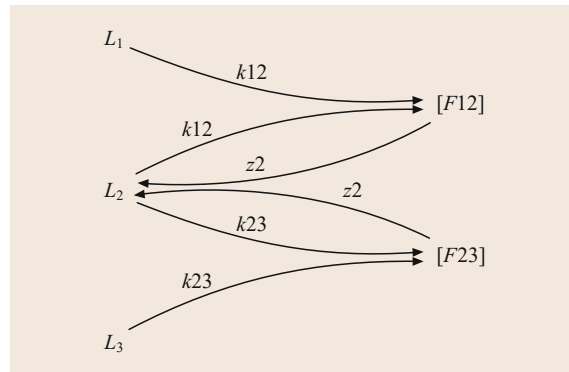


Fig. 20.19 Conflict set  $\{k12, z2, k23\}$

ysis becomes a nontrivial task; this is the consequence of the fact that this time the system under analysis is a highly interconnected dynamic one, it is described with nonlinear equations and there exists strong feedback in the system.

The Matlab/Simulink model of this system is shown in Fig. 20.14. Using Matlab/Simulink, one can simulate the expected correct behavior of the system. If some calculated variables are different from the measured values, an inconsistency is observed and the diagnostic procedure should be activated.

The CG for the example system is shown in Fig. 20.15. The CG can be generated automatically from the Matlab/Simulink model of the system with conflict generator application developed for experimental use. The application is described in more detail in Sect. 20.12.

After defining which of variables are measured ones, the program can generate PCSs. All poten-

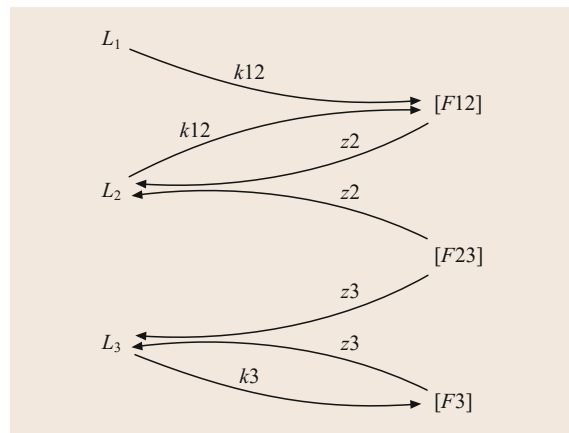


Fig. 20.20 Conflict set  $\{k12, z2, z3, k3\}$

tial conflicts for the three-tank system, that is,  $\{z1, k1, k12\}$ ,  $\{z1, k1, z2, k23\}$ ,  $\{z1, k1, z2, z3, k3\}$ ,  $\{k12, z2, k23\}$ ,  $\{k12, z2, z3, k3\}$ ,  $\{k23, z3, k3\}$  calculated by a conflict generator are shown in Fig. 20.22, and in the graphic way in Figs. 20.16–20.21. Notice that all PCSs can be calculated off-line. As the system is composed of strongly coupled subsystems (there are indirect feedback loops from  $L_2$  to  $L_1$  as well as from  $L_3$  to  $L_2$ ) the generated conflict sets are relatively complex. Assuming that the diagnostic system detected that real conflicts are:  $\{z1, k1, k12\}$ ,  $\{z1, k1, z2, k23\}$ ,  $\{z1, k1, z2, z3, k3\}$  based on Reiter’s theory, the following diagnoses may be calculated:  $\{z1\}$ ,  $\{k1\}$ ,  $\{k12, z2\}$ ,  $\{k12, k23, z3\}$ ,  $\{k12, k23, k3\}$ , so damaged is element  $z1$  or element  $k1$  or at the same time el-

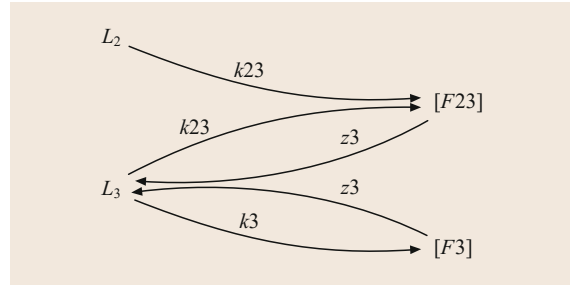


Fig. 20.21 Conflict set  $\{k23, z3, k3\}$

ements  $k12$  and  $z2$  or at the same time elements  $k12, k23, z3$ , or at the same time elements  $k12, k23, k3$  (Fig. 20.24).

### 20.11 Incremental Diagnosis

The algorithm of calculating diagnoses from conflict sets given by Reiter [20.22] requires rather complicated data structures, such as hitting sets (HSs) tree, and is quite difficult to implement. Now, a theorem, that allows us to calculate diagnoses for some of conflicts when there are known diagnoses for those conflict sets will be proposed. This theorem may be used, for instance, in the case when diagnoses are generated simultaneously for some subsystems of one system; in such a case, it can be used for combining together the separately generated sets of diagnoses. It can also be used for incremental generation of diagnoses for one set of conflict sets, and especially when one generates conflicts and diagnoses simultaneously.

Before the theorem is formulated, it is necessary to put forward the following definition.

**Definition 20.8**

Let  $\mathbf{A}$  is set of nonempty sets. The reduced set  $[\mathbf{A}]$  for  $\mathbf{A}$  is the set containing this elements from  $\mathbf{A}$  which are not supersets for other elements.

**Example 20.1**

Let  $\mathbf{A} = \{\{a, b\}, \{a, b, c\}\}$ . We have  $[\mathbf{A}] = \{\{a, b\}\}$ . Let  $\mathbf{B} = \{\{a\}, \{a, b\}, \{a, b, c\}\}$ . Then  $[\mathbf{B}] = \{\{a\}\}$ . Finally, let  $\mathbf{C} = \{\{a\}, \{a, b\}, \{a, b, c\}, \{b, c\}, \{d, e\}\}$ . In this case,  $[\mathbf{C}] = \{\{a\}, \{b, c\}, \{d, e\}\}$ .

Now, a special operator for combining diagnoses will be defined. The operator as its arguments takes the diagnoses for two different families of conflict sets.

**Definition 20.9**

Let  $\mathbf{C}_i$  denote sets of conflict sets,  $\mathbf{D}_i$  sets of diagnoses calculable from  $\mathbf{C}_i$ , and  $\mathbf{H}_i$  sets of all hitting sets for  $\mathbf{C}_i$ ,  $i = 1, 2$ . Let  $D_1 \in \mathbf{D}_1$ ,  $D_2 \in \mathbf{D}_2$ . The operator  $\oplus$  is defined as follows

$$D_1 \oplus D_2 = \begin{cases} \{D_1, D_2\} & D_1 \in \mathbf{H}_2 \text{ and } D_2 \in \mathbf{H}_1 \\ \{D_1\} & D_1 \in \mathbf{H}_2 \text{ and } D_2 \notin \mathbf{H}_1 \\ \{D_2\} & D_1 \notin \mathbf{H}_2 \text{ and } D_2 \in \mathbf{H}_1 \\ \{D_1 \cup D_2\} & D_1 \notin \mathbf{H}_2 \text{ and } D_2 \notin \mathbf{H}_1 \end{cases}$$

Note that the result of operation  $\oplus$  is a family of sets, which may contain one or two sets and each of these sets is a hitting set for  $\mathbf{C}_i$ ,  $i = 1, 2$ .

**Example 20.2**

Let us consider the following sets of conflict sets

$$\begin{aligned} \mathbf{C}_1 &= \{\{a, b, c\}, \{a, d\}\} \\ \mathbf{C}_2 &= \{\{a, c, d\}, \{b, e\}\} . \end{aligned}$$

The sets of diagnoses that can be generated are, respectively,

$$\begin{aligned} \mathbf{D}_1 &= \{\{a\}, \{b, d\}, \{c, d\}\}, \text{ and} \\ \mathbf{D}_2 &= \{\{a, b\}, \{a, e\}, \{b, c\}, \{c, e\}, \{b, d\}, \{d, e\}\} . \end{aligned}$$

We have

$$\begin{aligned} \{b, d\} \oplus \{a, b\} &= \{\{b, d\}, \{a, b\}\} , \\ \{a\} \oplus \{a, b\} &= \{\{a, b\}\} , \\ \{a\} \oplus \{b, c\} &= \{\{a, b, c\}\} . \end{aligned}$$

Now, let us define another operator that constitutes a kind of extension of the previous one.

**Definition 20.10**

Let  $\mathbf{C}_i$  denote sets of conflict sets, and  $\mathbf{D}_1 = \{D_1^1, D_1^2, \dots, D_1^m\}$ ,  $\mathbf{D}_2 = \{D_2^1, D_2^2, \dots, D_2^n\}$  be the sets of diagnoses calculable from  $\mathbf{C}_i$ ,  $i = 1, 2$ . We define operator  $\oplus$  as follows

$$\mathbf{D}_1 \oplus \mathbf{D}_2 = \bigcup_{\substack{i=m, j=n \\ i=1, j=1}} \{D_1^i \oplus D_2^j\}$$

In other words, by using  $\oplus$ , one makes a union of results of operations with  $\oplus$  for each diagnosis from  $\mathbf{D}_1$  with each diagnosis from  $\mathbf{D}_2$ .

Finally, the main theorem of this algebraic approach will be presented. The theorem is named a *composition theorem* since it allows for combining partial results (ones obtained separately or in turn) into the final set of diagnoses (proof in [20.30]):

**Theorem 20.2 (Composition theorem)**

Let  $\mathbf{C}_i$  denote sets of conflict sets and  $\mathbf{D}_i$  sets of diagnoses calculable from  $\mathbf{C}_i$ ,  $i = 1, 2, 3$ . If  $\mathbf{C}_3 = \mathbf{C}_1 \cup \mathbf{C}_2$ , then  $\mathbf{D}_3 = \lfloor \mathbf{D}_1 \oplus \mathbf{D}_2 \rfloor$ .

**Example 20.3**

Let

$$\mathbf{C}_1 = \{\{a, b, c\}, \{a, d\}\} \quad \mathbf{C}_2 = \{\{a, c, d\}, \{b, e\}\}.$$

The sets of diagnoses are

$$\begin{aligned} \mathbf{D}_1 &= \{\{a\}, \{b, d\}, \{c, d\}\}, \text{ and} \\ \mathbf{D}_2 &= \{\{a, b\}, \{a, e\}, \{b, c\}, \{c, e\}, \{b, d\}, \{d, e\}\}. \end{aligned}$$

Consider the combined set of conflict sets

$$\mathbf{C}_3 = \mathbf{C}_1 \cup \mathbf{C}_2 = \{\{a, b, c\}, \{a, d\}, \{a, c, d\}, \{b, e\}\}.$$

The set of diagnoses in this case is as follows

$$\mathbf{D}_3 = \{\{a, b\}, \{a, e\}, \{b, d\}, \{c, d, e\}\}.$$

Now, let us calculate the combination of diagnoses as

$$\begin{aligned} \mathbf{D}_1 \oplus \mathbf{D}_2 &= \{\{a, b\}, \{a, e\}, \{a, b, c\}, \\ &\{a, c, e\}, \{b, d\}, \{a, d, e\}, \{b, c, d\}, \{c, d, e\}\}. \end{aligned}$$

Let us reduce the combined set of diagnoses; then we obtain

$$\lfloor \mathbf{D}_1 \oplus \mathbf{D}_2 \rfloor = \{\{a, b\}, \{a, e\}, \{b, d\}, \{c, d, e\}\}.$$

It is easy to see that  $\mathbf{D}_3 = \lfloor \mathbf{D}_1 \oplus \mathbf{D}_2 \rfloor$ .

The composition theorem allows for calculating diagnoses for the sum of two families of conflict sets in the case where there are known diagnoses for each of these sets of conflicts. One does not need start generation of diagnoses from the beginning, that is, without using the known diagnoses for each individual family of conflicts. The application of this theorem may, therefore, significantly increase the efficiency of a procedure for the calculation of diagnoses.

The composition theorem may be easily generalized to the following theorem (proof in [20.30]):

**Theorem 20.3 (Generalized composition theorem)**

Let  $\mathbf{C}_i$  denote sets of conflict sets and  $\mathbf{D}_i$  sets of diagnoses calculable from  $\mathbf{C}_i$ ,  $i = 1, 2, \dots, n, n + 1$ . If

$$\mathbf{C}_{n+1} = \mathbf{C}_1 \cup \mathbf{C}_2 \cup \dots \cup \mathbf{C}_n,$$

then

$$\mathbf{D}_{n+1} = \lfloor \lfloor \lfloor \mathbf{D}_1 \oplus \mathbf{D}_2 \rfloor \oplus \mathbf{D}_3 \rfloor \oplus \dots \rfloor \oplus \mathbf{D}_n \rfloor.$$

## 20.12 Practical Example and Tools

For modeling dynamical systems, Matlab/Simulink was chosen as a pretty good standard. A diagnostic system was designed and implemented, which uses the presented theory. The system is a consistency-based diagnostic module and may be used to diagnose the class of the systems modeled with CGs. In particular, it needs a model of the diagnosed system made with Matlab/Simulink. The system is composed of two separate modules:

- Conflict generator
- Diagnostic module.

A conflict generator is an application implemented in C/C++. The main window of the system is shown in Fig. 20.22.

The conflict generator performs two main tasks, that is, the generation of CG from the model developed by using the Matlab/Simulink application and generation of all minimal, potential conflicts for such a graph. Con-

flict generation is done by the identification of all PCS in the given structure according to the previous considerations. The potential conflicts generated with the use of the conflict generator are then used by the diagnostic module.

The diagnostic module has been implemented in the Matlab/Simulink environment. It consists of three parts, which are in Matlab/Simulink as masked subsystems:

- A fault detector that has as input the values of measured variables of the diagnosed system and the corresponding values obtained from the model. The task of the fault detection subsystem is to detect inconsistency.
- A conflict verifier that takes potential conflicts as its input. Its task is to determine which of the conflicts are real ones.
- Diagnoses generator that calculates diagnoses in the form of minimal hitting sets for all real conflict sets. Its algorithm is based on Theorem 20.3.

A diagram of the diagnostic module is shown in Fig. 20.23.

Output of the system (main window) is shown in Fig. 20.24. One can see there:

- List of potential conflicts (in case there is fault detected real conflicts are marked)
- List of generated diagnoses
- Graphical visualization of the state of the system.

The graphical interface is quite important part of the application which supports monitoring of the diagnosed system and course of the diagnostic process. It can also be useful in teaching about diagnosis of static and dynamic systems, especially about MBR.

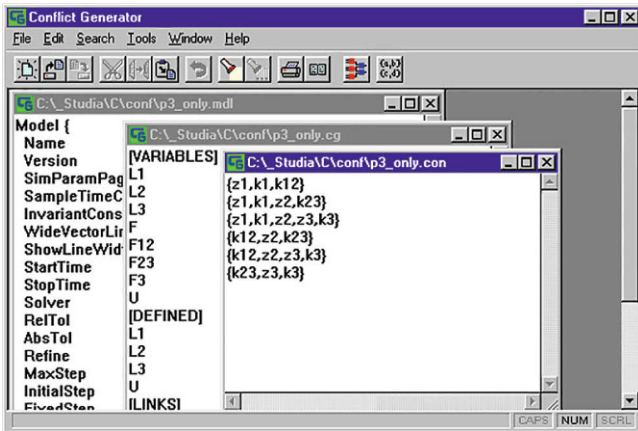


Fig. 20.22 The main window of the program Conflict Generator

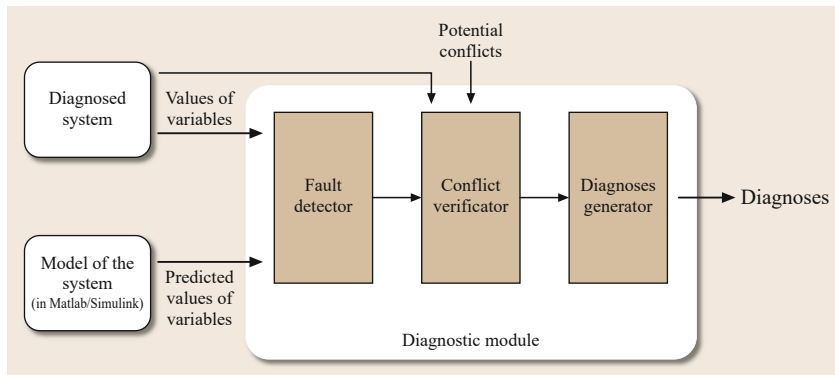


Fig. 20.23 Diagram of the diagnostic module



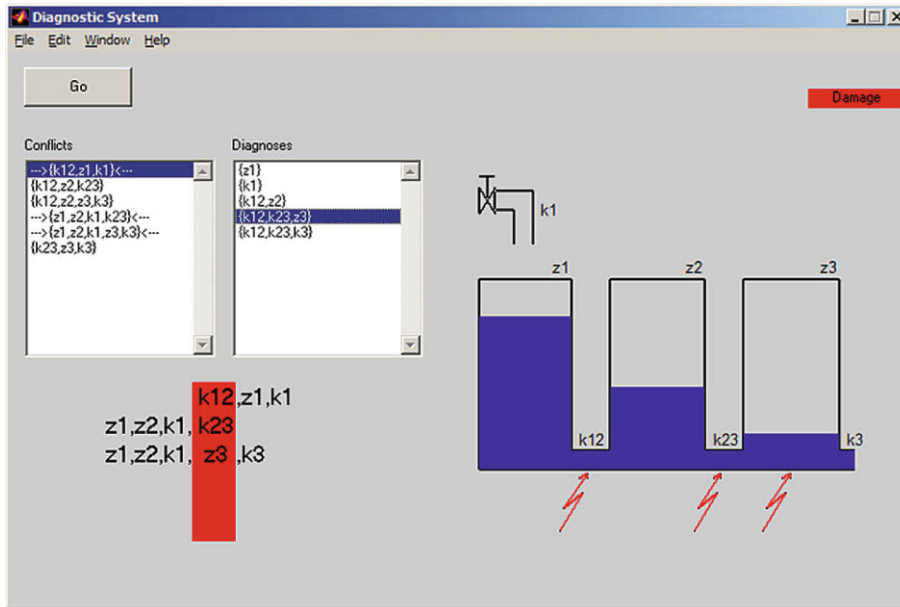


Fig. 20.24 Main window of the diagnostic module

## 20.13 Concluding Remarks

In this chapter, selected methods of knowledge engineering in application for the diagnosis of technical systems were presented in brief. These methods belong to the category of *model-based reasoning* since the analysis is performed with by using the system model. Contrary to the so-called *expert systems*, especially ones using the so-called *shallow knowledge* of expert being the result of experience and useful for diagnosis, and methods based on using model of the system do not require expert knowledge, experience, evidence, etc., but only the model of correct system behavior.

A classical example of the use of model-based diagnosis with the illustrative example of the multiplier–adder system is presented in detail. The methods used are based on *consistency-based reasoning* and *abductive inference*. Both the paradigms use models of the system. A typical example for consistency-based reasoning is Reiter’s theory. Abductive inference may use a model in the form of a CG. A detailed analysis of an example CG by abductive reasoning was shown. The produced results are consistency-based reasonings with the use of a system model.

The main idea for the search of *conflicts* which form *disjunctive conceptual faults* is the one of potential conflict structures; it was shown that PCS can be used to find all DCFs in an efficient way, and even to compile the diagnostic knowledge.

An elaborated analysis of *qualitative diagnoses* was presented in detail. Such an analysis yields more

detailed specification of diagnoses, and, simultaneously, may serve for the elimination of spurious behavior.

An example of the application of the theory to a dynamic system of three tanks was also presented in detail. In the case of dynamic systems, the on-line generation of conflicts may be necessary. As new information becomes accessible, the DCFs can be recalculated, thanks to the provided theorems. A note on practical diagnostic experiments and tools was provided as well.

The presented groups of methods have well-defined theoretical foundations. However, for efficient application, they require adjustment to the specific type of the diagnosed system. Moreover, new approaches to problem statement and new tools may open new diagnostic possibilities; those include embedding the diagnostic process within the framework of constraint programming and compilation of diagnostic knowledge [20.33, 34]. These methods may serve as a core of advanced diagnostic system, but in order to improve efficiency, they should be equipped with specific domain knowledge and heuristic knowledge. They may be also complementary to one another, and it seems reasonable to join expert knowledge based on experience with knowledge about the system model. It should allow for the diagnosis of new, unknown before failures, while the expert component should allow for improving reasoning efficiency.

## References

- 20.1 R. Davis, W. Hamscher: Model-based reasoning: Troubleshooting. In: *Readings in Model-Based Diagnosis*, ed. by W. Hamscher, L. Console, J. deKleer (Morgan Kaufmann, San Mateo 1992) pp. 3–24
- 20.2 W. Hamscher, L. Console, J. de Kleer (Eds.): *Readings in Model-Based Diagnosis* (Morgan Kaufmann, San Mateo 1992)
- 20.3 J. Korbicz, J.M. Kościelny, Z. Kowalczyk, W. Cholewa (Eds.): *Fault Diagnosis. Models, Artificial Intelligence, Applications* (Springer, Berlin 2004)
- 20.4 J. Liebowitz (Ed.): *The Handbook of Applied Expert Systems* (CRC, Boca Raton 1998)
- 20.5 D. Poole: Normality and faults in logic-based diagnosis, Proc. IJCAI-89, Detroit, ed. by N.S. Sridharan (Morgan Kaufmann, San Mateo 1989) pp. 1304–1310
- 20.6 A. Ligeza: *Logical Foundations for Rule-Based Systems* (Springer, Berlin, Heidelberg 2006)
- 20.7 M.R. Genesereth: The use of design descriptions in automated diagnosis, *Artificial Intell.* **24**, 411–436 (1984)
- 20.8 P. Fuster, A. Ligeza, J.A. Martin: Abductive diagnostic procedure based on an and/or/not graph for expected behaviour: Application to a gas turbine, Proc.10th Int. Congr. and Exhib. Cond. Monit. Diagn. Eng. Manag. (COMADEM), ed. by E. Jantunen, K. Holmberg, R.B.K. Rao (Valtion Teknillinen Tutkimuskeskus, Helsinki 1997) pp. 511–520
- 20.9 K. D. Althoff, E. Auriol, R. Barletta, M. Manago: A Review of Industrial Case-Based Tools, AI Intelligence Report (Oxford 1995)
- 20.10 C. Bach, D. Allemang: Case-based reasoning in diagnostic expert systems, *Artificial Intell. Commun.* **9**(2), 49–52 (1996)
- 20.11 I. Watson: *Applying Case-Based Reasoning: Techniques for Enterprise Systems* (Morgan Kaufmann, San Francisco 1997)
- 20.12 P.M. Frank: Fault diagnosis in dynamic systems using analytical and knowledge-based redundancy – A survey and some new results, *Automatica* **26**(3), 459–474 (1990)
- 20.13 P.M. Frank: Analytical and qualitative model-based fault diagnosis – A survey and some new results, *Eur. J. Control.* **2**, 6–28 (1996)
- 20.14 R. Paton, P. Frank, R. Clark: *Fault Diagnosis in Dynamic Systems. Theory and Applications* (Prentice Hall, USA 1989)
- 20.15 S.G. Tzafestas (Ed.): *Knowledge-Based System Diagnosis, Supervision and Control* (Plenum, New York, London 1989)
- 20.16 M.O. Cordier, P. Dague, M. Dumas, F. Lévy, J. Moutmain, M. Staroswiecki, L. Travé-Massuyès: AI and automatic control approaches of model-based diagnosis: Links and underlying hypotheses, Proc. 4th IFAC Symp. Fault Detection, Superv. Saf. Technical Process., ed. by A.M. Edelmayer (IFAC, Budapest 2000) pp. 274–279
- 20.17 M.O. Cordier, P. Dague, M. Dumas, F. Lévy, J. Moutmain, M. Staroswiecki, L. Travé-Massuyès: A comparative analysis of AI and control theory approaches to model-based diagnosis, Proc. 14th Eur. Conf. Artificial Intell. ECAI'2000, ed. by W. Horn (IOS, Berlin 2000) pp. 136–140
- 20.18 L. Travé-Massuyès: Bridges between diagnosis theories from control and AI perspectives. In: *Intelligent Systems in Technical and media Diagnosis*, ed. by J. Korbicz, M. Kowal (Springer, Heidelberg 2014) pp. 3–28
- 20.19 J. de Kleer, A.K. Mackworth, R. Reiter: Characterizing diagnoses and systems, *Artificial Intell.* **56**, 197–222 (1992)
- 20.20 R. Davis: Diagnostic reasoning based on structure and behavior, *Artificial Intell.* **24**, 347–410 (1984)
- 20.21 J. de Kleer, B.C. Williams: Diagnosing multiple faults, *Artificial Intell.* **32**, 97–130 (1987)
- 20.22 R. Reiter: A theory of diagnosis from first principles, *Artificial Intell.* **32**, 57–95 (1987)
- 20.23 P. Fuster-Parra: A Model for Causal Diagnostic Reasoning. Extended Inference Modes and Efficiency Problems, Ph.D. Thesis (Univ. Balearic Islands, Palma de Mallorca 1996)
- 20.24 A. Ligeza: *A Note on Systematic Conflict Generation in CA-EN-Type Causal Structures*, LAAS Report No. 96317 (LAAS, Toulouse 1996)
- 20.25 A. Ligeza, P. Fuster-Parra: AND/OR/NOT causal graphs – A model for diagnostic reasoning, *Appl. Math. Comput. Sci.* **7**(1), 185–203 (1997)
- 20.26 A. Ligeza, B. Górný: Systematic conflict generation in model-based diagnosis, Proc. 4th IFAC Symp. Fault Detection, Superv. Saf. Technical Process., Budapest, ed. by A.M. Edelmayer (IFAC, Budapest 2000) pp. 1103–1108
- 20.27 B. Górný, A. Ligeza: Model-based diagnosis of dynamic systems: Systematic conflict generation. In: *Model-Based Reasoning, Scientific Discovery, Technological Innovations, Values*, ed. by L. Magnani, N.J. Nersessian, C. Pizzi (Kluwer Academic, Dordrecht 2002) pp. 273–291
- 20.28 A. Ligeza: Selected methods of knowledge engineering in system diagnosis. In: *Fault Diagnosis. Models, Artificial Intelligence, Applications*, ed. by J. Korbicz, J.M. Kościelny, Z. Kowalczyk, W. Cholewa (Springer, Berlin 2004) pp. 633–668, Chap.16
- 20.29 A. Ligeza, J.M. Kościelny: A new approach to multiple fault diagnosis. Combination of diagnostic matrices, graphs, algebraic and rule-based models. The case of two-layer models, *Int. J. Appl. Math. Comput. Sci.* **18**(4), 465–476 (2008)
- 20.30 B. Górný: Consistency-Based Reasoning in Model-Based Diagnosis, Ph.D. Thesis (AGH, Kraków 2001)
- 20.31 B. Pulido, C.A. González: An alternative approach to dependency-recording engines in consistency-based diagnosis. In: *Artificial Intelligence: Methodology, Systems, and Applications, Lecture Notes in Artificial Intelligence*, Vol. 1904, ed. by S.A. Cerri, D. Dochev (Springer, Berlin, Heidelberg 2000) pp. 111–121
- 20.32 B. Pulido, C.A. González: Possible conflicts: A compilation technique for consistency-based diagno-

- sis, IEEE Trans. Systems Man and Cybernetics **34**(5), 2192–2206 (2004)
- 20.33 A. Ligęza: A constraint satisfaction framework for diagnostic problems. In: *Diagnosis of Processes and Systems*, ed. by Z. Kowalczyk (Pomeranian Science and Technology, Gdańsk 2009) pp. 255–262
- 20.34 A. Ligęza: Towards knowledge compilation for automated diagnosis: A qualitative, model-based approach with constraint programming. In: *Advanced and Intelligent Computations in Diagnosis and Control*, ed. by Z. Kowalczyk (Springer International, Switzerland 2016) pp. 355–367

## 21. Thought Experiments in Model-Based Reasoning

Margherita Arcangeli

Thought experimentation is at least as old as Western philosophy. Scholars have made much use of it in many disciplines. For instance, philosophical discussions on ethics, morality, knowledge, and language abound with thought experiments. Likewise, great scientific developments, such as in physics and mathematics, have been achieved via thought experimentation. This is true even long before the introduction of the term, between the late seventeenth and nineteenth centuries. But what is a thought experiment? Although giving a clear answer to this question is a very complicated task, it is quite common to consider thought experiments as pieces of reasoning about imaginary cases mainly performed with the aim of increasing our knowledge or understanding of the world. In this chapter, I review the lively debate on thought experiments. First, I introduce some famous examples and detail six of them (Sect. 21.1). Second, I give some historical background (Sect. 21.2). Then, I focus on three of the main questions asked in the literature, namely: What is a thought experiment? (Sect. 21.3), What is the function of thought experiments? (Sect. 21.4), How do thought experiments achieve their function? (Sect. 21.5). These issues will lead to tackle other important points, such as the relationship between real and thought experimentation, the differences between philosophical and scientific thought experimentation, the role played by intuitions and imagination in thought experimentation.

21.1	<b>Overview</b> .....	464
21.1.1	Galileo on Falling Bodies.....	464
21.1.2	Stevin's Chain Thought Experiment .....	465
21.1.3	Newton's Bucket Thought Experiment .....	465
21.1.4	Gettier's Thought Experiment .....	466
21.1.5	Twin Earth .....	466
21.1.6	Mary the Super-Scientist.....	467
21.2	<b>Historical Background</b> .....	467
21.2.1	The Rise of the Term .....	467
21.2.2	The Classical Phase .....	468
21.2.3	The Contemporary Phase .....	469
21.3	<b>What Is a Thought Experiment?</b> .....	469
21.3.1	Thought Experiments and the Experimental Realm .....	470
21.3.2	Thought Experiments and the Theoretical Realm.....	472
21.3.3	Thought Experiments and Their Features.....	473
21.4	<b>What Is the Function of Thought Experiments?</b> .....	475
21.4.1	Sorting Thought Experiments.....	476
21.4.2	Thought Experiments and Kinds of Knowledge .....	478
21.4.3	The Epistemological Status of Thought Experiments .....	480
21.5	<b>How Do Thought Experiments Achieve Their Function?</b> .....	484
21.5.1	A Cognitive Approach to Thought Experimentation.....	484
21.5.2	Imagination and Thought Experimentation .....	485
21.5.3	The Narrative Dimension of Thought Experimentation.....	486
	<b>References</b> .....	487

## 21.1 Overview

Thought experiments (TEs) have a rather curious history. Although thought experimentation is as old as philosophy, both the introduction of the term and the philosophical interest in thought experiments as such have a much more recent history. Since the beginning of Western philosophy, many famous thought experiments have been proposed and discussed, including *Plato's* ring of Gyges [21.1, 358a–360d], *Hilary Putnam's* brain in the vat [21.2], *John Locke's* inverted spectrum [21.3, II, Ch. 32, §15], *Galileo Galilei's* thought experiment on free-fall [21.4], *Étienne Bonnot de Condillac's* statue [21.5], *Immanuel Kant's* on handedness [21.6], *Charles Darwin's* thought experiment on the evolution of the eye [21.7], *Henri Poincaré's* disk world [21.8], *Albert Einstein's* lift [21.9], *Werner Heisenberg's*  $\gamma$ -ray microscope [21.10], *Tyler Burge's* arthritis [21.11], and *John Searle's* Chinese room [21.12]. These examples are only a sample of a vast production that spans a huge amount of time and issues (for other surveys and in-depth analyses see contributions in *Horowitz and Massey* [21.13]; *Casati et al.* [21.14]; *Ierodiakonou and Roux* [21.15]; *Frappier et al.* [21.16], and in the special issues of *Philosophica* – 72, 2003 – and that of *Perspectives on Science* – 2/22, 2014). But what is a thought experiment? After a brief historical introduction (Sect. 21.2), this chapter focuses on philosophical answers to this question (Sect. 21.3) and on other issues concerning thought experimentation, namely on the issues about the function of thought experiments (Sect. 21.4) and how thought experiments achieve their function (Sect. 21.5). Before turning to these questions, it is worth having a precise idea of what is under discussion by looking at some concrete examples. The remainder of the section will be devoted to the introduction of six thought experiments that can be seen as a good sample of the most quoted and discussed thought experiments in the literature.

### 21.1.1 Galileo on Falling Bodies

Galileo was a great thought experimenter. Via his thought experiments he forcefully undermined Aristotelian physics. One among them is widely quoted in the literature, namely the thought experiment against the Aristotelian idea that the speed of a body's free fall increases proportionally to its weight. As Galileo put it in his *Discorsi e dimostrazioni matematiche intorno a due nuove scienze* (*Discourses and Mathematical Demonstrations Relating to Two New Sciences*), the supposition that [21.4, p. 62]

“bodies differing in heaviness [*gravità*] are moved in the same medium with unequal speeds, which maintain to one another the same ratio as their weights [*gravità*].”

Galileo thought that this supposition was false and maintained that a heavy cannon ball of 100 or more pounds will not anticipate a half-pound musket ball both dropped from a height of 200 *arms*. In order to make his point, Galileo put forward a thought experiment and observed that the Aristotelian idea leads to a contradiction. His thought experiment runs as follows (Fig. 21.1).

Imagine two bodies (e.g., stones) that have unequal weights, and so speeds (e.g., the heavy stone falls with a rate of 8 and the light stone with a rate of 4). Suppose that these bodies are linked together (e.g., with a weightless chain) and, then, that one drops them from a certain height (e.g., the top of the Tower of Pisa). The Aristotelian thesis would entail that the velocity of the composite body will have: a) an intermediate value between the two, since the lighter body delays the heavier, and b) a higher value than the two, since both bodies are lighter than their union.

Galileo's conclusion is that large and small bodies fall with the same speed [21.4, p. 65]. He stresses that the slight differences we experience are due to external factors, such as the air resistance. Arguably, “in the vacuum their velocities would be completely identical” [21.4, p. 73].

Interestingly, it turns out that this hypothesis holds also for bodies made of different materials, for example, a hammer and a feather, as the well-known recreation of the experiment by Apollo 15 astronaut David Scott

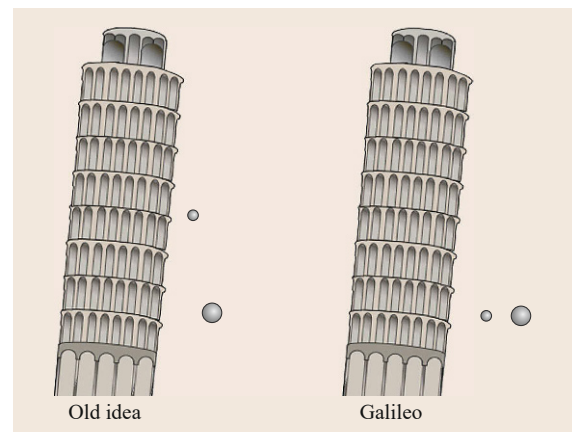


Fig. 21.1 Two bodies falling from the Tower of Pisa

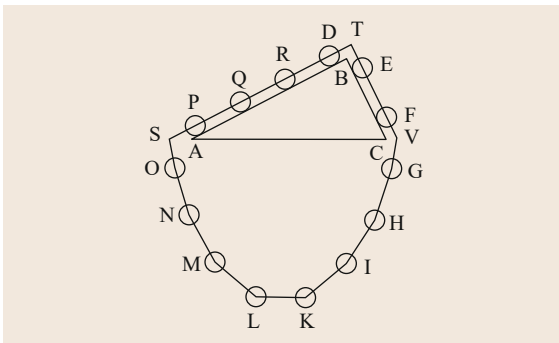
on the Moon showed. An Italian team of scientists has since conducted an atomistic version of Galileo's experiment, in which atoms of different weights fall in the vacuum at the same speed [21.17]. How could Galileo have seen so far using only his imagination?

### 21.1.2 Stevin's Chain Thought Experiment

In *De statica*, the fourth tome of Simon Stevin's *Hypomnemata mathematica*, Stevin was dealing with the force needed to keep an object on an inclined plane from sliding down, and he concluded that the force required is inversely proportional to the length of the plane. In order to demonstrate his result, he proposed the following thought experiment.

Imagine a triangular prism ABC (depicted in section in Fig. 21.2) whose basis (AC) is horizontal and the left side (AB) is twice the length of the right side (BC). A wreath of 14 balls of equal weight and size is draped over the prism, so that four balls are on AB (D, R, Q, P), two on BC (E, F), and the remaining eight beneath the base (G, H, I, K, L, M, N, O). The spheres are linked by a thread passing through their centers, so that they can move, but they must remain equally spaced. Moreover, the sides are frictionless, and S, T, and V are fixed points on which the thread can slide freely.

Stevin claimed that if D, R, Q, and P were not balanced by E and F, one of the two groups of spheres would pull the other. What would happen if this was so? Let us suppose that the block with eight balls (D, R, Q, P, O, N, M, and L) generates more force than the one with six balls (E, F, K, I, H, and G) and D, R, Q, and P slide down the left side. Thus, D will go down where O is and E, F, G, and H will take the place of P, Q, R, and D, while I and K, will take the place of E and



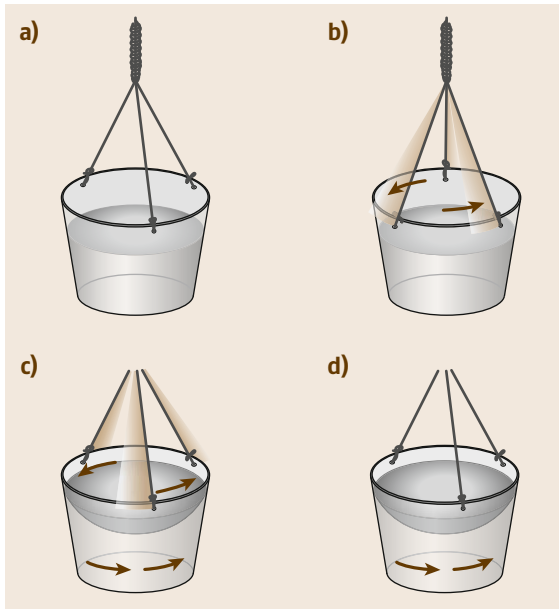
**Fig. 21.2** Stevin's chain: ABC is the section of the prism, whose basis (AC) is horizontal and the left side (AB) is twice the length of the right side (BC). Fourteen balls (D–R) compose the chain draped over the prism. S, T and V are fixed points on which the chain can slide freely

F. Nevertheless, as Stevin remarked, we are now facing the same setup of the beginning and for the same reason the four balls on the left side will slide down the left side and be replaced by other four balls, and so on and so forth. “These spheres will produce by themselves a continuum and eternal motion [*continuum et aeternum motum*], which is false” [21.18, p. 35]. Given the fact that a perpetual motion is absurd, we are led to conclude that the block with eight spheres and the one with six balance each other. Moreover, since the spheres beneath the basis are symmetrically arranged and pull equally in both directions, we can imagine cutting the string at the two lower corners without disturbing the equilibrium. Hence, two spheres (E, F) offset four spheres (D, R, Q, and P). From that Stevin concluded that the force required to keep an object from sliding down on an inclined plane varies inversely with the length of the plane and derived his law of the inclined plane: The ratio of the force to the weight is equal to the ratio of the height to the length of the plane.

### 21.1.3 Newton's Bucket Thought Experiment

When we are sitting in a train and realize that the train on the next platform changes its spatial relationship with our train, without any other external cue we are not immediately able to say which train is *really* moving. However, when a motion changes in speed or direction (i. e., accelerates), it is directly detectable, apparently without reference to any other object. Are accelerated motions a special type of motion or even in such cases is there an implicit contrast between relative and absolute motion? In his *Principia Mathematica* Isaac Newton [21.19] put forward a thought experiment in order to ascertain whether absolute and relative motions differ as regards their effects in the context of accelerated motions, more precisely whether only the former would involve centrifugal force. Let me explain step by step the thought experiment.

Imagine a bucket containing water, which hangs on a rope that is twisted as much as possible, and then released. We can distinguish four phases (Fig. 21.3): (a) both the water and the bucket are stationary, (b) only the bucket begins to move, (c) the water and the bucket are both moving, (d) the bucket stops, but the water keeps moving. According to Newton, the existence of relative motion between the water contained in the bucket and the bucket itself is not able to explain the changes in the surface shape of the water, when the water is rotating. Indeed, in both the first and the third phases, the water and the bucket are stationary relative to one another. Yet while in (a) the water's surface is flat, in (c) it is concave. Likewise, in both the second and the



**Fig. 21.3** Newton's bucket experiment: (a) both the water and the bucket are stationary, (b) only the bucket begins to move, (c) the water and the bucket are both moving, (d) the bucket stops, but the water keeps moving

fourth phases, the bucket is in motion from the point of view of the water and vice versa, but the water's surface is flat in (b) and concave in (d). Moreover, Newton suggests that the same would hold even in infinite empty space, where there is no object external to the device. This last remark gives to the thought experiment its unperformable flavor. In fact, although in the literature the bucket experiment has been mostly discussed as a thought experiment, Newton affirmed to have done it (clearly not in the empty space!).

According to the standard interpretation of the bucket experiment (be it a real or a thought experiment), the Newtonian theory could explain the phenomenon: Absolute space is what discriminates between absolute and relative motions, that is, between real and illusory motions. Absolute space would be the system with respect to which it is possible to understand that in the first and in the second phase the water surface is flat, because the water does not *really* move, whereas it does in the third and the fourth phases, thus climbing the bucket's wall.

#### 21.1.4 Gettier's Thought Experiment

In his article, *Is Justified True Belief Knowledge?*, Edmund Gettier [21.20] seriously undermined the classic definition of knowledge via thought experimentation. Almost from Plato to 1963, knowledge was considered

true and justified belief. A subject (S) knows the content of a proposition ( $p$ ) if and only if: (a)  $p$  is true, (b) S believes that  $p$ , and (c) S is justified in believing that  $p$ . Via his thought experiments Gettier pointed out how a justification may be flawed, thus implying that we can have true and justified beliefs that seem not to be states of knowledge (it should be noted that Bertrand Russell, in his *Human Knowledge: Its Scope and Its Limits*, anticipated this kind of example; see [21.21, pp. 170–171]). Many attempts have been made in order to account for Gettier's insight. Although none of these has been widely accepted, the overall debate has helped to clarify the concepts of knowledge and justification. Here is one of Gettier's original examples: Suppose that Smith has applied for a job, as well as Jones. Actually, the person in charge of hiring tells Smith that Jones will get the job. Moreover, Smith has just counted the number of coins in Jones' pocket, which is ten. Having such strong evidence, Smith believes that Jones is the man who will get the job and Jones has 10 coins in his pocket. Suppose that Smith infers from his belief another belief, namely the belief that the man who will get the job has 10 coins in his pocket. The same evidence grounds both beliefs.

But imagine that Smith gets the job, not Jones, and that by sheer chance, Smith unknowingly has 10 coins in his pocket. Although his belief that the man who will get the job has 10 coins in his pocket is both justified and true, Smith does not appear to know that the man who will get the job has 10 coins in his pocket.

After Gettier's original examples, a variety of cases à la Gettier have been proposed. A quite popular Gettier case is worth mentioning, namely the cow in the field problem [21.22]. Briefly, this example sets up an imaginary scenario in which a farmer is worried about his cow Daisy and gets relaxed when he sees her black and white shape in the field. It happens that Daisy was safely in the field, but hidden in a hollow. What the farmer mistook for his cow was a large sheet of black and white paper. Like Smith, the farmer has a justified, true belief which seems not straightforwardly to count as knowledge.

#### 21.1.5 Twin Earth

Among philosophers, a famous thought experiment is that of Twin Earth. Putnam [21.23, 24] proposed a thought experiment in order to show that psychological states conceived in a narrow (i. e., intensional) sense do not determine the references (i. e., extensions) of natural kind terms.

Suppose that somewhere in the Universe there is Twin Earth, which is a planet exactly like Earth, with

the exception that the chemical composition of what is called *water* on Twin Earth is not  $H_2O$ , but a very long and complicated formula that can be abbreviated as *XYZ*. Nevertheless, water and twin water (i. e., what is called *water* on Twin Earth) have the same visual appearance, flavor, odor, etc. Thus, if Earthlings ever visit Twin Earth, at first they will believe that the term *water* has the same meaning on both planets. But they will suitably revise their belief after discovering that *water* on Twin Earth refers to *XYZ*. And it would be the same for Twin Earthlings, if they ever visited Earth and discovered that *water* on Earth refers to  $H_2O$ .

Oscar-te is the Doppelgänger on Twin Earth of Oscar, a typical inhabitant of Earth, and we can suppose that both persons are perfect duplicates. Going back to 1750 (about 50 years before the discovery of the chemical composition of water on Earth and, by hypothesis, also of what is called *water* on Twin Earth), neither Oscar, nor Oscar-te had beliefs about the chemical elements of what they call *water*. Yet the term *water* referred to  $H_2O$  on Earth and to *XYZ* on Twin Earth, respectively, in 1750 as much as nowadays (i. e., the extension of the term did not change). Although by hypothesis Oscar and Oscar-te had all the same beliefs, and enjoyed the exact same psychological states with respect to the word *water*, they did not mean the same thing by *water*, because each use referred to a different substance. Therefore, Putnam argues that meanings of words are not in our heads.

## 21.2 Historical Background

Schematically, it is possible to divide the long debate on thought experiments into two phases, which we could call *classical* and *contemporary* (see [21.27] for a more fine-grained division in four stages). Thomas S. Kuhn, in his 1964 paper, *A Function for Thought Experiments*, marks the division between the two phases. Kuhn has the merit of having highlighted an epistemological problem hitherto underestimated. His essay opened a new way of viewing thought experiments as puzzles, by asking: How might a mere thought experiment yield new knowledge, without the input of new data? It should be noted that Kuhn's paper is mostly known in its 1977 version. This is the reason why some of his contemporaries (e.g., Carl Hempel) can be considered as belonging to the classical phase, even if their writings are posterior to Kuhn's essay.

The contemporary phase can, therefore, be separated quite sharply from the classical phase along two criteria:

1. A greater awareness of the problematic aspect of thought experiments tied to their epistemic function

### 21.1.6 Mary the Super-Scientist

Frank Jackson [21.25] proposed a famous thought experiment (also known as the knowledge argument) aimed at showing that physicalism (i. e., the view that everything can be physically explained) cannot account for our knowledge about what it feels like to be in a certain mental state. The thought experiment runs as follows.

Suppose that Mary is a brilliant neuroscientist who knows all physical facts about chromatic vision. For instance, she knows precisely which combinations of wavelengths from a tomato stimulate the retina, and “exactly how this produces via the central nervous system the contraction of the vocal chords and expulsion of air from the lungs” [21.25, p. 130], which results in the utterance of the judgment *the tomato is red*. Although she possesses all physical information concerning what happens when, for example, we see the redness of a tomato and use terms like *red*, she has never had the experience of seeing any color, because she has spent all her life in a black and white room. Now imagine that Mary could see a tomato, because either she is released from her room or provided with a color television monitor. Would she learn something or not? The intuitive answer seems to be *yes*. But does she really learn a new fact? These questions have sparked a lively debate [21.26], which led Jackson himself to revise his position and claim that, after all, Mary is not discovering a new fact, but rather a new way to represent it.

2. The extension of the analysis to thought experiments outside the realm of science (mainly physics), in particular to philosophical thought experimentation.

In this section, I shall review the most important steps of the history of thought experimentation. I begin with the origin of the term (Sect. 21.2.1). Then I briefly describe the two historical phases and finally introduce the main actors of each phase (Sect. 21.2.2 and Sect. 21.2.3, respectively).

#### 21.2.1 The Rise of the Term

At the end of the nineteenth century, the Austrian physicist and philosopher Ernst Mach wrote a paper entitled *Über Gedankenexperimente* [21.28], which popularized the German term *Gedankenexperiment* (thought experiment) and sparked a methodological debate on thought experiments, specifically in the scientific domain – physics in particular.



However, the authorship of the German term goes to the Danish physicist *Hans Ørsted* [21.29] and, perhaps, even before him to the German physicist and aphorist Georg Christoph Lichtenberg, who wrote about experimenting with thoughts [21.30, 31]. Actually, the very starting point of the philosophical interest about thought experimenting can be found in Immanuel Kant's philosophy [21.31], which is explicitly present in Ørsted's work [21.32], but also inspired Lichtenberg and many other philosophers [21.33].

Despite the controversy about who coined the term (on the topic see, for example, [21.31–37]), what is striking is that thought experimentation has a much older history – in philosophy, as well as in natural sciences. According to Nicholas Rescher, Pre-Socratics “invented thought-experimentation as a cognitive procedure and [...] practiced it with great dedication and versatility” ([21.38, p. 31] – see also [21.39, 40], for a critique of Rescher's view; on the topic of ancient thought experiment; see [21.41] and several contributions in [21.15]). Likewise, *Imre Lakatos* [21.42] located in Ancient Greece the beginning of thought experimentation, specifically in formal Euclidean mathematics [21.43]. Afterward the practice of thought experimentation continued (see [21.44–46] for discussions on thought experiments during Middle Ages) and flourished, mainly thanks to exceptional thought experimenters, such as Galileo, René Descartes, Locke, Newton and Gottfried Leibniz (on thought experimentation between Middle Ages and the introduction of the term, see contributions in *Horowitz and Massey* [21.13], *Casati et al.* [21.14], and *Ierodiakonou and Roux* [21.15]; on historical perspectives about Galileo's thought experimentation, see also *Prudovsky* [21.47], *Atkinson and Peijnenburg* [21.48], and *Palmieri* [21.49]).

### 21.2.2 The Classical Phase

In the classical phase, the first reflections on thought experiments are due to Lichtenberg and Ørsted (also Novalis, according to *Daiber* [21.50] and *Fehige* [21.51]; see also *Fehige* and *Stuart* [21.33] for an in-depth discussion of Lichtenberg's, Novalis', and Kant's analyses of thought experimentation). It is worth noting that former analyses mainly focused on thought experimentation in the scientific domain. Curiously, though, milestone scientific thought experiments in the history of physics, such as Galileo's on free-fall (Sect. 21.1.1) or Newton's bucket (Sect. 21.1.3), are rarely mentioned. Ørsted, for instance, mainly addresses thought experimentation in geometry and, when it comes to physics, only Kant's philosophy is mentioned.

One might resist counting Mach among the *classical* authors, given his systematic analysis of thought ex-

perimentation, which seems to go beyond the scientific domain. Indeed, when he lists thought experimenters, he speaks of dreamers and novelists – whilst not of philosophers [21.28, p. 451 of the 1973 English translation]). He also wrote that “Experimenting in thought is extremely important for cognitive development” and that “the thought experiment not only is of importance in the field of physics, but on the contrary, in all fields of knowledge” [21.28, pp. 455–456]. Still he analyses only thought experiments in physics and mathematics.

Moreover, in Mach's work we can find two traits typical of the classical phase: (i) it is not always clear how to distinguish between genuine thought experimenting and merely imagining about real experiments (REs) [21.52, p. 74]; and (ii) Mach does not worry about the distinct sphere of autonomy of thought experiments as such, and eventually he brings back the latter to real experiments. This is not intended to underestimate in any way the great significance played by Mach's analysis of thought experiments on the subsequent debate. Indeed, credit is due to his examination of the features proper to thought experiments (as well as to real experiments) and inquiry on how they function, even from a psychological point of view (Sect. 21.5).

*Alexius Meinong* first noticed the quite broad notion of thought experiment introduced by Mach. In his discussion of thought experiment [21.53], he addresses the distinction between experiments carried out *on* thoughts (i. e., psychological experiments based on the subjects' thoughts) and *within* thought (e. g., mathematical thought experiments). But he himself did not acknowledge that thought experiments should also be differentiated from imaginings about real experiments. The same can be said about another stakeholder of this stage, namely Pierre Duhem.

Scepticism about the epistemic value of thought experiments is common to Meinong and *Duhem* [21.54]. Duhem is known to have been a strenuous critic of the idea that thought experiments could play a legitimate role in the development of scientific thinking [21.34, 55, 56]. However, Duhem's thesis is not to reject thought experiments in toto [21.36, 57], but a certain use of them: Thought experiments should not be used as if they were real experiments just in order to conceal hypotheses arisen from pure supposition and abstraction.

Scepticism about thought experimentation is also present in both Karl Popper's and Hempel's inquiries, but it is accompanied by the acknowledgement of the positive uses of thought experiments. Their analyses have led these authors to suggest useful taxonomies of thought experiments [21.58, 59] (see Sect. 21.4.1). By contrast *Alexandre Koyré*, more in line with Lichtenberg, Ørsted and Mach, focused almost exclusively on the positive role played by thought experiments in sci-

entific inquiry (only in an appendix, he considers their misuse – see [21.60]).

### 21.2.3 The Contemporary Phase

In his milestone 1964 paper, *Kuhn* raised three questions that set the subsequent debate on thought experiments:

1. To what extent must the imagined situation be one that can be (or has been) found in nature, that is, what conditions of verisimilitude are thought experiments subject to?
2. How, “relying exclusively upon familiar data, can a thought experiment lead to new knowledge or to new understanding of nature?” [21.61, p. 241].
3. What, if any, kind of knowledge do thought experiments produce?

The authors of the classical phase have analyzed only thought experiments within the scientific domain, physics in particular. Although *Kuhn* is no exception, the many attempts that have been made to answer his questions have extended the enquiry to thought experiments in philosophy.

The contemporary phase is characterized by both a proliferation of works on thought experiments, due to the impact of *Kuhn*'s paper, and a very considerable production of thought experiments. Philosophy was the main protagonist of this new season of thought experimentation. Just to give some examples, between the 1970s and the 1980s: *Judith Thomson* [21.62] questioned on the concept of the right to life and how it differs from the concept of the right to what is needed to sustain life, through a bizarre kidnapping of a violinist by the Society of Music Lovers; *Put-*

*nam* wondered what would happen to the meaning of the word *water* in a twin Earth where what is called *water* has a different chemical composition from H<sub>2</sub>O (Sect. 21.1.5); *Searle* [21.12] attempted to refute the idea that the mind is a suitably programmed computer, by imagining himself in a room that receives and returns input in Chinese; *Derek Parfit* [21.63] *racked his brain* on personal identity and tried to show that the concept of identity is less important than that of survival by imagining a person splitting like an amoeba (for concise but exhaustive descriptions of *Thomson*'s, *Searle*'s and *Parfit*'s thought experiments, see [21.64]). As hinted earlier (Sect. 21.2.1), there have been philosophical thought experiments long before, but much of contemporary philosophy makes heavy use of thought experiments and it would be severely impoverished without them.

However, not all contemporary analyses of thought experiments deal with both scientific and philosophical examples (Sect. 21.4.3). As underlined by *Rachel Cooper*, although many authors have restricted their study to scientific thought experiments, a fine-grained theory of thought experimentation should cover both the sciences and the humanities [21.65–67].

Different answers to *Kuhn*'s questions have emerged in a stream of literature since the 1990s, with two polar positions: *James R. Brown* and *John D. Norton*. “The views of *Brown* and *Norton* represent the extremes of platonic rationalism and classic empiricism, respectively” [21.34, p. 69]. The best way to get into this still flourishing debate is by tackling three of the main issues which the contemporary phase has sought to clarify further: What is a thought experiment? What is the function of thought experiments? How do thought experiments achieve their function?

## 21.3 What Is a Thought Experiment?

The aforementioned examples give an idea of what a thought experiment is, but is it possible to offer a precise and comprehensive definition? The lively debate among contemporary philosophers has not led to a unanimous definition. Indeed, thought experiments are characterized variously as sometimes being arguments [21.68], specific ordered pairs [21.69], glimpses into a Platonic world [21.55, 70–72], experiments whose aim can be achieved without the benefit of execution [21.52], forms of “simulative model-based reasoning” [21.73], icons [21.74], abstract entities that are not particularly experimental, but rather an exploration and a refinement of theoretical models [21.75], guided contemplations [21.76] and pieces of counterfactual reasoning with experiment-like features [21.77].

Thus, it is a hard task to give a clear definition of thought experiment that does not widen the concept so as to make it practically useless. Too broad a definition would capture phenomena that intuitively are not thought experiments [21.78]. It has been claimed that we do not need a definition after all [21.55, 79]. *Geordie McComb* [21.80] suggests to see thought experiment as a cluster concept.

At a closer look, we can impose some order into the spectrum of definitions, which ranges from definitions that classify thought experiments as belonging to the theoretical realm, to those placing thought experiments within the experimental realm. In this section, I shall, first, review what has been said about the experimental side of thought experimentation (Sect. 21.3.1)

and, then, turn to the opposite end, namely the arguments given in favor of the theoretical nature of thought experimentation (Sect. 21.3.2). Finally, I shall dwell on the main features that should make thought experiments easily identifiable (Sect. 21.3.3).

### 21.3.1 Thought Experiments and the Experimental Realm

The issue about the relationship between thought and real experiments is a much discussed topic in the debate. The trend, as shown in Table 21.1, has been to underline a continuity between thought experiments and real experiments.

However, very often the experimental side of thought experimentation has not been evaluated per se: Thought experiments have been judged from the standards of real experiments, rather than on the basis of a broad definition of experiment that can include both types of experimentation. According to many scholars, thought experiments are not kinds of experiment, but tend to proceed as if they were.

The analysis of the experimental side of thought experiments seems to be influenced by a widespread

**Table 21.1** Simplified overview of some of the positions in the debate on the continuity or discontinuity between TE and RE. Strong continuists are authors who explicitly and extensively talk about the common features between the two types of experimentations. Even though Paul Humphreys admits a parallelism between thought and real experiments, he is considered as a strong discontinuist, because he sharply distinguishes between the theoretical realm of the former and the empirical realm of the latter [21.75, pp. 218–219]. In italics are highlighted philosophers who can be classified in the classical phase of the analysis of thought experiments. More will be said in the following about both Duhem and Marco Buzzoni

Status	Degree	
	Weak	Strong
Continuity between TE and RE	<i>Ørsted</i> [21.29] <i>Popper</i> [21.58] <i>Hempel</i> [21.59] Kuhn [21.61] Brown [21.55, 70] Szabó Gendler [21.81] Bokulich [21.56] Peijnenburg and Atkinson [21.79] Buzzoni [21.57] Brendel [21.82]	<i>Mach</i> [21.28, 83] <i>Koyré</i> [21.60] Sorensen [21.52] Nersessian [21.84] Gooding [21.85] Häggqvist [21.86] Wilkes [21.87] Bishop [21.88] Cohen [21.64]
Discontinuity between TE and RE	<i>Duhem</i> [21.54]	Hull [21.89, 90] Norton [21.68] Humphreys [21.75] Hacking [21.74]

preconception about the intrinsic epistemological superiority of real experiments. By following this preconception, we risk focusing on the features proper to *true* experiments (i. e., real experiments) that thought experiments lack. For instance, a typical plea for real experimentation would stress that, insofar as thought experimentation does not directly examine nature, it is less reliable and lacks justificatory power (Sect. 21.4.3). The upshot of this line of reasoning is that thought experiments should be employed only when real experiments are not available, otherwise they are useless.

Following *Roy Sorensen* [21.52], we might say that the problem is rooted in how the adjective *thought* should be interpreted in the expression *thought experiment*. A terminological attitude that can fall prey to the aforementioned preconception is to consider thought experiments as mere imaginary visualizations of experiments. In the works of many philosophers (particularly belonging to the classical phase), the expression *imaginary experiment* is frequently used as a synonym for thought experiment. However, substituting *thought* with *imaginary* can be misleading (see also *Krimsky* [21.91], who claims that all imaginary experiments are thought experiments but not vice versa, and *Wilkes* [21.87]). The imaginary unit and the imaginary number for mathematicians, as well as the social imaginary and the child imaginary for psychologists, are not degraded entities. Still *imaginary* is commonly used as an adjective that falsifies or somehow discredits the phenomenon to which it refers. An imaginary friend, imaginary worlds, imaginary fears and beliefs are understood as fictional entities. The emphasis is on the negative aspects, on what they lack in order to be real friends, worlds, fears or beliefs.

It is no coincidence that Duhem called thought experiments *expériences fictives* (fictitious experiments), given his harsh critique of thought experiments used as if they were real experiments. Similarly, albeit not motivated by the same causticity, Hempel spoke of imaginative experiments and seemed to complain about the fact that thought experiments tend to be merely heuristic, instead of providing purported evidence to be further validated. Hempel had been influenced, more or less explicitly, by the neo-positivist distinction between *context of discovery* and *context of justification* [21.55, p. 89]; for the relevant distinction, see *Hans Reichenbach* [21.92]). Contrary to real experimentation, thought experimentation would be confined to the domain of discovery, that is, the processes through which a hypothesis has been formulated, rather than how such hypothesis could be controlled and confirmed. The dichotomy between the context of discovery and that of justification seems to have influenced much of the analysis on thought exper-

iments belonging to the classical phase, but also some reflections of the contemporary phase, such as that of Norton (according to *Brown* [21.55]) and of David Hull (Sect. 21.4.3).

In his *Plea for real examples* [21.89, 90], Hull puts forward a critique of thought experimentation even harsher than that of Duhem. As suggested before (Sect. 21.2.2), Duhem's criticism should not be seen as a total rejection of thought experimentation. Although Duhem is commonly presented in the literature as a detractor of the role of thought experiments in scientific practice (thus, as a discontinuist – see Table 21.1), his remarks should lead us to re-evaluate (but not dismiss) thought experimentation (also suggested in *Daly* [21.93, p. 114]). He criticized a naive view of real experimentation and put forward very innovative ideas about it (e.g., real experiments are not a matter of mere observation without theory; they are subject to underdetermination in the choice of a theory, that is, they cannot test isolated hypotheses). Duhem focused his attention on the negative aspects of thought experimentation that convey such a naive view without realizing that thought experimentation itself can be seen in a less simplistic way and be subject to the same conceptual revision he was advancing for real experimentation. Most likely Duhem is a continuist, who undermined a naive view of experimentation as a whole, including thought experimentation.

Hull is more likely to be considered as a discontinuist. According to him, thought experiments are mostly useless, and real experiments should be preferred to them (on real examples that go beyond our imaginative ability; see [21.89, p. 312] and [21.90, p. 435]; on Hull's view see also Sect. 21.4.3). In fact, Hull seems to admit that thought experiments can have scientific value, but only if they involve an imaginary situation which is as plausible and detailed as possible (on the importance of a detailed scenario see also *Brendel* [21.82] and *Häggqvist* [21.86], which ties with issues discussed in Sects. 21.3.3 and 21.4.3). Moreover, he seems to take for granted that thought experiments must become, sooner or later, real experiments (on the issue about whether real experiments can resolve thought experiments, see *Arthur* [21.94]; Sect. 21.3.3). However, *Hull* is not willing to concede that thought experiments can be of value in all scientific fields, but only for those which are “well-articulated” ([21.90, p. 431], where he quotes other detractors of thought experimentation such as *Wilkes* [21.87], and *Fodor* [21.95]; other sceptics are in line with Hull's view – for example, *Feyerabend* [21.96], *Quine* [21.97] and *Thagard* [21.98, 99]). For example, he argues that in biology thought experiments cannot play any role; rather they risk to create only confusion (for a critique of this position,

see [21.100]; for an analysis of thought experiments in biology, see [21.101, 102]; see also [21.103], for a discussion on artificial life and thought experimentation in biology, and [21.104], for a recent discussion on the relationship between real and thought experimentation in biology).

Recalling what was said earlier with respect to how adjectives can transform the value of the name they modify, it is interesting to notice that Hull mostly calls thought experiments *fictitious examples*, but when he emphasizes their positive aspect, he employs the expression *hypothetical examples*. *Hypothetical* does not convey a negative value as *imaginary* does. Still it is more cautious than *thought*. For example, a hypothetical buyer is not really a buyer, she may become a buyer, but at that point she will be a *real* buyer and not anymore hypothetical. Thus, Hull's view on thought experiments is nicely exemplified by his use of the adjective *hypothetical* as a synonym for *thought*.

It should be noted that positive continuist views also oversimplify thought experimentation when they anchor it to real experimentation. For instance, it has been stressed that Mach talks about thought experiments as if they always lose to real experiments [21.52, p. 74]. *Mach* acknowledged that the outcome of some thought experiments, such as Galileo's thought experiment on falling bodies (Sect. 21.1.1), is strictly determined, so that the thought experimenter is led to consider superfluous “any further test by means of a physical experiment, whether rightly or wrongly” [21.28, p. 452]. However, he seems to complain about thought experimenters that avoid further real experimentation and take the result of thought experiments as conclusive. In this regard, Mach's critique of Newton's bucket thought experiment (Sect. 21.1.3) is a good example. According to Mach, Newton violated his rule of *hypotheses non fingere* (to feign no hypotheses), since he used the bucket thought experiment in order to show what was actually presupposed by the thought experiment itself (i. e., the existence of absolute space).

Similarly, even *Koyré's* analysis [21.60] might seem driven by considering real experimentation as the benchmark: Thought experiments are positive, since they accentuate positive features of real experiments. In this respect, *Koyré's* and *Duhem's* approaches can even be seen as complementary, more than opposed.

More recent views have tried to investigate thought experimentation as a genuine experimental practice on a par with real experimentation. At least two analyses are worth mentioning. *Sorensen* argues that thought experiments are a limiting case of real experiments. Both types of experimentation have very similar purposes and also share methods for assessing such purposes (Sect. 21.3.3 and 21.4). Clearly, they differ insofar as

thought experimentation emphasizes the design aspect at the expense of the execution aspect (other scholars have followed Sorensen on this point – for example [21.82, 85, 105]). Sorensen goes further and argues that, although historically thought experiments have become autonomous, their origin has to be individuated in the mental component of real experimentation. They should be seen as the result of an evolutionary process: a “selective pressure” would have deprived real experimentation of the execution aspect, emphasizing the design aspect ([21.52, pp. 186 and 212], [21.106] for a critique of Sorensen’s evolutionary explanation; on Sorensen’s view, see also contributions in the special issue of *Informal Logic* [21.107–110]).

Marco Buzzoni argues that Sorensen underestimates “the technological-operational dimension of the scientific experiment” and supports a concept of real experiment as a mathematical function ([21.57, p. 175]; see also [21.36]). Buzzoni [21.36, 57, 111] develops a Kantian framework according to which from an empirical point of view (i. e., exactly with respect to the *technological–operational* dimension) real and thought experimentations coincide, but they are complementary from a transcendental point of view. Thus, one type of experimentation without the other is unproductive for scientific purposes (see [21.112, 113] for objections to Buzzoni’s account and Buzzoni’s reply in [21.114]). Many other scholars have underlined that thought experimentation shows an action-practical component (e.g., [21.85, 101]; see Sect. 21.3.2 and 21.5). Moreover, this component has been adduced as one of the main arguments against views that confine thought experiments into the theoretical realm.

### 21.3.2 Thought Experiments and the Theoretical Realm

Instead of considering thought experiments either as rough copies of real experiments or as peculiar experiments, it might be claimed that they do not belong at all to the experimental realm. There are two major proponents of this point of view: Norton and Humphreys (Ian Hacking’s view might also belong to this interpretative current, since he sharply distinguishes thought from real experiments and considers the former as *static* entities. However, he seems to concede that the embodiment aspect is important for thought experimentation [21.74]; Sect. 21.5).

Norton argues that thought experimentation cannot be a type of experimentation, because it lacks the essential element proper to the latter, namely interaction with the natural world [21.115]. According to Norton,

thought experiments are disguised arguments [21.68, 115–117]: a good thought experiment should be a sound argument that increases our knowledge. In other words, without epistemic loss a thought experiment can be reconstructed (translated, or reduced) into an argument – i. e., a list of propositions, of premises and assumptions, leading to a conclusion via (inductive or deductive) inferences. Thought experiments are often rhetorically embellished and frequently they do not make explicit all the assumptions on which they rely: these features conceal their argumentative nature. Along with this *reconstruction thesis*, Norton suggests an *elimination thesis*: Thought experimentation is a dispensable epistemic tool (see Gendler Szabó [21.76, 118] for a fine-grained analysis of Norton’s elimination thesis; see also Timothy Williamson’s view [21.119] – which seems in line with Norton’s view, except for the role granted to imagination – Sect. 21.5).

Humphreys claims that thought experiments “lie much closer to theory than to the world” [21.75, p. 218]. He admits that they can be assimilated to that type of real experiments that isolate “those features of the world that are represented in a theoretical model” and approximate “the idealizations that are employed therein” [21.75, p. 218]. But nowadays, according to him, this function is fulfilled by theories. In support of his argument, he compares thought experiments to computer simulations (or numerical experiments). Both methods involve refinements of theories, adjustments to conform conditions, parameters, approximations and idealizations to empirical data, and can deliberately alter parameters in order to produce laws different from those of our world.

Actually, the parallelism between thought and numerical experiments can show that the issue of the experimental nature of thought experiments is still open. Indeed, a lively debate among philosophers of science on the status of computer simulations has led to considerations very similar to those discussed earlier (Sect. 21.3.1) about thought experiments. For instance, some (e.g., [21.120]) argue that real experiments and numerical experiments could not possibly differ from each other more and that the latter can be seen as arguments [21.121]; while others regard numerical experiments as a genuine experimental practice (e.g., [21.122, 123]), whose analysis has been often influenced by a bias for real experiments [21.124, 125].

Several parallelisms between thought experiments and numerical experiments can be drawn and a comparative analysis can shed light on both topics at once. Even though some authors have suggested that numerical experiments can be seen as type of thought experiments ([21.126]; see also [21.103], for a specific

biological case, and [21.127], for a provocative view according to which computer modeling will replace thought experimentation), the *trading zone* between thought experiments and numerical experiments has been sparsely considered by current works on either thought or numerical experiments. These works have primarily focused their attention on the links between these two scientific tools and real experiments (in-depth analysis can be found in [21.128–131]; in passing other authors have commented on the parallelism between thought experimentation and numerical experimentation – for example, [21.36, 43, 52, 57, 64, 65, 73, 101, 132]; see also the related topic of video games as executable thought experiments [21.133]).

Much criticism has been raised against theoretical views about thought experimentation, especially against Norton’s argument view. Although some have found the latter too liberal (e.g., [21.134]), most philosophers have found it too restrictive and have offered four main objections. First, Norton’s translation of thought experiments into arguments would lose some important aspects proper to thought experimentation (e.g., [21.71–73, 82, 101, 135–138]), such as its nonpropositional dimension or an action-practical component. These aspects should not be neglected, for they play an epistemic role, rather than being merely picturesque. *Tamar Gendler Szabó* [21.76, 118], for instance, maintains that Galileo’s thought experiments on falling bodies (Sect. 21.1.1) cannot be fully reconstructed into arguments (see also [21.71, 72, 139] for other examples; see [21.140] for a reconstruction of Galileo’s thought experiment within a nonclassical logical framework). It has been pointed out that the same holds for thought experiments that rely heavily on sensory imagination or spatial reasoning (e.g., [21.65]).

A second related objection concerns the cognitive underpinnings of thought experimentation. The same conclusions can be drawn from a thought experiment and from a logical argument, but constructing and performing the former are different from producing and carrying out the latter. This is so even if we take for granted that all thought experiments are translatable into arguments and that such a translation procedure is epistemically advantageous (Sect. 21.5).

Third, it has been pointed out that thought experiments may feature in the argumentation steps, but this does not mean that they are arguments. Likewise for real experiments. Real experiments can play a role in or be rephrased as arguments, but typically they are not considered as arguments and it is unlikely that someone would claim that they are dispensable. We should not confuse an experiment of whatever kind with its published description [21.86, 101].

Finally, *Michael Bishop* ([21.88]; see also [21.141]) has offered a counterexample to Norton’s view: The same thought experiment can be reconstructed in two different arguments. Often this is the case when scholars disagree about the upshot of a thought experiment; otherwise it would be impossible to compare their views and determine who is right. The main example given by Bishop is the debate between Einstein and Niels Bohr on an Einsteinian thought experiment, namely the clock-in-the-box thought experiment (see the volume dedicated to Einstein of *The Library of Living Philosophers*, edited by Arthur Schlipp, for a complete description of this thought experiment and also for Bohr’s objections and Einstein’s reply [21.142]).

Nevertheless, one might think that Norton’s argumentative reconstruction thesis is valuable, while maintaining that thought experiments are not arguments and/or rejecting the elimination thesis. Once translated into arguments, thought experiments can make explicit their implicit assumptions. As pointed out by *Richard Arthur*, “the reformulation of thought experiments as arguments is a vital part of the scientific process” ([21.136, p. 228]; see also [21.101]; see [21.143] for a proposal which takes into account both the experimental and the argumentative sides of thought experimentation).

### 21.3.3 Thought Experiments and Their Features

Despite the fact that there is not a unanimous definition of thought experiment and different views push thought experimentation toward either the empirical or the theoretical realms, there are some features common to most thought experiments. It should be noted that, although discussions about these features often lead to draw several parallelisms between thought and real experimentations, they are not committed to the experimental nature of thought experimentation. After all it might be profitable to study thought experiments as if they were experiments, even if they are not [21.52, 132].

In what follows I shall focus, on three features common to both thought and real experimentation:

1. *The method of variation* (i. e., isolation of variables, *manipulation* and *observation*)
2. Fallibility and
3. Theoretical underdetermination.

I shall then turn to the main feature proper to thought experimentation only (i. e., the mental nature of its laboratory) and some connected features.

### The Method of Variation, Fallibility, and Theoretical Underdetermination

Mach [21.28] maintained that the leading principles of real experimentation must hold for thought experimentation as well. According to Mach, the experimental practice in toto is based on the *method of variation*. He wrote that [21.28, p. 452]

“It can be seen that the basic method of the thought experiment is just like that of a physical experiment, namely, the method of variation. By varying circumstances (continuously, if possible) the range of validity of an idea (expectation) related to these circumstances is increased.”

Many philosophers agree with Mach and have stressed that the method of variation is a core feature of thought experimentation [21.36, 57, 75, 82, 86, 87, 132]. It has been pointed out that such a method is common also to numerical experimentation [21.128].

The method of variation can be seen as a three-step procedure describing what in general a thought experimenter has to do: (i) select and isolate the features which act as variables, (ii) *manipulate* these variables, that is, make them interact, and (iii) finally observe what consequently happens.

The first step leads to the question: *what are the variables involved?* Here answers diverge, since they are hostage to the held view about the nature of thought experimentation (Sects. 21.3.1 and 21.3.2). Can thought experimentation examine nature or does it merely explore theoretical models? Anyway, it seems that thought experimentation deals more with abstract representations [21.128] or suppositions ([21.52]; see also Goffi and Roux [21.144], who speak of beliefs) than natural circumstances and concrete entities.

As far as the second step is concerned, one might be uneasy with the fact that in thought experimentation an experimenter is not literally *manipulating* the variables in question. Despite philology, however, manipulating is not merely influencing manually. Things can be rotated and moved also in our imagination. Expert chess players or Rubik’s Cube solvers do perform such kind of mental manipulation [21.52, 145]. This consideration leads us to the third step of the method of variation, namely the observation of the interactions among the variables ([21.64, 73, 85, 146–148] have particularly stressed both (ii) and (iii)).

Observation or visualization has seemed to many a necessary condition of thought experimentation, as well as of real experimentation [21.36, 57, 71, 72, 85, 147, 149, 150]. The problem is that it is not clear whether observation means the same thing in both contexts. In thought experimentation, observation seems

not to be grounded in perceptual experiences as in real experimentation. Often enough authors speaking of observation in thought experimentation seem to refer to a representation of the described situation in the mind’s eye and to imagination as a vehicle for quasi-observation [21.64, 90, 103, 145, 150]. Making observations, however, can be interpreted in a less perceptual sense. John Gilbert and Miriam Reiner [21.148], for instance, take thought experimental observations as outcomes produced by logical laws, without denying that in a thought experiment a world containing objects is imagined (this issue relates to that of the underpinnings of thought experimentation, see Sect. 21.5).

Other features of thought experimentation have been pointed out. Two of them are worth mentioning, which are still common to thought and real experimentation, namely fallibility and theoretical underdetermination.

Like real experiments, thought experiments can fail. Alan Janis [21.151] has underlined three different ways in which real and thought experimentation may fail. First, an experiment can fail because of its incompleteness, due to inadequacy of the equipment, or to external factors (see also [21.64, 65]). It is difficult to give examples of this category, because usually they are not published. Second, an experiment can fail when its results are incorrect (e.g., Einstein’s clock-in-the-box, according to Janis). Third, an experiment can yield correct results, but ones that elude the question which motivated the experiment in the first place (Janis gives as example the Einstein–Podolsky–Rosen (EPR) thought experiment – see next paragraph). Arguably, thought experimenters, similarly to real experimenters, need to engage in error management in order to improve their results (e.g., the thought experimental debate between Einstein and Bohr, as well as that between Darwin and Fleeming Jenkin – as for the latter, see [21.100, 101]).

Alisa Bokulich [21.56] has suggested that thought experiments can show theoretical underdetermination in that they cannot discriminate between different theoretical frameworks (see also [21.86, particularly ch. 6], [21.152] and [21.153]). EPR, a thought experiment by Einstein et al. [21.154], is an example of a thought experiment that can be *rethought* from the perspective of different and incompatible theories [21.56, p. 299]. Sidestepping the technical details of such thought experiment, what is important to keep in mind is that its upshot is a *dangerous* correlation between two physical quantities (position and momentum), which would undermine quantum mechanics (QM) (in its Copenhagen version). Commonly EPR has been interpreted as a failed demonstration of the incompleteness of QM and as an argument in embryo form for a determinis-

tic completion of QM, which is indeterministic. Indeed, precisely on the basis of EPR, David Bohm developed the most famous deterministic version of QM. However, per se EPR is not a crucial thought experiment that can help us to decide between QM and its deterministic rivals (EPR is widely discussed in the literature on thought experiments – for example, see the debate between *Atkinson*, [21.155], and *Stöltzner*, [21.43]). This feature of thought experiments undermines the idea that they do not have a life of their own [21.74], that is the ability to evolve and be adapted to different theories and ends (also philosophical thought experiments seem to have a life of their own – for example, see *Twin Earth* chronicles in [21.156]).

### The Laboratory of the Mind

The most striking feature of thought experiments is that they are conducted in the “laboratory of the mind” [21.55, 70]. Thought experimentation seems to be grounded in imagination (Sect. 21.5). The fact that we can *experiment* within our mind has its advantages. As remarked by *Mach*, “Our own ideas are more easily and readily at our disposal than physical facts. We experiment with thought, so to say, at little expense” ([21.28, p. 452]; see also [21.52] for a discussion of the advantages of thought experimentation over real experimentation). However, the fact that thought experiments are not in direct contact with natural phenomena and are merely a product of our imagination has its shortcomings (Sects. 21.3.1 and 21.4.3).

Two other features are tied to the mental nature of thought experimentation. First, thought experimentation seems not able to give quantitative outcomes, since it does not involve instrumental apparatus. However, at least scientific thought experiments can give quantitative results (e.g., Ronald Fisher’s thought experiments that explained the influence of natural selection on sex ratio – [21.52, p. 250]; Sect. 21.4.3). Still, the outcomes of real experimentation seem to be fixed and possible to be determined in a way that the outcomes of thought experimentation cannot ([21.157]; see also [21.52, p. 247], on the unavoidable incompleteness of thought experimentation, which ties with the issue about philosophical thought experimentation heavily relying on unclear background assumptions – Sect. 21.4.3).

Second, it has been argued that a genuine thought experiment should not require a concrete implementation, which can even be impossible for practical or ethical–political reasons. *Sorensen* [21.52], for instance, conceives thought experiments as experiments in which the design aspect is accentuated at the expense of the execution aspect (Sect. 21.3.1). Furthermore, he has identified three reasons (impossibility, unimprovableness, unaffordability) that explain why thought experiments need not to be concretely performed (see the previous discussions on this issue in [21.54, 83]). *Sorensen*’s view can be summed up by means of a spectrum: There would be nonimplementation, on one extreme, due to the maximization of benefits and, on the other, due to containment of losses.

To put it in another way, even when possible, a real performance of a thought experiment would be irrelevant to the purpose of the thought experiment [21.72]. This does not mean that thought experimentation cannot lead to real experimentation. Arguably, a thought experiment can open new lines of inquiry, which can be explored by means of real experiments. Still the resulting real experiment should not be seen as the realization of the initial thought experiment.

Some authors disagree and claim that, at least some, thought experiments can be concretely implemented and that, more generally, thought experimentation should be resolved into real experimentation [21.78, 79, 89, 90, 155, 158]. A classical example given in favor of this view is *Alan Aspect* and colleagues’ real experiments, whose results have been published in a paper titled *Experimental Realization of Einstein–Podolsky–Rosen–Bohm Gedanken Experiment: A New Violation of Bell’s Inequality* [21.159]. However it is possible to contend this interpretation of EPR. Despite part of the title of Aspect and colleagues’ paper, the real experiment they conducted can be considered as an empirical test of a hypothesis suggested by John Stuart Bell, who found it by studying Bohm’s version of EPR ([21.152]; see also [21.43, 160–162] for similar views on EPR).

On a moderate view, some thought experiments can be concretely performed, without denying a genuine status to thought experimentation (on this topic see recent discussions in [21.16]; Sect. 21.3.1).

## 21.4 What Is the Function of Thought Experiments?

Independent of the adequate definition of thought experiment, scholars widely agree on what it should do: to increase our knowledge. Thought experiments have generated a lot of epistemological interest, mainly

sparked by Kuhn’s inquiry (see Sect. 21.2.2). The main epistemological questions addressed in the literature are the following: What kind of knowledge do thought experiments really produce? To what ex-



tent are thought experiments a reliable source of information? What role do thought experiments play in processes of rational choice? The last question is strongly connected to the issue about how to classify thought experiments according to their epistemic functions. I shall begin by presenting some proposed taxonomies (Sect. 21.4.1), then I shall turn to the questions about the type of knowledge (if any) produced by thought experimentation (Sect. 21.4.2) and about its status (Sect. 21.4.3).

### 21.4.1 Sorting Thought Experiments

It might be useful to have an efficient classificatory scheme of immediate understanding, in order to put some order in the domain of thought experimentation and to try to understand it better. However, this is not an easy task. Thought experiments are employed in so many disciplines. Moreover, their interpretation can depend on historical factors ([21.163]; see also [21.164]) and on the intention of the thought experimenter, indeed they can even be rethought for different purposes ([21.56]; Sect. 21.3.3).

Thought experiments can be classified along several dimensions (e.g., by domains such as science vs philosophy, by type of reasoning such as inductive vs deductive). However, most taxonomies classify thought experiments according to their functions with respect to a group of hypotheses or a theory (several taxonomies are put forward by different scholars in [21.13]). None of these taxonomies seems to be definitive, but one has become quite popular, namely the taxonomy proposed by Brown ([21.81]; see [21.165] for a critique of this taxonomy).

Brown firstly divides thought experiments into two general types, destructive and constructive. A thought experiment falling within the former category is “a picturesque *reductio ad absurdum*” ([21.55, p. 34]; see also [21.70, p. 123]) devised in order to reject, or at least seriously undermine, some hypotheses or a theory. Here Brown rejoins Popper’s taxonomy and his critical use of thought experiment [21.58], which in turn is analogous to Hempel’s *theoretical* thought experiments – although the latter category goes beyond thought experiments against theories and encompasses all thought experiments that explicitly make fruitful predictions [21.59].

There are different ways of undermining a theory, thus suggesting different subcategories of destructive thought experiments. At least two subcategories can be offered. First, a thought experiment can show a problem *internal* to a given theoretical framework. This is the case, for instance, in Galileo’s falling bodies thought experiment (Sect. 21.1.1), since it shows an inconsis-

tency within Aristotle’s account of motion (due to the Aristotelian hypothesis that speed is proportional to weight).

Second, a thought experiment can show a problem *external* to a given theoretical framework, that is, between the latter and other assumptions or theoretical frameworks. Erwin Schrödinger’s cat thought experiment is such an example, since it underlined how QM (in its Copenhagen interpretation) was in conflict with our beliefs about the macroscopic level.

According to the Copenhagen interpretation of QM, a physical system can be in a very special state which actually is a simultaneous superimposition of different states. Once observed or measured, the physical system collapses into one of the superimposed states. This physical phenomenon occurs only at the quantum or microscopic level, but the problem is precisely where to draw the divide between the latter and the macroscopic level (i. e., the object of study of classical physics). As pointed out by Schrödinger, macroscopic objects like cats are not likely to be at the same time dead and alive (Fig. 21.4).

Following Brown and Fehige [21.27], a third subcategory might be added, namely “counter thought experiments” [21.166] or “thought-experiment/anti-thought-experiment pairs” [21.115, 117]. Counter or anti-thought experiments target thought experiments, more than theoretical frameworks. Examples of this category are Lucretius’ thought experiment (originally introduced by the Pythagorean Archytas; see [21.115, 167]), meant to undermine an Aristotelian thought experiment on finiteness of space, and Mach’s version of Newton’s bucket (Sect. 21.1.3) aimed at showing that centrifugal forces are due to the rotatory motion relative to the terrestrial mass and other celestial bodies ([21.168]; on this topic see, for example, [21.31, 116, 169]). Popper stressed that counter-thought experiments run the risk to be unacceptable, because unfair with respect to the opponent’s position [21.58, p. 466]. This is the apologetic use condemned by Popper (on the latter and Popper’s critical use, see [21.170]).

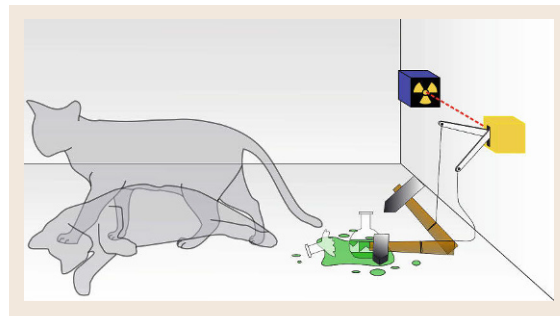


Fig. 21.4 Schrödinger’s cat thought experiment

Constructive thought experiments aim to support a theory or theoretical hypothesis, but they can do so in very different ways. Thus, Brown divides this category into three further types, namely mediative, conjectural, and direct. Mediative thought experiments have a pedagogic or illustrative role (generally on the pedagogical role played by thought experiments see, for example, [21.148, 171–176]). Indeed, they help us to better understand the conclusions that can be drawn from a specific theory. Brown gives as an example James Clerk Maxwell’s demon thought experiment. According to the kinetic theory of Maxwell, there is a probability, albeit very small, that heat moves from a cold body to a hot one. The second law of thermodynamics, however, implies the impossibility of such an event. To show the logical possibility of violating classical thermodynamics, *Maxwell* proposed his thought experiment of the demon [21.177].

Imagine two interconnected boxes; one filled with cold gas (C) and the other with hot gas (H). A very small door controlled by a demon is in between the two boxes (Fig. 21.5). The demon lets fast molecules go from C to H, and slows molecules go from H to C. In this way, while the average speed of the molecules in H would increase, the average speed of the molecules in C would decrease. Since according to Maxwell’s theory, heat is nothing more than the average speed of the molecules, the thought experiment shows the possibility of the flow of heat moving from a cold body to a hot body.

Given their illustrative and expository role, Brown’s mediative thought experiments recall *Popper’s heuristic* use of thought experiment [21.58] – a category profoundly similar to *Hempel’s inductive* thought experiments [21.59]. According to Brown, however, positive thought experiments can do more than merely illustrate a theory; they can help in constructing a theory. This is precisely what both conjectural and direct thought experiments aim to do. Contrary to mediative thought experiments, they do not start from a specific theory, but they end with one. What distinguishes conjectural from direct thought experiments is that they make up conjectured phenomena and put forward theories in order to explain them. Brown gives Newton’s bucket experiment (Sect. 21.1.3) as an example of con-

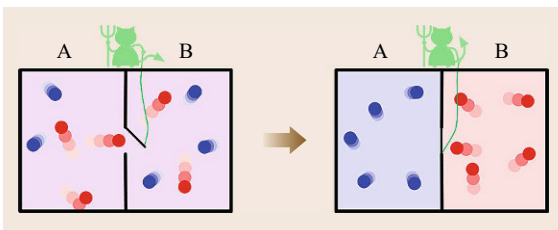


Fig. 21.5 Maxwell’s demon

jectural thought experiment, since it advanced a problem (i. e., things being equal from a relative point of view, there can be different effects on the surface of some water contained in a rotating bucket) and its solution (i. e., we should distinguish between relative and absolute motions, where the latter refer to the absolute space). It should be noted that Mach would have disagreed with such an interpretation of Newton’s bucket thought experiment, which will turn to be a mediative more than a conjectural thought experiment. According to *Mach* [21.168], from the thought experiment, we can reach the conclusion that absolute motions and space do exist, only if we accept from the beginning the existence of absolute space and the distinction between absolute and relative motions. Moreover, if we consider Newton’s bucket as a thought experiment run against relativist theories of motion (such as Descartes’s and Leibniz’s ones), it can also be seen as a destructive thought experiment.

Direct thought experiments establish new theories starting with unproblematic phenomena. An example of this category is Stevin’s chain thought experiment, since it introduced Stevin’s law of the inclined plane – i. e., the force to the weight is equal to the ratio of the height to the length of the plane (Sect. 21.1.2; this thought experiment can be seen also as destructive [21.153]).

Finally, according to *Brown* some thought experiments are both destructive and direct-constructive, these are platonic thought experiments [21.55, 70–72]. An example of this category is Galileo’s thought experiment on falling bodies (Sect. 21.1.1), since at the same time it undermined the Aristotelian theory of motion and put forward a new theoretical framework. As *Koyré* [21.60] remarked, such a thought experiment seems an example of good physics made a priori. This is precisely what characterises platonic thought experiments; they are vehicles of a priori knowledge (Sect. 21.4.2).

The following schema (Fig. 21.6) sums up Brown’s taxonomy.

Although Brown applies his taxonomy only to scientific thought experiments, it might be extended to

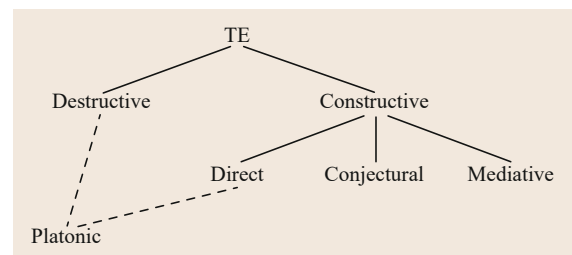


Fig. 21.6 Brown’s taxonomy of thought experiments (after [21.55])

encompass philosophical thought experiments too (as suggested in [21.27]). For example, Twin Earth thought experiment (Sect. 21.1.5) may count as internal-destructive, since it targets a hypothesis assumed by internalist theories of meaning (i. e., identical psychological states imply identical references). The same thought experiment can also be seen as a conjectural-constructive one, since it highlights a problematic phenomenon (i. e., the situation in which identical psychological states imply different references) and suggests a solution (i. e., meanings are constrained by natural kinds). The fact that a thought experiment can be difficult to classify is specific neither to philosophical thought experiments (see what was said earlier about Newton’s bucket), nor to Brown’s taxonomy (as will become clearer in the following). This is nothing but a symptom of how sorting thought experiments is difficult in itself.

Two more taxonomies are worth mentioning, which have sought to integrate also philosophical thought experiments. *Sorensen* has proposed a taxonomy of thought experiments driven by the idea that thought experiments are “stylized” paradoxes [21.52, p. 165], which serve as “alethic refuters” [21.52, p. 135]. In other words, thought experiments can be seen as “expeditions to possible worlds,” whose mission is “to refute a source statement that has an implication about the constituents of these worlds” [21.52, p. 135]. In analogy with the two alethic modalities, *Sorensen* divides thought experiments into two categories: refuters of possibility and refuters of necessity. Both argue against a theory or theoretical framework: the former by pointing out inauthentic possibilities wrongly considered as authentic, the latter by revealing neglected genuine possibilities. *Sorensen* gives many examples, for instance, while *Gettier*’s thought experiments (Sect. 21.1.4) and *Maxwell*’s demon are necessity refuters, *Schrödinger*’s cat and *Mark Johnston*’s [21.178] thought experiment against views on personal identity that make the latter dependent on future events are possibility refuters (see [21.86] for a critique of *Sorensen*’ logical regimentation, which is at the bottom of his taxonomy; it might be interesting to explore the link between *Sorensen*’s taxonomy and the one proposed in [21.179], pivoting on the idea that while some thought experiments enlarge the domain of properties pertaining to the actual world, others restrict such a domain).

*Gendler* [21.118] proposes a tripartite taxonomy of thought experiments pivoting on three different questions that can arise from a thought experiment. First, one may wonder what would happen if the imaginary scenario take place. This is what factive thought experiments ask. Second, there are conceptual thought experiments that pose the question of how what takes

place in the imaginary scenario should be described. Finally, valuational thought experiments assess the appropriate, moral, or aesthetic, evaluation of the envisaged situation. However, this taxonomy can be reduced to two broad categories. *Gendler* takes thought experiments as contemplations of imaginary cases that force us to account for the represented exceptional episodes and identifies two strategies for doing that, namely exception-driven and norm-driven. On the one hand, the exception helps us to establish the norms of a theory (e.g., *Galileo* on falling bodies – Sect. 21.1.1) and, on the other hand, norms guide us in evaluating the exception (e.g., the thought experiment of the ship of *Theseus*, which questions criteria for identity). While factive thought experiments typically fall within the former strategy, conceptual and valuational thought experiments fall within the latter.

Note that some authors might complain that all these taxonomies neglect to consider other specific types of thought experiment. First, some thought experiments seem neither to refute, nor to support a theory, but rather to be part of the theory itself. These thought experiments have been called *functional*, precisely because they have “a specific function within a theory” [21.180, p. 384]. In psychological testing, for instance, thought experiments advancing brainwashing procedure made possible to apply a frequentist conception of probability [21.180, p. 384]. It has been pointed out that functional thought experiments can also be found in modern physics [21.43].

Second, some thought experiments can start by clarifying conceptual issues and then turning to provide the basis for normative judgments. This seems to be the case with respect to the concept of money in economics. *Julian Reiss* has labeled such thought experiments “genealogical” ([21.181]; see also [21.182] for parallels between thought experiments in physics and in politics – on political thought experimentation, see also [21.183]).

Third, by analyzing thought experimentation in quantum gravity, *Mark Shumelda* [21.184] stresses that thought experiments can also be used in order to impose logical constraints on future scientific theories.

### 21.4.2 Thought Experiments and Kinds of Knowledge

A good thought experiment should be conducive to a new justified belief about the world or, at least, our interpretation of the latter. For instance, thanks to *Galileo*’s thought experiment on free fall (Sect. 21.1.1), we know that speed is not proportional to weight. Hence, we have, on the one hand, reasons against the Aristotelian theory of motion and, on the other hand, ev-

*idence* in favor of Galileo's theory, according to which speed is proportional to time.

Nevertheless, the fact that thought experimentation can produce knowledge is not a simple issue, and it turns out to be more complicated than in the case of real experimentation. Leaving aside arguments to the effect that thought experiments do not at all increase our knowledge, disagreements arise when philosophers want to specify the kind of knowledge that we gain through thought experimentation. This is easily seen with the problem of informativeness raised by Kuhn (Sect. 21.2.2): How can a thought experiment yield new empirical knowledge without the input of new data? This question has a paradoxical flavor due to the fact that only real experimentation is in direct contact with the world, from which it directly derives new materials. By contrast, thought experimentation is bound to use only old data, stored in the mind of the thought experimenter. How, therefore, can thought experiments provide us with new knowledge or understanding of nature? And what kind of *new* knowledge would they produce?

Very different stances have emerged in the many attempts to answer this epistemological question since the 1990s. Two among them can be seen as the polar positions of the relevant logical space: Brown and Norton, respectively, claiming that there is, and that there is not, new knowledge.

Following Koyré, Brown identifies a set of a priori thought experiments. These are, in Brown's terminology, Platonic thought experiments (Sect. 21.4.1), since they are neither based on new empirical data, nor simply inferred from old ones [21.55, 70–72]. These thought experiments are to be considered constitutively a priori and a source of knowledge independent of experience. How are we to explain, for example, the transition from the Aristotelian theory to the Galilean theory of motion? The right answer cannot be new sensory data, since none has been added. According to Brown, it is not even possible either to invoke any logical truth, that allows us to infer that all bodies fall at the same speed, or to appeal to other criteria, such as aesthetic ones (e.g., that of simplicity). Platonic thought experiments allow us to *see* the laws of nature. Many authors have criticized Brown's aprioristic account, above all his risked extension of Platonism from mathematics to physics [21.65, 75, 136].

According to Norton, pure thought is totally unable to generate any kind of knowledge, except from logical truths, and can only transform what the subject already possesses [21.115, p. 49]. Moreover, he criticizes a fundamental assumption in Brown's account:

the parallelism between visual and platonic perception [21.116]. Norton highlights that we have good criteria for assessing the unreliability of the former, but the same does not hold for the latter, which relies on both imagination and intuition [21.71, 72, 150]. Finally, since he argues that thought experiments can be reconstructed into arguments without epistemic loss (see Sect. 21.3.2), Norton denies that thought experimentation has a distinguishing kind of epistemic force. Thus, thought experiments can increase our knowledge, but only in the way logically sound arguments can do.

Between these two extremes, many scholars agree in thinking that new knowledge can be gained via thought experimentation. For example, *Humphreys* [21.75] maintains that thought experimentation provides a better understanding of the conditions needed for a theoretical model to hold. By contrast, *Gendler* [21.118, 145] claims that via thought experiments, we can get either new justified beliefs about contingent aspects of the natural world or new justifications for old beliefs. Gendler's reflections are clearly influenced by Kuhn, who already emphasized the importance of thought experiments in conceptual reconfiguration. Moreover, like others (e.g., [21.73, 82]), she also follows some insights of *Mach*, who argued that thought experiments make explicit a kind of inarticulate knowledge, not yet organized in theoretical frameworks, though stored in memory [21.28, 83]). In line with Mach's view, it has been claimed that successful thought experiments transform ability knowledge into propositional knowledge [21.185].

As far as the kind of knowledge is concerned, while some authors have maintained that thought experimentation involves both a priori and a posteriori knowledge (e.g., [21.186]; see also [21.136] for an aprioristic account of thought experiments), others have strongly criticized any aprioristic account of thought experimentation. For instance, although *Rodney Snooks* [21.78] agrees with Brown in thinking that thought experiments are a direct vehicle for the laws of nature, he argues that they do not give us access to a priori truths (see also *Hopp* [21.138], where a phenomenological approach is defended according to which thought experimentation can lead us to intuit universals and relations among them).

The debate has tackled also other kinds of knowledge, for instance, both modal and counterfactual knowledge [21.119, 187].

The issue about the kind of knowledge (e.g., new/old, a priori/a posteriori, universal/contingent, conceptual/empirical) gained via thought experiments is not the only thorny problem. Also, the issue about the status of such knowledge remains open.

### 21.4.3 The Epistemological Status of Thought Experiments

Is the knowledge gained via thought experimentation valid or reliable? And more generally we may ask: Are thought experiments indispensable epistemic tools? These questions, on the one hand, take us back to the comparison between real and thought experimentation (Sect. 21.3.1) and, on the other hand, open the issue about thought experiments in philosophy. In what follows, I shall address these issues respectively and devote a final subsection to the topic of intuitions.

#### The Proper Functions of Thought Experimentation

As we have seen (Sect. 21.3.1), thought experimentation is often considered as being of a rank lower than real experimentation, as if they were competing strategies. That is, the two types of experimentations perform the same function and the real type is to be preferred when possible (Sect. 21.3.3). And indeed, real and thought experimentations seem to play very similar roles in the evaluation of theories: Both test hypotheses, help to refine theories and similarly may fail in achieving these goals. However, one might ask whether there is a functional difference between thought and real experiments.

Some authors argue that, in fact, contrary to real experiments, thought experiments cannot have a justificatory role, but only an illustrative or explanatory role [21.89, 90, 98, 99]. However, this position does not do justice to both types of experimentations. After all, even real experiments are not only means of theoretical justification. Moreover, reasoning along this line tends to focus only on the justificatory inadequacy and to see it as the major limit and deficiency of thought experimentation. Once again the running idea seems to be that thought experimentation has no role to play within the context of justification and should be confined to the context of discovery (see Sect. 21.3.1). Once again we run the risk of underestimating the peculiarities of both types of experimentation. What, then, is the proper function of thought experiments, which sets them apart from real ones and is also what motivates us to use them instead of the latter?

In the literature, answers given to this question are not crystal clear. Let us mention some of them. It has been emphasized that thought experimentation provides us with idealization and modeling of reality to a higher degree compared to real experiments (e.g., *Koyré* [21.60]); idealization can also be a source of unreliability, for a discussion on this topic, see *Sorensen* [21.188]). However, it is questionable whether this really answers the above question, or merely

changes the focus to *how* thought experimentation functions (Sect. 21.5). *Gendler* [21.118] has proposed to see the functional difference in the type of results. Both thought experiments (at least scientific ones) and real experiments tell us about the real physical world, but via the former we obtain *intuitions*, whereas via the latter *data* (Sect. 21.3.3). A question arises: Do we really make use of thought experiments because we are in search of intuition rather than data?

Inspired by Kuhn, *Bokulich* [21.56] has suggested that thought experimentation tests the nonempirical virtues of theories, such as (internal or external) coherence, simplicity, and fruitfulness (for the notion of nonempirical virtues, see [21.189]; similar notions can be found in [21.190, 191]). On this view, Galileo's thought experiments showed an incoherence internal to Aristotle's theory of motion, stemming from an ambiguous use of the concepts of *speed* and *weight*. He also dared to go beyond the impasse, and to propose a new theoretical framework within which he could account for the phenomena. Bokulich's conclusion on thought experimentation in physics finds a parallel in the work of *James Lennox* [21.101, 102] on Darwin's thought experiments (see also the claim that thought experiments in science test how unified a theory is in [21.132]). Indeed, Lennox argues that thought experiments are functionally experiments, but we appeal to them under special conditions: "thought experiments are especially important when the issue at hand is the theory's *potential* to explain *as*, and *what*, it claims it will" [21.101, p. 236]. It has been proposed to extend a similar approach to philosophical thought experiments [21.152].

It should be noted that most authors who challenge the epistemic validity of thought experimentation do not object to scientific thought experiments. *George Bealer* [21.192] has proposed to formalize this view terminologically. According to him, the expression *thought experiment* should refer only to those hypothetical situations designed to generate intuitions about the natural world, in other words to scientific thought experiments. Likewise, others, among the fiercest critics of the epistemological role of thought experiments, have always rescued scientific thought experiments, specifically within physics. For example, both *Hull* [21.90] and *Snooks* [21.78] limit the power of thought experiments to well-articulated scientific fields, that is physics. Generally, there is a sharp scepticism about philosophical thought experiments.

#### The Status of Philosophical Thought Experimentation

*Rachel B. Cooper* [21.65] has pointed out that much of the analysis on thought experimentation is restricted to

scientific thought experiments, probably due to a strategy of caution (Sect. 21.2.3). In the last decade, the number of analyses considering only philosophical thought experiments has also grown – mainly due to the debate about the role of intuition in philosophy (see in the following). Although *Brown* does not deal with examples of philosophical thought experimentation, he is aware of this shortcoming and wonders whether scientific and philosophical thought experiment can be accommodated by a single theory [21.55, p. 28–31]. He also claims that perhaps this kind of conceptual analysis reveals a core common to philosophy and physics. If we were able to know more about such a common core, we would probably learn more about physics and philosophy, as well as thought experimentation (it is worth underlining that also in mathematics thought experimentation seems an important, and perhaps the only, form of experimentation; on the topic see, for example, [21.42], contributions in [21.13, 193–195]).

*Cooper* claims that often what distinguishes scientific from philosophical works are only the journals in which they are published. But papers on scientific thought experiments, such as Schrödinger's cat (Sect. 21.4.1), can be found both in scientific and philosophical journals. Similarly, a philosophical thought experiment such as Searle's Chinese room (Sect. 21.2.3) is tackled by philosophers, as well as by psychologists. "It is hard to distinguish science from philosophy and even harder to distinguish philosophical from scientific thought experiments" [21.65, p. 329]. *Cooper* suggests that a comprehensive analysis of thought experimentation cannot avoid to give an account of both philosophical and scientific thought experiments.

Besides *Cooper*, other scholars have tried, more or less extensively, to analyze both scientific and philosophical thought experiments, without raising a barrier between them [21.52, 64, 79, 82, 118, 141, 152, 161].

Arguably science and philosophy are intertwined, but one might think that the latter is more hostage to speculation and boundless imagination than the former. Precisely, for this reason, *Kathleen Wilkes* maintains that a good thought experimenter should envisage a scenario not too far from reality and specify all conditions relevant to its understanding ([21.87, p. 9]; see also [21.28]). Thus, she considers thought experimentation fruitful only in the scientific domain, because the latter, contrary to the philosophical domain, cannot deviate too much from reality and must invoke a type of thought experimentation more akin to real experimentation (this point brings us back to the issue about the biased continuity between real and thought experimentation see Sect. 21.3.1; see recent discussion in [21.196] about the idea that also thought experimentation in science is weakened by being dependent on imagination –

Sect. 21.5). According to *Wilkes*, there are two conditions for any experimenter (thought or real): first, to aim at testing a theory by varying key parameters and maintaining constant other relevant parameters (Sect. 21.3.3) and, second, not to violate natural laws. This second condition should further distinguish scientific from philosophical thought experiments. On this condition, however, the philosophical thought experiment of the brain in the vat [21.2] would be acceptable, since it is not obviously nomologically impossible, whereas the thought experiment of Einstein chasing a light beam would not be (see [21.164] for an in-depth analysis of this thought experiment). Moreover, as rightly stressed by *Brown* [21.55, p. 30–31]

"Too often thought experiments are used to find the laws of nature themselves; they are tools for unearthing the theoretically or nomologically possible. Stipulating the laws in advance and requiring thought experiments not to violate them would simply undermine their use as powerful tools for the investigation of nature"

See also what will be said about *Cooper* on this point Sect. 21.5.2).

Philosophical thought experiments are generally pictured by their detractors as fairy tales, which do not deserve to be taken seriously. The underlining idea seems to be that philosophy is too much prone to conceptual ruminations, involving idealization and approximation, and based on a methodology less strict than the scientific one. Philosophical thought experimentation would be paradigmatic of these flaws. Following *Hull's* discussions ([21.89, 90]; Sect. 21.3.1), which sum up criticisms against philosophical thought experiments very well, there are four negative aspects, which make philosophical thought experimentation less effective than scientific thought experimentation.

First, philosophical thought experiments lack well-defined theoretical frameworks. According to *Hull* [21.90, pp. 432, 434 and 438] this is the fundamental difference between philosophical and scientific thought experiments and, probably, the reason for the disparity (at the time of his writings) between many excellent analyses of scientific thought experiments and poor accounts of philosophical thought experiments. *Hull* thinks that thought experiments made within analytic philosophy well exemplify the lack of a theoretical framework that allows us to set up the imagined scenario. Provocatively, he writes: "If no such context exists, philosophers need to construct one. [...] If Jane Austen can do it, so can Hilary Putnam" [21.90, p. 434]. If a theoretical and technical background is missing, as much as one tries to refine the details of the given thought experiment, it will remain hopelessly incom-

plete (Sect. 21.3.3) and of poor cognitive value. Furthermore, without a reliable theoretical background, the usefulness of philosophical thought experimentation is also undermined, in so far as thought experimentation cannot exploit a fruitful interdependence between observations and theories [21.89, p. 311].

Second, philosophical thought experiments are used in order to justify or provide evidence in favor of theoretical hypotheses, though they should be used only for descriptive purposes ([21.89, pp. 315–316]; see also [21.90, p. 438 and p. 453]). According to Hull, the fact that philosophical thought experimentation relies more on common sense than on scientific data weakens its justificatory power and is also the reason why they cannot offer the same degree of technical specificity as scientific thought experimentation and real experimentation.

Third, Hull maintains that, contrary to real experimentation, thought experimentation requires a theory of conceivability as a vehicle for possibility. Thus, thought experimentation should adopt a strong standard of conceivability. Unfortunately, “Too often, the decisions that philosophers make rest heavily on intuitions about what sounds right” [21.90, p. 435]. In a nutshell, we settle for weak requirements for assessing the plausibility of the conclusions reached via thought experimentation (there is a vast philosophical literature on the link between conceivability and possibility; see, for example, contributions in [21.197], which also touch upon the issue about thought experimentation; see also [21.198]).

Finally, misleading intuitions seriously undermine the efficacy of thought experimentation. These intuitions are culturally variable, being dependent on our cultural beliefs. The latter can help us in exploring possible worlds, but also be narrow-minded and inhibit innovation [21.90, p. 431 and p. 446]. It does emerge, from the catastrophic picture of the critics about philosophical thought experiments, that the latter are a source of, in Hull words, “conceptual morass” [21.89, p. 315], rather than a prelude to its remediation.

The lack of a strong standard of conceivability, the contextual vagueness, and the consequent paucity of interrelation between empirical and theoretical data are errors that can be traced to a single source: misleading intuitions. Hence, criticisms against philosophical thought experimentation can be reduced to 2. Philosophical thought experimentation, on the one hand, relies on questionable intuitions and, on the other hand, purports to bring evidence in defence of a philosophical claim or theory. Without calling into question, the plausibility of such a view, a further difficulty arises from the fact that the meaning of the term *intuition* is not crystal clear. There is not a consensus either on what intuitions are or on what we can reasonably ex-

pect from them. This is so not only in the debate on thought experimentation, but also in the specific debate about the nature of intuitions (a good starting point on the topic are the contributions in [21.199]; see also Cappelen [21.200], Chudnoff [21.201] and Booth and Rowbottom [21.202]).

### Intuitions and Thought Experimentation

Intuitions seem to be an integral part in the processes of rational choice. Psychology and related disciplines have been investigating the formation and variation of our daily choices for a long time [21.203]. The picture that has emerged is that our decisions are highly sensitive to many elements which, at first sight, appear irrelevant, such as the framing of the context. Similar considerations seem to apply to intuitions as well. This line of research is a warning to a standard philosophical practice that uses intuitions generated in response to thought experiments as evidence in the assessment of a philosophical thesis (e.g., in ethics – see, [21.204–206]). The lesson would be that a more rigorous program, whose goal is to observe responses obtained via thought experiments and to study the nature of the intuitions involved, is needed. A new philosophical movement known as *Experimental Philosophy* (or *X-Phi*) aims precisely at meeting this challenge, by making use of the critical methods proper to social experimental psychology (for an introduction to X-phi, see [21.207] and contributions in [21.208]).

Note that the methodology of the experimental philosophers has been highly criticized (e.g., [21.209] and [21.210]), *Williamson* is also sceptical about the role of intuitions in thought experimentation – [21.119, 211, 212] – without being sceptical about thought experimentation itself like [21.213]). Leaving aside such critiques, X-Phi studies have shown that some philosophical thought experiments typically considered as universally acceptable (e.g., Gettier’s cases, *Keith Lehrer’s* Mr. Truetemp thought experiment [21.214], Putnam’s brain in a vat, *Saul Kripke’s* thought experiment on Gödel and Schmidt – [21.215]) evoke variable intuitions both *inter-* and *intra-*subjectively [21.216, 217]. The fact that thought experiments produce unstable intuitions makes thought experimentation itself shaky. However, thought experimentation in the scientific domain is commonly considered efficient. Therefore, it is legitimate to ask whether only philosophical thought experiments evoke poor intuitions and, if not, whether scientific thought experiments have other resources in order to make their intuitions more useful.

Some contemporary philosophers have explicitly pointed out that all thought experiments, both philosophical and scientific, evoke and make use of intuitions. In *Brown’s* account, for example, the state of

seeing the laws of nature is interpreted in terms of having intuitions [21.55, 72]. On this view, in Galileo's thought experiment (Sect. 21.1.1) the desired scientific conclusion comes as a result of our having the intuition that the two bodies fall at the same speed. However, it seems that in the scientific domain as well, thought experimentation can elicit misleading and unreliable intuitions (e.g., in EPR thought experiment, which is generally interpreted as a failed thought experiment – Sect. 21.3.3). Still it is an open question whether and how these intuitions can be properly used in the scientific domain.

One way to answer this question is to argue that philosophical and scientific thought experiments do not involve the same type of intuitions. Bealer [21.192] seems to hold this view. He distinguishes between rational and physical intuitions. The former would be *sui generis* intellectual seemings and would arise when considering the (logical or metaphysical) possibility of an imagined scenario or the applicability of a given concept to such a scenario. The author gives as an example of this type of intuition Gettier's cases (Sect. 21.1.4), which trigger two rational intuitions: A first intuition confirms that the case is possible, while a second intuition that we cannot ascribe to the imagined subject a state of knowledge. Physical intuitions deal with what would happen if the given imagined scenario were actual, rather than with its plausibility. Newton's rotating bucket thought experiment (Sect. 21.1.3) would exemplify this type of intuition, since in this case a physical intuition has to answer the question: "Would water creep up the side of the bucket (assuming that the physical laws remained unchanged)?" [21.192, p. 207]. According to Bealer, another feature distinguishes rational from physical intuitions: only the former present themselves as necessary. As Bealer would say, necessarily if a subject S intuits that the given imagined scenario is not a case of knowledge, it seems to S that the given imagined scenario is not a case of knowledge and also that necessarily the given imagined scenario is not a case of knowledge. By contrast, it does not seem that the water must crawl up the side of the bucket, though it is possible. Finally, Bealer claims that the expression *thought experiment* should be used to refer only to hypothetical situations that generate physical intuitions – i. e., to scientific thought experiments (mathematics and logic excluded).

Beyond the plausibility of a distinction between physical and rational intuitions drawn on the distinction between possibility and necessity, its relevance for a corresponding distinction between scientific

and philosophical thought experiments is questionable, since the alleged self-evidence of intuitions produced by philosophical thought experimentation has been seriously challenged.

Daniel Dennett [21.218] defined thought experiments as *intuition pumps*. Generally, in the literature on thought experiments, this expression is interpreted in a negative sense. Indeed, Dennett does not consider (at least philosophical) thought experiments highly. However, the philosopher seems to acknowledge that thought experiments can be useful when he writes that [21.218, p. 18]

"Philosophy with intuition pumps is not science at all, but in its own informal way it is a valuable – even occasionally necessary – companion to science."

Following Peter Swirski [21.219], it can be argued that the fact that thought experiments, both scientific and philosophical, are intuition pumps, and that these intuitions are unstable, is not negative per se. The epistemic force of a thought experiment seems precisely to arise from the fact that it depicts an exceptional case and forces us to account for the latter. Perhaps, the problem lies in an overestimation of what "can reasonably be expected of such experiments" [21.219, p. 105]. For example, it might be an exaggeration to consider thought experimentation a canonical procedure of justification [21.192], as if a single thought experiment could lead us to accept or to reject a theory. After all, even real experiments seem not capable of doing so much.

It is also possible to argue that it is a mistake "to describe the sort of knowledge involved in these thought experiments as intuitions" [21.56, p. 300]. The idea would be that intuitions are a component in the cognitive process of thought experimenting, more than its upshot. It might be even argued that intuitions are dispensable in thought experimenting [21.31, 119, 211, 212, 220]. According to many authors [21.71, 72, 82, 137, 161, 204, 221], however, intuitions play a crucial role in thought experimenting. Through these intuitions, in conjunction with other components (e.g., theoretical assumptions, empirical data), thought experiments lead us to acquire knowledge. The power of thought experiments would rely on optimizing the combination between data, theories, and intuitions. It can be argued that scientific and philosophical thought experimentations are functionally similar: The latter have the same potential as the former in order to make interact data, theories, and intuitions [21.152].



## 21.5 How Do Thought Experiments Achieve Their Function?

The discussion on the intuitions involved and generated by thought experimentation (Sect. 21.4.3) highlights another important issue in the debate on thought experimentation, namely the question about how the latter obtains its results. This question is intimately connected to the cognitive side of thought experimentation. In what follows, first, I shall outline the motivation for a cognitive approach to thought experimentation and briefly review what has been said about the cognitive underpinnings of thought experimentation (Sect. 21.5.1). Second, I shall focus on the role played by imagination in thought experimentation, since almost all scholars agree in thinking that imaginative capacities are recruited by thought experimenters (Sect. 21.5.1). Third, I shall draw on issues connected to the latter, namely the narrative dimension of thought experimentation.

### 21.5.1 A Cognitive Approach to Thought Experimentation

Most analyses in the debate have been primarily concerned with epistemological issues that aim to analyze thought experiments with respect to their outputs, more than their cognitive underpinnings. A cognitive approach would, however, be worth pursuing, because we could consider not only the result of a thought experiment, but also what happens in the head of someone performing a thought experiment (on the advantages and disadvantages of the study of thought experimentation via cognitive sciences see the debate between *Stuart* [21.222] and *Thagard* [21.99]).

*Mason Myers* [21.185] complained about both the lack of a deep investigation of the basis of thought experimental reasoning and the epistemic aspects of (philosophical) thought experiments. However, an epistemological approach to thought experiments is not cognitive, or not necessarily so. The difference between these two ways of studying thought experiments lies in the specific issues addressed, as well as in the fact that while the epistemological approach is generally normative, the cognitive approach is more descriptive. A fine-grained and comprehensive analysis of thought experiments should acknowledge the differences between these two approaches and pursue them together insofar as they are complementary [21.223].

In the literature, there have been several attempts to describe the stages of a thought experiment or how it works. For instance, *Reiner* and *Gilbert* [21.148] argue that there are six stages to thought experimenting:

1. A problem or a hypothesis is stated
2. An imaginary world that contains objects and laws is made up
3. The thought experiment is designed
4. The thought experiment is run
5. *Observations* are made (i. e., an outcome produced with the use of the laws of logic) and
6. Conclusions are drawn (see also [21.52, 86] and more recent contributions in [21.15]).

However, it is not always clear if these analyses focus on what really happens in the heads of thought experimenters or if they are about the argumentative structure of thought experiments. Indeed, it seems that thought experiments are pieces of reasoning, so that it should be possible to organize them into a premise-conclusion structure, involving certain inferential rules (Sect. 21.3.2).

Anyway, the question of how thought experiments fulfil their functions has led many to tackle the issue about the kind of reasoning underlying thought experimentation. On this issue, authors disagree and have appealed to different, but sometimes compatible, kinds of reasoning: hypothetical [21.38], counterfactual [21.77, 119], deductive [21.115, 117, 134], inductive [21.115, 117], abductive [21.103], simulative model-based [21.73], propositional [21.119], nonpropositional [21.73, 136], and heuristic (e.g., *De Mey* [21.224], who pleads for an approach who integrates both the heuristic value and the demonstrative force of thought experiments).

A related debate concerns the role played by intuitions in thought experimentation (Sect. 21.4.3). *Jeanne Peijnenburg* and *David Atkinson* claimed that, although there is not a unanimous definition of what thought experiments are, there is unanimity about what they should do, that is to give “a sudden and exhilarating insight” ([21.79, p. 306]; see also the definition in the *Encyclopedia of Cognitive Science* – [21.81]). Indeed, many philosophers in the debate on thought experiments have claimed that intuitions are an important component of the thought experimental process, as well as the type of its outcome. However, the precise role played by intuitions in thought experimentation is an open question. As previously stressed, while some authors have argued that intuitions cannot explain by themselves the epistemic role of thought experiments [21.56], others have based their scepticism about the thought experimental practice precisely on the fact that they mainly involve (deceptive) intuitions [21.79, 89, 90, 99].

Despite this disagreement, almost all authors involved in the debate on thought experiments agree in considering thought experiments as epistemic tools, which involve imagination in order to provide insights on a certain hypothesis or theory (see the definition in the *Encyclopedia of Cognitive Science* [21.81]).

### 21.5.2 Imagination and Thought Experimentation

*Mach* [21.28, 83, 168] was the first to argue that imagination plays a pivotal role in thought experimentation. According to him, performing a thought experiment is to “combine circumstances” in imagination [21.28, p. 452]. Some passages of his writings have led authors to maintain that Mach conceived imagination as visualization. Among these authors is *Gendler* [21.145], who attributes to Mach the idea that it is visual imagery (i. e., visual imagination) that is primarily at work in thought experimenting (see also [21.52]; for a critique [21.223]). *Gendler* herself has tried to pursue Mach’s approach. She establishes a link between research of cognitive scientists and philosophers on visual imagery and the analysis of *Stevin’s* thought experiment on the inclined plan (Sect. 21.1.2), and finds that at least in some thought experiments the role of visual imagery is epistemically crucial (see also [21.105] on this point).

It is possible to consider the *model-based approach*, which calls on the literature of model-based reasoning in cognitive science, as belonging to the *Machian tradition* too, that is, they pursue Mach’s aim of analyzing thought experiments with the help of a psychological theory (for a detailed analysis, see [21.86, Chap. 4]). Among the authors advocating this approach [21.85, 132, 225, 226], three are to be considered as its main developers, namely *Miščević* [21.137, 149], *Nersessian* [21.73, 84, 135, 227] and *Cooper* [21.65]. These authors agree in maintaining that in thought experimentation we gain new knowledge through manipulating a model. They have, however, advanced different theses pivoting on different notions of model.

*Miščević* and *Nersessian* appeal to the cognitive literature concerning *mental modeling*, and more specifically to the notion of the mental model proposed by *Philip Johnson-Laird* [21.228, 229]. In a nutshell, a mental model is a structure stored in short or long term memory and it is defined by cognitive scientists as a third type of mental representation, half way between propositional and pictorial. Indeed, mental models are structurally analogous to that which they represent, but not all such models can be visualized. This is clearly seen in *Nersessian’s* account [21.86], whereas *Miščević* holds a more pictorialist view about mental models. Indeed, *Miščević* claims that the mental model

is a “quasi-spatial picture” and has a “concrete and quasi-spatial character” [21.149, p. 220].

By contrast, *Nersessian* argues that the mental model manipulated in thought experimentation is neither a picture in the head nor a linguistic representation. Following *Johnson-Laird*, she maintains that it is rather a structural analog of the situation depicted in the thought experimental narrative [21.73, p. 297]. Nevertheless, *Nersessian* highlights the role of nonpropositional representations much more than *Miščević*: In her view the reasoning proper to thought experiments is entirely rather than partially nonpropositional. For that reason, *Nersessian* maintains that deductive and inductive inferences do not have a central part in thought experimentation.

*Cooper* disagrees with both *Miščević* and *Nersessian* and holds a much more liberal view. On the one hand, she maintains that the hypothesis of mental models is debatable as it is based on contestable empirical data. On the other hand, she argues that [21.65, p. 341]

“whether the thought experimenter reasons through the situation via manipulating a set of propositions, or a mental picture, or even plasticine characters makes no difference.”

For *Cooper*, a thought experimenter can manipulate physical models in addition to mental representations, and she can carry out deductive or inductive inferences, as well as diagrammatical ones. Another point of disagreement is that for *Nersessian* and *Miščević*, “models are restricted to simulating the way in which phenomena would unfold in the real world”. *Cooper* replies that “the thought experimenter may model a world in which some laws of nature are suspended or altered” ([21.65, p. 341] – see what has been said about *Wilkes* on this point Sect. 21.4.3).

Two other views that fall within the mental-model approach are worth mentioning. First, by following *Nersessian’s* account, *Michael Bishop* [21.132] has put forward a very liberal view according to which mental models of actualized experiments and at least some computer simulations count as thought experiments (see what has been said about numerical experiments Sect. 21.3.2). Second, *David Gooding* [21.85, 147] was in line with *Nersessian’s* approach too. However, he did not rely on the notion of a mental model and developed an embodied view on thought experiments where the bodily and visual components play a central role (see [21.230] for a more phenomenological and less naturalistic embodied approach on thought experimentation).

According to the *Machian tradition*, thought experiments are species of simulative-based reasoning. This

idea is implicit in Gendler's analysis, but it is only in the model-based approach that it is made fully explicit and linked to the notion of a (physical or mental) model. *John Zeimbekis* [21.231] has criticized simulationist approaches to thought experiments by arguing that we should distinguish between two kinds of mental simulation: mental–mental simulation and mental–physical simulation (see also the similar distinction between recreative and icastic imagination drawn in [21.223]). He claims that while only the latter is captured by mental models, the former is a source of epistemic bias, at least, for moral thought experiments. Zeimbekis grounds his argumentation on the literature in philosophy of mind about simulation theory. However, it is not entirely clear, whether he is dealing with high-level or low-level mental simulations. While the former comes at the personal level and can be interpreted as conscious imagination, the latter comes at the subpersonal level and is realized by mirroring processes, for example, the activation of mirror neurons in the observation mode [21.232, 233].

Imagination is often cited by all these authors, but it is not crystal clear how imagination is defined and what the link is between mental or physical models and imagination. More generally, the role of imagination in thought experiment is a controversial topic (see [21.196]; on imagination and thought experiments see also [21.234, 235]). Indeed, some authors, like *Gendler* [21.145], give it a central role, while others, such as *Norton* [21.115], maintain that thought experimenters could and should do without it, imagination being here a source of error (see also [21.54]). Moreover, an additional complication arises once we acknowledge that accounts of imagination provided by the cognitive literature have pointed out that imagination comes in many varieties, for instance there would be sensory and nonsensory forms of imagination (e.g., [21.232]). A closer look at the expressions used in the literature to describe thought experimentation suggests that most authors think of imagination as the means by which the thought experimenter gains access to a scenario which is not directly accessible to her senses: in thought experimenting she *quasi-observes*. For instance, *Brown* [21.55, 70–72] speaks of *seeing* the laws of nature and he claims that the pictorial and sensory aspects are essential to thought experimentation (see also [21.136] on that point). *Nersessian* [21.73] stresses that when we perform a thought experiment, we feel ourselves as observers. *Martin Cohen* takes the second rule of good thought experimenting to be that the thought experiment must be imaginable, that is, “the clearer the picture, the stronger the image, the better the experiment” [21.64, p. 106]. *Gooding* [21.85] claims that visualization is a necessary and sufficient condition

to most if not all thought experiments. *Norton* [21.115] admits that thought experiments involve visualization, though he denies its epistemic role. Thought experimentation, thus, seems to involve a sensory – specifically, visual – variety of imagination. Although they are in the minority, other authors have suggested that non-sensory forms of imagination may be necessary to the thought experimenter, like supposition [21.236] or conceiving [21.237]. Indeed, Mach himself seems to have given to imagination in all its forms a role in thought experimenting [21.223].

### 21.5.3 The Narrative Dimension of Thought Experimentation

The model-based approach has underlined a rather neglected aspect of thought experimentation, namely its narrative dimension [21.73, 85]. Thought experiments are extremely important because they are intentional products related to the sharing and the spreading of knowledge. Moreover, they are publicly presented to different audiences through narratives [21.219]. In disagreement with Norton, *Nersessian* has stressed that the aesthetic details in thought experimental narratives are not simply rhetorical, but “serve to reinforce crucial aspects of the [thought] experiment” ([21.73, p. 296] – she also sees a parallel between thought and real experiments, since also the latter when published are presented in a narrative form). Still, according to *Lawrence Souder* [21.238], even in *Nersessian's* account the role played by the narrative aspect of thought experimentation is underestimated. The same holds for other views that deny to thought experimentation a life of its own (the reference is mainly to [21.74], as we have seen before Sect. 21.3.3).

The narrative dimension of thought experimentation has led some authors to conclude that the reasoning underlying thought experimentation is closely related to the one used in the consumption of fiction [21.73, 149, 186, 234]. Some have proposed to consider thought experiments as a genre, like science fiction [21.239]. This view is in line with *David Davies' one*. Indeed, he argues [21.240, 241] that both philosophical and scientific thought experimentation meet two necessary and sufficient conditions for the fictionality of a narrative. That is, first, they involve to make believe, rather than to believe, that the state of affairs described holds (this point brings us back to the issue about the role played by imagination in thought experimentation, since make-believing is a form of imagining, see [21.234, 242], Sect. 21.5.2); second, they involve a narrative constrained by some specific purpose “such as entertaining or perhaps instructing readers in certain specific ways” ([21.241]; in [21.240] he specifies that

the imaginary world should not be constrained by actual events).

If, on the one hand, we can put thought experimentation on the level of literary fiction, on the other hand, we can also do it the other way around and put the latter on the level of the former. Fictions themselves can be seen as thought experiments aiming at enriching the subject's knowledge via *journeys* in more or less far possible worlds [21.219, 243]. Precisely, on this basis, both Carroll [21.244] and Elgin [21.245, 246] have tried to defend literary cognitivism, that is, the view according to which fictional narratives can be a source

of knowledge or understanding of the real world (in [21.246], the author also argues that the process of *exemplification* is common to literary fiction, thought and real experimentation). Some caveats to this move have been raised by Davies ([21.241]; see also [21.80]), in particular when applied to films [21.247].

**Acknowledgments.** I am very grateful to Marco Buzoni, Jérôme Dokic, Yiftach Fehige, and Michael Stuart for helpful comments on earlier versions of this chapter. I would also like to thank the editor in charge of the part D, Nora Schwartz, for her support.

## References

- 21.1 Plato: *The Republic* (Cosimo, New York 2008), translated by Benjamin Jowett
- 21.2 H. Putnam: *Reason, Truth and History* (Cambridge University Press, Cambridge 1981)
- 21.3 J. Locke: *An Essay Concerning Human Understanding* (Thomas Dring Samuel Manship, London 1690/ 1694)
- 21.4 G. Galilei: *Discorsi e Dimostrazioni Matematiche Intorno a Due Nuove Scienze* (Louis Elsevier, Leida 1638)
- 21.5 E. Condillac: *Traité des Sensations* (Chez de Bure, London/Paris 1754)
- 21.6 I. Kant: Von dem ersten Grunde des Unterschieds der Gegenden im Raume. In: *Vorkritische Schriften*, ed. by A. Buchenau (Bruno Cassirer, Berlin 1768) pp. 375–383
- 21.7 C. Darwin: *The Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life* (John Murray, London 1859)
- 21.8 H. Poincaré: *La Science et l'hypothèse* (E. Flammarion, Paris 1908)
- 21.9 A. Einstein, L. Infeld: *The Evolution of Physics. The Growth of Ideas from Early Concepts to Relativity and Quanta* (Simon Schuster, New York 1938)
- 21.10 W. Heisenberg: *Physikalische Prinzipien der Quantentheorie* (Hirzel, Leipzig 1930)
- 21.11 T. Burge: Individualism and the mental, *Midwest Stud. Philos.* **4**, 73–121 (1979)
- 21.12 J. Searle: Minds, brains, and programs, *Behav. Brain Sci.* **3**, 417–457 (1980)
- 21.13 T. Horowitz, G. Massey (Eds.): *Thought Experiments in Science and Philosophy* (Rowman Littlefield, Lanham 1991)
- 21.14 R. Casati, A. Jacomuzzi, P. Kobau (Eds.): *Esperimenti Mentali* (Rosenberg Sellier, Turin 2009)
- 21.15 K. Ierodiakonou, S. Roux (Eds.): *Thought Experiments in Methodological and Historical Contexts* (Brill, Leiden–Boston 2011)
- 21.16 M. Frappier, L. Meynell, J.R. Brown (Eds.): *Thought Experiments in Philosophy, Science, and the Arts* (Routledge, London, New York 2013)
- 21.17 M.G. Tarallo, T. Mazzoni, N. Poli, D.V. Sutyryn, X. Zhang, G.M. Tino: Test of Einstein Equivalence Principle for 0-spin and half-integer-spin atoms: Search for spin-gravity coupling effects, *Phys. Rev. Lett.* **113**, 023005–1–023005–5 (2014)
- 21.18 Stevinus: *De Staticae Elementis* (Ioannis Patii, Leiden 1605/ 1608)
- 21.19 I. Newton: *Philosophiae Naturalis Principia Mathematica* (Joseph Streater, London 1687)
- 21.20 E. Gettier: Is justified true belief knowledge?, *Analysis* **23**, 121–123 (1963)
- 21.21 B. Russell: *Human Knowledge: Its Scope and Its Limits* (Allen Unwin, London 1948)
- 21.22 M. Cohen: *101 Philosophy Problems* (Routledge, London, New York 1999)
- 21.23 H. Putnam: Meaning and Reference, *J. Philos.* **70**, 699–711 (1973)
- 21.24 H. Putnam: Meaning of “Meaning”, *Minnesota Stud. Philos. Sci.* **7**, 131–193 (1975)
- 21.25 F. Jackson: Epiphenomenal qualia, *Philos. Q.* **32**, 27–36 (1982)
- 21.26 P. Ludlow, Y. Nagasawa, D. Stoljar (Eds.): *There's Something About Mary. Essays on Phenomenal Consciousness and Franck Jackson's Knowledge Argument* (MIT Press, Cambridge, London 2004)
- 21.27 J.R. Brown, Y. Fehige: Thought experiments. In: *The Stanford Encyclopedia of Philosophy*, ed. by E.N. Zalta, <http://plato.stanford.edu/archives/fall2011/entries/thought-experiment/> (2014)
- 21.28 E. Mach: Über Gedankenexperimente, *Z. Phys. Chem. Unterr.* **10**, 1–5 (1896), Translated by W.O. Price, S. Krimsky: On thought experiments, *Philosophical Forum* **4**(3), 446–457 (1973)
- 21.29 H.C. Ørsted: *Første Indledning til den Almindelige Naturlære* (J.S. Schultz, Copenhagen 1811)
- 21.30 C. Schildknecht: *Philosophische Masken: Literarische Formen der Philosophie bei Platon, Descartes, Wolff und Lichtenberg* (Metzler, Stuttgart 1990)
- 21.31 U. Kühne: *Die Methode des Gedankenexperimentes* (Suhrkamp, Frankfurt 2005)
- 21.32 J. Witt-Hansen: H.C. Ørsted, Immanuel Kant, and the thought experiment, *Danish Yearb. Philos.* **13**, 48–65 (1976)
- 21.33 Y. Fehige, M.T. Stuart: On the origins of the philosophy of thought experiments: The forerun,

- 21.34 Perspect. Sci. **22**, 179–220 (2014)
- 21.34 A.S. Moue, K.A. Masavetas, H. Karayianni: Tracing the development of thought experiments in the philosophy of natural sciences, *J. Gen. Philos. Sci.* **37**, 61–75 (2006)
- 21.35 D. Cohnitz: Ørsted's Gedankenexperiment: Eine Kantianische Fundierung der Infinitesimalrechnung? Ein Beitrag zur Begriffsgeschichte von "Gedankenexperiment" und zur Mathematikgeschichte des frühen 19. Jahrhunderts, *Kant-Studien* **99**, 407–433 (2008)
- 21.36 M. Buzzoni: *Thought Experiment in the Natural Sciences: An Operational and Reflexive-Transcendental Conception* (Königshausen Neumann, Würzburg 2008)
- 21.37 S. Roux: Introduction: The emergence of the notion of thought experiments. In: *Thought Experiments in Methodological and Historical Contexts*, ed. by K. Ierodiakonou, S. Roux (Brill, Leiden-Boston 2011) pp. 1–33
- 21.38 N. Rescher: Thought experimentation in presocratic philosophy. In: *Thought Experiments in Science and Philosophy*, ed. by T. Horowitz, G. Massey (Rowman Littlefield, Lanham 1991) pp. 31–42
- 21.39 N. Rescher: *What If?: Thought Experimentation in Philosophy* (Transaction Publishers, New Brunswick 2005)
- 21.40 A. Irvine: Thought experiments in scientific reasoning. In: *Thought Experiments in Science and Philosophy*, ed. by T. Horowitz, G. Massey (Rowman Littlefield, Lanham 1991) pp. 149–166
- 21.41 K. Ierodiakonou: Ancient thought experiments: A first approach, *Ancient Philos.* **25**, 125–140 (2005)
- 21.42 I. Lakatos: *Proofs and Refutations. The Logic of Mathematical Discovery* (Cambridge University Press, Cambridge 1976)
- 21.43 M. Stöltzner: The dynamics of thought experiments – Comment to Atkinson. In: *Observation and Experiment in the Natural and Social Sciences*, ed. by M. Galavotti (Kluwer Academic Publishers, Dordrecht 2003) pp. 243–258
- 21.44 P. King: Mediaeval thought-experiments: The metamethodology of mediaeval science. In: *Thought Experiments in Science and Philosophy*, ed. by T. Horowitz, G. Massey (Rowman Littlefield, Lanham 1991) pp. 43–64
- 21.45 D. Perler: Thought experiments: The methodological function of angels in late medieval epistemology. In: *Angels in Medieval Philosophical Inquiry*, ed. by I. Iribarren, M. Lenz (Ashgate, Aldershot 2008) pp. 143–153
- 21.46 C. Grellard: Thought experiments in late medieval debates on atomism. In: *Thought Experiments in Methodological and Historical Contexts*, ed. by K. Ierodiakonou, S. Roux (Brill, Leiden–Boston 2011) pp. 65–82
- 21.47 G. Prudovsky: The confirmation of the superposition principle: The role of a constructive thought experiment in Galileo's Discorsi, *Stud. Hist. Philos. Sci.* **20**, 453–468 (1989)
- 21.48 D. Atkinson, J. Peijnenburg: Galileo and prior philosophy, *Stud. Hist. Philos. Sci.* **35**, 115–136 (2004)
- 21.49 P. Palmieri: "Spuntur lo scoglio più duro": Did Galileo ever think the most beautiful thought experiment in the history of science?, *Stud. Hist. Philos. Sci.* **36**, 305–322 (2005)
- 21.50 J. Daiber: *Experimentalphysik des Geistes: Novalis und das Romantische Experiment* (Vadenhoeck Ruprecht, Göttingen 2001)
- 21.51 Y. Fehige: Poems of productive imagination: Thought experiments, theology, and science in Novalis, *Neue Z. Syst. Theol. Religionsphilos.* **55**, 54–83 (2013)
- 21.52 R. Sorensen: *Thought Experiments* (Oxford University Press, Oxford 1992)
- 21.53 A. Meinong: *Über die Stellung der Gegenstandstheorie im System der Wissenschaften* (Voigtländer, Leipzig 1907)
- 21.54 P. Duhem: *La Théorie Physique: Son Objet, sa Structure* (Vrin, Paris 1914)
- 21.55 J.R. Brown: *The Laboratory of the Mind: Thought Experiments in the Natural Sciences* (Routledge, London 1991)
- 21.56 A. Bokulich: Rethinking thought experiments, *Perspect. Sci.* **9**, 285–307 (2001)
- 21.57 M. Buzzoni: *Esperimento ed Esperimento Mentale* (FrancoAngeli, Milano 2004)
- 21.58 K. Popper: On the use and misuse of imaginary experiments, especially in quantum theory. In: *The Logic of Scientific Discovery* (Hutchinson, London 1959) pp. 442–456
- 21.59 C. Hempel: Typological methods in the natural and the social sciences. In: *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science* (Free Press, New York 1965) pp. 155–171
- 21.60 A. Koyré: Galileo's treatise de motu gravium: The use and the abuse of imaginary experiment, *Rev. Hist. Sci. Paris.* **13**, 197–245 (1960)
- 21.61 T.S. Kuhn: A function for thought experiments. In: *L'aventure de la Science, Mélanges Alexandre Koyré*, ed. by I.B. Cohen, R. Taton (Hermann, Paris 1964) pp. 307–343
- 21.62 J.J. Thomson: A defense of abortion, *Philos. Public Aff.* **1**, 47–66 (1971)
- 21.63 D. Parfit: *Reasons and persons* (Clarendon Press, Oxford 1984)
- 21.64 M. Cohen: *Wittgenstein's Beetle and Other Classic Thought Experiments* (Blackwell, Oxford 2005)
- 21.65 R. Cooper: Thought Experiments, *Metaphilosophy* **36**, 328–347 (2005)
- 21.66 G. Boniolo: On a unified theory of models and thought experiments in natural sciences, *Int. Stud. Philos. Sci.* **11**, 121–142 (1997)
- 21.67 U. Gähde: Gedankenexperimente in Erkenntnistheorie und Physik: Strukturelle Parallelen. In: *Rationalität, Realismus, Revision*, ed. by J. Nida-Rümlin (de Gruyter, Berlin 2000) pp. 457–464
- 21.68 J. Norton: Thought experiments in Einstein's work. In: *Thought Experiments in Science and Philosophy*, ed. by T. Horowitz, G. Massey (Rowman Littlefield, Lanham 1991) pp. 129–148

- 21.69 R. Laymon: Thought experiments by Stevin, Mach and Gouy: Thought experiments as ideal limits and as semantic domains. In: *Thought Experiments in Science and Philosophy*, ed. by T. Horowitz, G. Massey (Rowman Littlefield, Lanham 1991) pp. 167–192
- 21.70 J.R. Brown: Thought experiments: A Platonic account. In: *Thought Experiments in Science and Philosophy*, ed. by T. Horowitz, G. Massey (Rowman Littlefield, Lanham 1991) pp. 119–128
- 21.71 J.R. Brown: Why thought experiments transcend experience. In: *Contemporary Debates in the Philosophy of Science*, ed. by C. Hitchcock (Blackwell, Oxford 2004) pp. 23–43
- 21.72 J.R. Brown: Peeking into Plato's Heaven, *Philos. Sci.* **71**, 1126–1138 (2004)
- 21.73 N.J. Nersessian: In the theoretician's laboratory: Thought experimenting as mental modelling. In: *PSA 1992*, ed. by D. Hull, M. Forbes, K. Okruhlik (Philosophy of Science Association, East Lansing 1993) pp. 291–301
- 21.74 I. Hacking: Do thought experiments have a life of their own? Comments on James Brown, Nancy Nersessian and David Gooding. In: *PSA 1992*, ed. by D. Hull, M. Forbes, K. Okruhlik (Philosophy of Science Association, East Lansing 1993) pp. 302–308
- 21.75 P. Humphreys: Seven theses on thought experiments. In: *Philosophical Problems of the Internal and External World: Essays on the Philosophy of Adolf Grunbaum*, ed. by J. Earman, A. Janis, J. Massey, N. Rescher (University of Pittsburgh Press/Universitätsverlag Konstanz, Pittsburgh/Konstanz 1993) pp. 205–227
- 21.76 T. Gendler Szabó: Galileo and the indispensability of scientific thought experiment, *Br. J. Philos. Sci.* **49**, 397–424 (1998)
- 21.77 E. Weber, T. De Mey: Explanation and thought experiments in history, *Hist. Theory* **42**, 28–38 (2003)
- 21.78 R. Snooks: Another scientific practice separating chemistry from physics: Thought experiments, *Found. Chem.* **8**, 255–270 (2006)
- 21.79 J. Peijnenburg, D. Atkinson: When are Thought Experiments Poor Ones?, *J. Gen. Philos. Sci.* **34**, 305–322 (2003)
- 21.80 G. McComb: Thought experiment, definition, and literary fiction. In: *Thought Experiments in Philosophy, Science, and the Arts*, ed. by M. Frappier, L. Meynell, J.R. Brown (Routledge, London, New York 2013) pp. 207–222
- 21.81 T. Gendler Szabó: Thought experiment. In: *Encyclopedia of Cognitive Science*, ed. by L. Nadel (New York/London, Nature/Routledge 2002) pp. 388–394
- 21.82 E. Brendel: Intuition pumps and the proper use of thought experiments, *Dialectica* **58**, 88–108 (2004)
- 21.83 E. Mach: *Erkenntnis und Irrtum* (Barili, Leipzig 1905)
- 21.84 N.J. Nersessian: How do scientists think? Capturing the dynamics of conceptual change in science. In: *Cognitive Models of Science*, ed. by R.N. Giere (University of Minnesota Press, Minneapolis 1992) pp. 3–44
- 21.85 D. Gooding: What is experimental about thought experiments? In: *PSA 1992*, ed. by D. Hull, M. Forbes, K. Okruhlik (Philosophy of Science Association, East Lansing 1993) pp. 280–290
- 21.86 S. Häggqvist: *Thought Experiments in Philosophy* (Almqvist Wiksel, Stockholm 1996)
- 21.87 K. Wilkes: *Real People: Personal Identity Without Thought Experiments* (Clarendon Press, Oxford 1988)
- 21.88 M. Bishop: Why thought experiments are not arguments, *Philos. Sci.* **66**, 534–541 (1999)
- 21.89 D. Hull: A Function for actual examples in philosophy of science. In: *What the Philosophy of Biology is: Essays Dedicated to David Hull*, ed. by M. Ruse (Kluwer Academic Publishers, Dordrecht 1989) pp. 309–321
- 21.90 D. Hull: That just don't sound right: A plea for real examples. In: *The Cosmos of Science: Essays of Exploration*, ed. by J. Earman, J.D. Norton (University of Pittsburgh Press, Pittsburgh 1997) pp. 430–457
- 21.91 S. Krimsky: The Nature and Function of "Gedankenexperimente" in Physics, Ph.D. Thesis (University of Michigan, Ann Arbor 1970)
- 21.92 H. Reichenbach: *Experience and Prediction. An Analysis of the Foundations and the Structure of Knowledge* (University of Chicago Press, Chicago 1938)
- 21.93 C. Daly: *An Introduction to Philosophical Methods* (Broadview Press, Peterborough 2010)
- 21.94 R. Arthur: Can thought experiments be resolved by experiment? The case of Aristotle's wheel. In: *Thought Experiments in Philosophy, Science, and the Arts*, ed. by M. Frappier, L. Meynell, J.R. Brown (Routledge, London, New York 2013) pp. 107–122
- 21.95 J. Fodor: On knowing what we would say, *Philos. Rev.* **73**, 198–212 (1964)
- 21.96 P. Feyerabend: *Against Method* (Verso, London 1978)
- 21.97 W.V. Quine: Review of identity and individuation, *J. Philos.* **69**, 488–497 (1972)
- 21.98 P. Thagard: *The Brain and the Meaning of Life* (Princeton University Press, Princeton 2010)
- 21.99 P. Thagard: Thought experiments considered harmful, *Persp. Sci.* **22**, 288–305 (2014)
- 21.100 M. Arcangeli: Il posto delle favole. In: *Rivista di Estetica, s.i. Esperimenti mentali*, Vol. 42, ed. by R. Casati, A. Jacomuzzi, P. Kobau (Rosenberg Sellier, Turin 2009) pp. 3–19
- 21.101 J. Lennox: Darwinian thought experiments: A function for just-so stories. In: *Thought Experiments in Science and Philosophy*, ed. by T. Horowitz, G. Massey (Rowman and Littlefield 1991) pp. 223–245
- 21.102 J. Lennox: Darwin's methodological evolution, *J. Hist. Biol.* **38**, 85–99 (2005)
- 21.103 L.S. Swan: Synthesizing insight: Artificial life as thought experimentation in biology, *Biol. Philos.* **24**, 687–701 (2009)

- 21.104 F. Doolittle: Craig venter's new life. The realization of some thought experiments in biological ontology. In: *Thought Experiments in Philosophy, Science, and the Arts*, ed. by M. Frappier, L. Meynell, J.R. Brown (Routledge, London, New York 2013) pp. 160–176
- 21.105 J. McAllister: Thought experiments and the belief in phenomena, *PSA 2002*, *Philos. Sci.* **71**, 1164–1175 (2004)
- 21.106 J. Maffie: “Just-so” stories about “inner cognitive Africa”: Some doubts about Sorensen's evolutionary epistemology of thought experiments, *Biol. Philos.* **12**, 207–224 (1997)
- 21.107 R. Sorensen: Precis of thought experiments, *Informal Log.* **17**(3), 385–387 (1995)
- 21.108 M. Bunzl: Bunzl on Sorensen's thought experiments, *Informal Log.* **17**(3), 389–393 (1995)
- 21.109 R. Feldman: Feldman on Sorensen's thought experiments, *Informal Log.* **17**(3), 394–398 (1995)
- 21.110 R. Sorensen: Sorensen's reply to Bunzl and Feldman, *Informal Log.* **17**(3), 399–405 (1995)
- 21.111 M. Buzzoni: Thought experiments from a Kantian point of view. In: *Thought Experiments in Philosophy, Science and the Arts*, ed. by M. Frappier, L. Meynell, J.R. Brown (Routledge, London, New York 2013) pp. 90–106
- 21.112 Y. Fehige: Experiments of pure reason. Kantianism and thought experiments in science, *Epistemologia Ital. J. Philos. Sci.* **35**, 141–160 (2012)
- 21.113 Y. Fehige: The relativized a priori and the laboratory of the mind: Towards a neo-kantian account of thought experiments in science, *Epistemologia Ital. J. Philos. Sci.* **36**, 55–63 (2013)
- 21.114 M. Buzzoni: On thought experiments and the Kantian a priori in the natural sciences: A reply to Yiftach J. H. Fehige, *Epistemologia Ital. J. Philos. Sci.* **36**, 277–293 (2013)
- 21.115 J. Norton: Why thought experiments do not transcend empiricism. In: *Contemporary Debates in the Philosophy of Science*, ed. by C. Hitchcock (Blackwell, Oxford 2004) pp. 44–66
- 21.116 J. Norton: Are thought experiments just what you thought?, *Canad. J. Philos.* **26**, 333–366 (1996)
- 21.117 J. Norton: On thought experiments: Is there more to the argument?, *PSA 2002*, *Philos. Sci.* **71**, 1139–1151 (2004)
- 21.118 T. Gendler Szabó: *Thought Experiment: On the Powers and Limits of Imaginary Cases* (Garland Press, New York 2000)
- 21.119 T. Williamson: *The Philosophy of Philosophy* (Blackwell, Malden 2007)
- 21.120 N. Gilbert, K. Troitzsch: *Simulation for the Social Scientist* (Open University Press, Philadelphia 1999)
- 21.121 C. Beisbart, J. Norton: Why Monte Carlo simulations are inferences and not experiments, *Int. Stud. Philos. Sci.* **26**, 403–422 (2012)
- 21.122 D. Dowling: Experimenting on theories, *Sci. Context* **12**(2), 261–273 (1999)
- 21.123 A. Barberousse, S. Franceschelli, C. Imbert: Computer simulations as experiments, *Synthese* **169**(3), 557–574 (2009)
- 21.124 E. Winsberg: A tale of two methods, *Synthese* **169**(3), 575–592 (2009)
- 21.125 W. Parker: Does matter really matter? Computer simulations, experiments and materiality, *Synthese* **169**(3), 483–496 (2009)
- 21.126 E.A. Di Paolo, J. Noble, S. Bullock: Simulation models as opaque thought experiments. In: *Proceedings of the Seventh International Conference on Artificial Life*, ed. by M.A. Bedau, J.S. McCaskill, N.H. Packard, S. Rasmussen (MIT Press, Cambridge 2000) pp. 497–506
- 21.127 S. Chandrasekharan, N.J. Nersessian, V. Subramanian: Computational modeling: Is this the end of thought experiments in science? In: *Thought Experiments in Philosophy, Science, and the Arts*, ed. by M. Frappier, L. Meynell, J.R. Brown (Routledge, London, New York 2013) pp. 239–260
- 21.128 F. Staudner: Virtuelle Erfahrung. Eine Untersuchung über den Erkenntniswert von Gedankenexperimenten und Computersimulationen in den Naturwissenschaften, Ph.D. Thesis (Friedrich Schiller Universität, Jena 1998)
- 21.129 M. Velasco: Experimentación y técnicas computacionales, *Theoria* **17**, 317–331 (2002)
- 21.130 J. Lenhard: Epistemologie der Iteration: Gedankenexperimente und Simulationsexperimente, *Dtsch. Z. Philos.* **59**, 131–154 (2011)
- 21.131 R. El Skaf, C. Imbert: Unfolding in the empirical sciences: Experiments, thought experiments and computer simulations, *Synthese* **190**, 3451–3474 (2013)
- 21.132 M. Bishop: An epistemological role for thought experiments. In: *Idealization in Contemporary Physics*, ed. by N. Shanks (Rodopi, Amsterdam, Atlanta 1998) pp. 19–33
- 21.133 M. Schulzke: Simulating philosophy: Interpreting video games as executable thought experiments, *Philos. Technol.* **27**, 251–265 (2014)
- 21.134 M. Bunzl: The logic of thought experiments, *Synthese* **106**, 227–240 (1996)
- 21.135 N.J. Nersessian: Thought experiments as mental modelling: Empiricism without logic, *Croat. J. Philos.* **7**(20), 125–161 (2007)
- 21.136 R. Arthur: On thought experiments as a priori science, *Int. Stud. Philos. Sci.* **13**, 215–229 (1999)
- 21.137 N. Mišćević: Modelling intuitions and thought experiments, *Croat. J. Philos.* **20**, 181–214 (2007)
- 21.138 W. Hopp: Experiments in thought, *Perspect. Sci.* **22**, 242–263 (2014)
- 21.139 J.R. Brown: Why empiricism won't work. In: *PSA 1992*, ed. by D. Hull, M. Forbes, K. Okruhlik (Philosophy of Science Association, East Lansing 1993) pp. 271–279
- 21.140 R. Urbaniak: “Platonic” thought experiments: How on Earth?, *Synthese* **187**, 731–752 (2012)
- 21.141 S. Häggqvist: A model for thought experiments, *Canad. J. Philos.* **39**, 56–76 (2009)
- 21.142 P.A. Schilpp (Ed.): *Albert Einstein: Philosopher-Scientist* (Open Court, La Salle 1949)
- 21.143 T. De Mey: The dual nature view of thought experiments, *Philosophica* **72**, 61–78 (2003)

- 21.144 J.Y. Goffi, S. Roux: On the very idea of a thought experiment. In: *Thought Experiments in Methodological and Historical Contexts*, ed. by K. Ierodiakonou, S. Roux (Brill, Leiden-Boston 2011) pp. 165–192
- 21.145 T. Gendler Szabó: Thought experiments rethought – And re-perceived, *Philos. Sci.* **71**, 1152–1164 (2004)
- 21.146 D. Gooding: *Experiment and the Making of Meaning: Human Agency in Scientific Observation and Experiment* (Kluwer Academic Publishers, Dordrecht, Boston 1990)
- 21.147 D. Gooding: The procedural turn; or, why do thought experiments work? In: *Cognitive Models of Science*, ed. by R.N. Giere (University of Minnesota Press, Minneapolis 1992) pp. 45–76
- 21.148 M. Reiner, J. Gilbert: Epistemological resources for thought experimentation in science learning, *Int. J. Sci. Educ.* **22**(5), 489–506 (2000)
- 21.149 N. Mišćević: Mental Models and Thought Experiments, *Int. Stud. Philos. Sci.* **6**, 215–226 (1992)
- 21.150 J.R. Brown: What Do we see in a thought experiment? In: *Thought Experiments in Philosophy, Science, and the Arts*, ed. by M. Frappier, L. Meynell, J.R. Brown (Routledge, London, New York 2013) pp. 53–68
- 21.151 A. Janis: Can thought experiments fail? In: *Thought Experiments in Science and Philosophy*, ed. by T. Horowitz, G. Massey (Rowman Littlefield, Lanham 1991) pp. 113–118
- 21.152 M. Arcangeli: Poveri esperimenti mentali. In: *Analisi. Annuario e Bollettino della Società Italiana di Filosofia Analitica SIFA*, ed. by R. Davies (Mimesis, Milano/Udine 2011) pp. 277–290
- 21.153 D. Rowbottom: Intuitions in science: Thought experiments as argument pumps. In: *Intuitions*, ed. by A. Booth, D. Rowbottom (Oxford Univ. Press, Oxford 2014) pp. 119–134
- 21.154 A. Einstein, B. Podolsky, N. Rosen: Can quantum-mechanical description of physical reality be considered complete?, *Phys. Rev.* **47**, 777–780 (1935)
- 21.155 D. Atkinson: Experiments and thought experiments in natural science. In: *Observation and Experiment in the Natural and Social Sciences*, ed. by M.C. Galavotti (Kluwer Academic Publishers, Dordrecht 2003) pp. 209–225
- 21.156 A. Pessin, S. Goldberg (Eds.): *The Twin Earth Chronicles: Twenty Years of Reflection on Hilary Putnam's "The Meaning of Meaning"* (M. E. Sharpe, New York 1996)
- 21.157 G.N. Schlesinger: The power of thought experiments, *Found. Phys.* **26**(4), 467–482 (1996)
- 21.158 G. Boniolo: *On Scientific Representations From Kant to a New Philosophy of Science* (Palgrave Macmillan, New York 2008)
- 21.159 A. Aspect, P. Grangier, G. Roger: Experimental realization of Einstein-Podolsky-Rosen-Bohm Gedankenexperiment: A new violation of Bell's inequalities, *Phys. Rev. Lett.* **49**(2), 91–94 (1982)
- 21.160 M. Rédei: Thinking about thought experiments in physics. Comment on 'Experiments and thought experiments by David Atkinson'. In: *Observation and Experiment in the Natural and Social Sciences*, ed. by M. Galavotti (Kluwer Academic, Dordrecht 2003) pp. 237–241
- 21.161 D. Cohnitz: Poor thought experiments? A comment on Peijnenburg and Atkinson, *J. Gen. Philos. Sci.* **37**, 373–392 (2006)
- 21.162 M. Dorato: Dalla freccia di Lucrezio all'ascensore di Einstein: Alcune considerazioni sul ruolo degli esperimenti mentali nella scienza. In: *Rivista di Estetica, s.i. Esperimenti mentali*, Vol. 42, ed. by R. Casati, A. Jacomuzzi, P. Kobau (Rosenberg Sellier, Turin 2009) pp. 21–37
- 21.163 J. McAllister: The evidential significance of thought experiment in science, *Stud. Hist. Philos. Sci. Part A* **27**, 233–250 (1996)
- 21.164 J. Norton: Chasing the light: Einstein's most famous thought experiment. In: *Thought Experiments in Philosophy, Science, and the Arts*, ed. by M. Frappier, L. Meynell, J.R. Brown (Routledge, London, New York 2013) pp. 123–140
- 21.165 J. Norton: Seeing the laws of nature, *Metascience* **3**, 33–38 (1993)
- 21.166 J.R. Brown: Counter thought experiments, *R. Inst. Philos. Suppl.* **61**(82), 155–177 (2007)
- 21.167 K. Ierodiakonou: Remarks on the history of an ancient thought experiment. In: *Thought Experiments in Methodological and Historical Contexts* ed. by K. Ierodiakonou, S. Roux (Brill, Leiden-Boston 2011) pp. 37–50
- 21.168 E. Mach: *Die Mechanik in ihrer Entwicklung historisch-kritisch dargestellt* (Brockhaus, Leipzig 1883)
- 21.169 S. Krimsky: The multiple-world thought experiment and absolute space, *Noûs* **6**(3), 266–273 (1972)
- 21.170 S. Krimsky: The use and misuse of critical Gedankenexperimente, *Z. Allg. Wissenschaftstheor.* **4**, 323–334 (1973)
- 21.171 H. Helm, J. Gilbert: Thought experiments and physics education – Part 1, *Phys. Educ.* **20**, 124–131 (1985)
- 21.172 H. Helm, J. Gilbert, M.D. Watts: Thought experiments and physics education – Part 2, *Phys. Educ.* **20**, 211–217 (1985)
- 21.173 S. Klassen: The science thought experiment: How might it be used profitably in the classroom?, *Interchange* **37**, 77–96 (2006)
- 21.174 A. Velentzas, K. Halkia, C. Skordoulis: Thought experiments in the theory of relativity and in quantum mechanics: Their presence in textbooks and in popular science books, *Sci. Educ.* **16**, 353–370 (2007)
- 21.175 M. Toscano: Thought experimentation and modelling in the science classroom, *AARE Conf. Proc.* (2007), TOS0741
- 21.176 R. Casati: *Dov'è il Sole di notte?* (Raffaello Cortina Editore, Milano 2013)
- 21.177 J.C. Maxwell: *Theory of Heat* (Longman, London 1871)
- 21.178 M. Johnston: Human beings, *J. Philos.* **84**, 59–83 (1987)
- 21.179 R. Casati, J. Dokic: *La Philosophie du Son* (Chambron, Nîmes 1994)



- 21.180 D. Borsboom, G.J. Mellenbergh, J. Van Heerden: Functional thought experiments, *Synthese* **130**, 379–387 (2002)
- 21.181 J. Reiss: Genealogical thought experiments in economics. In: *Thought Experiments in Philosophy, Science, and the Arts*, ed. by M. Frappier, L. Meynell, J.R. Brown (Routledge, London, New York 2013) pp. 177–190
- 21.182 N. Mišćević: Political thought experiments from Plato to Rawls. In: *Thought Experiments in Philosophy, Science, and the Arts*, ed. by M. Frappier, L. Meynell, J.R. Brown (Routledge, London, New York 2013) pp. 191–206
- 21.183 M.W. Jackson: The government of reason, *J. Value Inq.* **26**(2), 163–174 (1992)
- 21.184 M. Shumelda: At the limits of possibility: Thought experiment in quantum gravity. In: *Thought Experiments in Philosophy, Science, and the Arts*, ed. by M. Frappier, L. Meynell, J.R. Brown (Routledge, London, New York 2013) pp. 141–159
- 21.185 C. Myers: Analytical thought experiments, *Metaphilosophy* **17**, 109–118 (1986)
- 21.186 J. Ichikawa, B. Jarvis: Thought-experiment intuitions and truth in fiction, *Philos. Stud.* **142**, 221–246 (2009)
- 21.187 P. Engel: Philosophical thought experiments: In or out of the armchair? In: *Thought Experiments in Methodological and Historical Contexts*, ed. by K. Ierodiakonou, S. Roux (Brill, Leiden–Boston 2011) pp. 145–163
- 21.188 R. Sorensen: Veridical idealizations. In: *Thought Experiments in Philosophy, Science, and the Arts*, ed. by M. Frappier, L. Meynell, J.R. Brown (Routledge, London, New York 2013) pp. 30–52
- 21.189 T.S. Kuhn: Objectivity, value judgement, and theory choice. In: *The Essential Tension*, (University of Chicago Press, Chicago 1973) pp. 320–339
- 21.190 P.M. Churchland: The ontological status of observables: In praise of the superempirical virtues. In: *Images of Science: Essays on Realism and Empiricism*, ed. by P.M. Churchland, C.A. Hooker (University of Chicago Press, Chicago 1985) pp. 35–47
- 21.191 J.R. Griesemer, W.C. Wimsatt: Picturing weismannism: A case study of conceptual evolution. In: *What the Philosophy of Biology Is: Essays Dedicated to David Hull*, ed. by M. Ruse (Kluwer Academic Publishers, Dordrecht 1989) pp. 75–137
- 21.192 G. Bealer: Intuition and the autonomy of philosophy. In: *Rethinking Intuition. The Psychology of Intuition and Its Role in Philosophical Inquiry*, ed. by M. DePaul, W. Ramsey (Rowman Littlefield, Lanham 1998) pp. 201–239
- 21.193 K. Shrader-Frechette: Using a thought experiment to clarify a radiobiological controversy, *Synthese* **128**, 319–342 (2001)
- 21.194 V. Giardino: Sperimentare con i Triangoli. In: *Rivista di Estetica, s.i. Esperimenti mentali*, Vol. 42, ed. by R. Casati, A. Jacomuzzi, P. Kobau (Rosenberg, Sellier, Turin 2009) pp. 39–54
- 21.195 M. Buzzoni: On mathematical thought experiments, *Epistemol. Ital. J. Philos. Sci.* **34**, 61–88 (2011)
- 21.196 J. McAllister: Thought experiment and the exercise of imagination in science. In: *Thought Experiments in Philosophy, Science, and the Arts*, ed. by M. Frappier, L. Meynell, J.R. Brown (Routledge, London, New York 2013) pp. 11–29
- 21.197 T. Gendler Szabó, J. Hawthorne (Eds.): *Conceivability and Possibility* (Clarendon/OUP, New York/Oxford 2002)
- 21.198 D.H.M. Brooks: The method of thought experiment, *Metaphilosophy* **25**(1), 71–83 (1994)
- 21.199 M. DePaul, W. Ramsey (Eds.): *Rethinking Intuition. The Psychology of Intuition and Its Role in Philosophical Inquiry* (Rowman Littlefield, Lanham 1998)
- 21.200 H. Cappelen: *Philosophy Without Intuitions* (Oxford Univ. Press, Oxford 2012)
- 21.201 E. Chudnoff: *Intuition* (Oxford Univ. Press, Oxford 2013)
- 21.202 A. Booth, D. Rowbottom: *Intuition* (Oxford Univ. Press, Oxford 2013)
- 21.203 A. Tversky, D. Kahneman: The framing of decisions and the psychology of choice, *Science* **211**, 453–463 (1981)
- 21.204 T. Horowitz: Philosophical intuitions and psychological theory, *Ethics* **108**, 367–385 (1998)
- 21.205 S. Black, J. Tweedale: Responsibility and alternative possibilities: The use and abuse of examples, *J. Ethics* **6**, 281–303 (2002)
- 21.206 J.M. Doris, S.P. Stich: As a matter of fact: Empirical perspectives on ethics. In: *The Oxford Handbook of Contemporary Philosophy*, ed. by F. Jackson, M. Smith (OUP, Oxford 2005) pp. 114–152
- 21.207 J. Alexander, J. Weinberg: Analytic epistemology and experimental philosophy, *Phil. Compass* **2**, 56–80 (2007)
- 21.208 J. Knobe, S. Nichols (Eds.): *Experimental Philosophy* (OUP, Oxford 2008)
- 21.209 K. Ludwig: The epistemology of thought experiments: First person versus third person approaches, *Midwest Stud. Philos.* **31**, 128–159 (2007)
- 21.210 T. Williamson: Replies to Ichikawa, Martin and Weinberg, *Philos. Stud.* **145**, 465–476 (2009)
- 21.211 T. Williamson: Philosophical ‘intuitions’ and scepticism about judgement, *Dialectica* **58**, 109–153 (2004)
- 21.212 T. Williamson: Armchair philosophy, metaphysical modality and counterfactual thinking, *Proc. Aristot. Soc.* **105**, 1–23 (2005)
- 21.213 E. Machery: Thought experiments and philosophical knowledge, *Metaphilosophy* **42**(3), 191–214 (2011)
- 21.214 K. Lehrer: *Theory of Knowledge* (Westview Press, Boulder 1990)
- 21.215 S. Kripke: *Naming and Necessity* (Blackwell, Oxford 1980)
- 21.216 J. Weinberg, S. Nichols, S. Stich: Normativity and epistemic intuitions, *Phil. Topic* **29**, 429–460 (2001)
- 21.217 E. Machery, R. Mallon, S. Nichols, S. Stich: Semantics, cross-cultural style, *Cognition* **92**, B1–B12 (2004)

- 21.218 D. Dennett: *Elbow Room: The Varieties of Free Will Worth Wanting* (MIT Press, Cambridge 1984)
- 21.219 P. Swirski: *Of Literature and Knowledge: Explorations in Narrative Thought Experiments, Evolution and Game Theory* (Routledge, London, New York 2007)
- 21.220 M.T. Stuart: Philosophical conceptual analysis as an experimental method. In: *Meaning, Frames and Conceptual Representation*, ed. by T. Gärdenfors, D. Gerland, R. Osswald, W. Petersen (Düsseldorf University Press, Düsseldorf 2015) pp. 161–186
- 21.221 T. Gendler Szabó: Philosophical thought experiments, intuitions and cognitive equilibrium, *Midwest Stud. Philos.* **31**, 68–89 (2007)
- 21.222 M.T. Stuart: Cognitive science and thought experiments: A refutation of Paul Thagard's skepticism, *Perspect. Sci.* **22**, 264–287 (2014)
- 21.223 M. Arcangeli: Imagination in thought experimentation: Sketching a cognitive approach to thought experiments. In: *Model-Based Reasoning in Science and Technology*, ed. by L. Magnani, W. Carnielli, C. Pizzi (Springer, Dordrecht 2010) pp. 571–587
- 21.224 T. De Mey: Imagination's grip on science, *Metaphilosophy* **37**, 222–239 (2006)
- 21.225 E. McMullin: Galilean Idealization, *Stud. Hist. Philos. Sci.* **16**, 247–273 (1985)
- 21.226 P. Palmieri: Mental models in Galileo's early mathematization of nature, *Stud. Hist. Philos. Sci.* **34**, 229–264 (2003)
- 21.227 N.J. Nersessian: *Creating Scientific Concepts* (MIT Press, Cambridge 2008)
- 21.228 P.N. Johnson-Laird: *Mental Models: Toward a Cognitive Science of Language, Inference and Consciousness* (Harvard Univ. Press, Cambridge 1983)
- 21.229 P.N. Johnson-Laird: The history of mental models. In: *Psychology of Reasoning: Theoretical and Historical Perspectives*, ed. by K. Manktelow, M.C. Chung (Psychology Press, New York 2004) pp. 179–212
- 21.230 Y. Fehige, H. Wilsche: The body, thought experiments, and phenomenology. In: *Thought Experiments in Philosophy, Science, and the Arts*, ed. by M. Frappier, L. Meynell, J.R. Brown (Routledge, London, New York 2013) pp. 69–89
- 21.231 J. Zeimbekis: Thought experiments and mental simulations. In: *Thought Experiments in Methodological and Historical Contexts*, ed. by K. Ierodiakonou, S. Roux (Brill, Leiden–Boston 2011) pp. 193–215
- 21.232 A. Goldman: *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading* (OUP, Oxford 2006)
- 21.233 A. Goldman: Mirroring, mindreading, and simulation. In: *Mirror Neuron Systems: The Role of Mirroring Processes in Social Cognition*, ed. by J.A. Pineda (Humana Press, New York 2009) pp. 311–330
- 21.234 L. Meynell: Imagination and insight: A new account of the content of thought experiments, *Synthese* **191**(17), 4149–4168 (2014)
- 21.235 M.T. Stuart: Imagination: A 'Sine Qua Non' of scientific understanding, *Croat. J. Philos.* (2015), forthcoming
- 21.236 K. Mulligan: La varietà e l'unità dell'immaginazione, *Riv. Estet.* **11**, 53–67 (1999)
- 21.237 M. Balcerack Jackson: On the epistemic value of imagining, supposing, and conceiving. In: *Knowledge through Imagination*, ed. by A. Kind, P. Kung (Oxford Univ. Press, Oxford 2016) pp. 41–60
- 21.238 L. Souder: What are we to think about thought experiments?, *Argumentation* **17**, 203–217 (2003)
- 21.239 J. Weinberg: Configuring the cognitive imagination. In: *New Waves in Aesthetics*, ed. by K. Stock, K. Thomson-Jones (Palgrave Macmillan, Houndmills, Basingstoke 2008) pp. 203–223
- 21.240 D. Davies: Thought experiments and fictional narratives, *Croat. J. Philos.* **7**, 29–45 (2007)
- 21.241 D. Davies: Learning through fictional narratives in art and science. In: *Beyond Mimesis and Convention*, ed. by R. Frigg, M.C. Hunter (Kluwer Academic Publishers, Dordrecht 2010) pp. 51–70
- 21.242 K. Walton: *Mimesis as Make-Believe: On the Foundations of the Representational Arts* (Harvard University Press, Harvard 1990)
- 21.243 E.A. Davenport: Literature as thought experiment (On aiding and abetting the muse, *Philos. Soc. Sci.* **13**(3), 279–306 (1983)
- 21.244 N. Carroll: The wheel of virtue: art, literature, and moral knowledge, *J. Aesthet. Art Crit.* **60**(1), 3–26 (2002)
- 21.245 C.Z. Elgin: The laboratory of the mind. In: *A Sense of the World: Essays on Fiction, Narrative, and Knowledge*, ed. by W. Huerner, J. Gibson, L. Pocci (Routledge, London 2007) pp. 43–54
- 21.246 C.Z. Elgin: Fiction as thought experiment, *Perspect. Sci.* **22**, 221–241 (2014)
- 21.247 D. Davies: Can philosophical thought experiment be "screened"? In: *Thought Experiments in Philosophy, Science, and the Arts*, ed. by M. Frappier, L. Meynell, J.R. Brown (Routledge, London, New York 2013) pp. 223–238

---

# Models in Mathematics

## Part E Models in Mathematics

Ed. by Albrecht Heeffer

**22 Diagrammatic Reasoning  
in Mathematics**

Valeria Giardino, Nancy Cedex, France

**23 Deduction, Diagrams  
and Model-Based Reasoning**

John Mumma, San Bernadino, USA

**24 Model-Based Reasoning  
in Mathematical Practice**

Joachim Frans, Brussels, Belgium

Isar Goyvaerts, Torino, Italy

Bart Van Kerkhove, Brussels, Belgium

**25 Abduction and the Emergence  
of Necessary Mathematical Knowledge**

Ferdinand Rivera, San Jose, USA

The use of models in mathematics can broadly be distinguished as two categories. Most commonly, mathematical models are applied to the formal sciences and engineering, as well as to social sciences and other modes of quantitative reasoning. Secondly, models are also used within formal mathematics and in mathematical practice. This part of the book is mostly concerned with the latter use of models. In this sense, models are tools for reasoning in mathematics or teaching mathematics, and as a consequence also a means for understanding mathematical practice. The oldest and most common use of models in mathematical practice are applications of extended cognition. Chinese counting rods, the Roman abacus, medieval *jetons* or reckoning counters, or modern day computers are all material aids that allow us to delegate part of the cognitive load in performing complex calculations to the external environment. The operations that we carry out using these contrivances represent specific calculating procedures and in that sense they act as specific models of algorithms in arithmetic.

Diagrams are a more interesting application of extended cognition in mathematics. It is rather surprising that after more than 2000 years of practice with diagrams in mathematics, a study of their precise meaning, use, and function in mathematical reasoning has become a subject of serious study only during the past 20 years. **Chapter 22** by *Valeria Giardino* provides us with a comprehensive state of the art in recent research on mathematical diagrams. The new approach from the philosophy of mathematical practice – studying what mathematicians are actually doing when producing mathematics – has focused on the role of diagrams as a reasoning model rather than a visual representation of a mathematical object. Especially, the classic lettered diagram from Euclidean geometry has come under scrutiny, with the role of its constituent elements, their ontology, their epistemic functions, and their relation with the text and inherent ambiguities being inspected. It turns out that the diagram is not just a static object in a textbook, but as *Giardino* calls it, the diagram is *kinaesthetic*, as the text referring to it treats it as a constructed and manipulative object in which its inherent ambiguities become productive and open up modes of reasoning which are inhibited in more formal representations.

**Chapter 23** is a contribution by *John Mumma* that builds further on research from the philosophy on mathematical practice on Euclidean diagrams, in particular the work of *Ken Manders* (2008). *Mumma* presents a formalization of the model-based reasoning involved in mathematical diagrams, not only accounting for the

construction of the geometrical diagram but also its use in the demonstration.

A third application of extended cognition worth mentioning, while not treated as a dedicated subject below, are symbolic representations. It might be less obvious to view mathematical symbolism as a form of model-based reasoning, but recent research indicates that symbolism is not just a game of meaningless symbols and that modern symbolism, as it is taught and practiced today, relies on the visual processing of elements in a way that is similar to the interpretation of diagrams. Spatial organization, directionality, grouping, and mental operations like *picking up* and moving elements across symbolic expressions, appear to be crucial for our understanding of mathematical symbolism (*Heffer* 2014).

**Chapter 24** is the reflection of a joint project by two philosophers of mathematical practice, *Joachim Frans* and *Bart Van Kerkhove*, and the mathematician *Isar Goyvaerts*. This contribution provides a more general account of model-based reasoning in mathematical practice, concentrating on the processes that are required to arrive at higher levels of abstraction in mathematics. Three examples from different mathematical disciplines show how models facilitate additional layers of abstraction. The first one is from Euclidean geometry and deals with the abstraction from concrete shapes and measures to mathematical objects to which deductive reasoning can be applied. The second one is from approximation theory, in which functions can be modeled by other, more nicer and simpler functions. The third example is more technical and describes how the highest level of abstraction is achieved by the modeling of the inferences on certain algebraic objects by category theory.

**Chapter 25** deals with abduction, which is well suited for modeling mathematical inferences within the context of discovery. Abduction is the response to observations or findings which appear surprising or anomalous, and formulating hypotheses which allow us to evaluate their consequences. Abductive reasoning can thus lead to the emergence of new mathematical objects, ideas, or even theories. An earlier historical case study has shown how imaginary numbers appear from practice in Renaissance algebra (*Heffer* 2007). By means of empirical case studies, *Ferdie Rivera* shows how abductive processes are prominent within students' understanding of mathematics in the classroom, while previous studies have only focused on deductive or inductive reasoning. Four concrete suggestions are formulated to illustrate how abduction can be applied to mathematics education in a more systematic way.

---

## References

- K. Manders: The Euclidean diagram. In: *Philosophy of Mathematical Practice*, ed. by P. Mancosu (Clarendon, Oxford 2008), 112–183
- A. Heeffer: Epistemic justification and operational symbolism, *Found. Sci.* **19**(1), 89–113 (2014)
- A. Heeffer: Abduction as a Strategy for concept formation in mathematics: Cardano postulating a negative. In: *Abduction and the Process of Scientific Discovery*, ed. by O. Pombo, A. Gerner (Centro de Filosofia das Ciências da Universidade de Lisboa, Lisboa 2007), 179–194, Colecção Documenta

## 22. Diagrammatic Reasoning in Mathematics

Valeria Giardino

The objective of the present chapter will be to review the most recent studies about diagrammatic reasoning in mathematics. Section 22.3 will focus on the very much discussed topic of the role and of the features of diagrams and diagrammatic reasoning in Euclidean geometry. Section 22.4 will be devoted to the proposal of considering diagrams as representations that are introduced in support of other symbolic practices and whose power resides in their ambiguity. In Sect. 22.5, the attention will turn toward studies discussing diagrammatic reasoning in contemporary mathematics. In Sect. 22.6, computational perspectives on how to implement diagrammatic reasoning in computer programs will be introduced, both for Euclidean geometry and theory of numbers. In Sect. 22.7, it will be discussed how the study of diagrammatic reasoning can shed light onto the nature of mathematical thinking in general. Finally, in Sect. 22.8, some brief conclusions about diagrammatic reasoning in mathematics will be drawn. The choice of reviewing the research about diagrammatic reasoning along these lines is of course at least in part arbitrary. The aim of such a regrouping is to provide the reader with a map that can be helpful for exploring the various and already copious literature that has been recently produced on the subject. The ambition is that such a map will be as extensive as possible.

22.1	<b>Diagrams as Cognitive Tools</b> .....	499
22.2	<b>Diagrams and (the Philosophy of) Mathematical Practice</b> .....	501
22.3	<b>The Euclidean Diagram</b> .....	503
22.3.1	The (Greek) Lettered Diagram.....	504
22.3.2	Exact and Co-Exact Properties.....	505
22.3.3	Reasoning in the Diagram.....	506
22.3.4	Concrete Diagrams and Quasi-Concrete Geometrical Objects.....	508
22.4	<b>The Productive Ambiguity of Diagrams</b> .....	509
22.5	<b>Diagrams in Contemporary Mathematics</b> .....	510
22.5.1	Analysis.....	511
22.5.2	Algebra.....	513
22.5.3	Topology.....	514
22.6	<b>Computational Approaches</b> .....	515
22.6.1	(Manders') Euclid Reloaded.....	516
22.6.2	Theorem Provers.....	517
22.7	<b>Mathematical Thinking: Beyond Binary Classifications</b> .....	518
22.8	<b>Conclusions</b> .....	520
	<b>References</b> .....	521

### 22.1 Diagrams as Cognitive Tools

In his *Parallel Lives*, Plutarch famously reported the murder of Archimedes. He relates three different versions of the circumstances that brought about his death. According to the first one, Archimedes was so intent upon inspecting a diagram to work out some problem that he never noticed the incursion of the Romans, nor that the city was taken. His absorption in study and contemplation of the diagram was so deep that he declined to follow a soldier who had unexpectedly come up to him and commanded him to do so. Given his refusal, the soldier drew his sword and ran him through. In the

same spirit, in one of the most celebrated frescoes of the Italian Renaissance, *The School of Athens*, Raphael depicts a group of men attentively watching a scholar – most likely to be interpreted as Archimedes or Euclid – while he draws a geometrical figure on a clay tablet.

The mathematician is thus often portrayed as intently working on a diagram; this popular image attests to what extent the resource to diagrams, figures, or sketches – among other possible available instruments – is commonly considered as an outstanding element of the practice of mathematics. Is this picture true to the

facts? Are diagrams really part and parcel of the mathematical practice? And if it is so, what can be said about their features, use, and relations with other elements of the same practice? The objective of the present chapter is to introduce the most recent works on diagrammatic reasoning in mathematics and to review the answers that have been proposed so far for these questions. In this first section, the domain of inquiry – diagrammatic reasoning in mathematics – and the issues at stake in exploring it will be defined.

First of all, a clarification is needed on the meaning of the term *diagram* in diagrammatic reasoning, so as to avoid misinterpretations. Throughout the chapter – and possibly in contrast with other views – the term will be used in a very broad sense, that is, to include all cases of two-dimensional representations where their two dimensionality is relevant for the way in which information is displayed and read off from them. This seemingly too vague definition is actually appropriate to refer to many different phenomena that are found in mathematics. Moreover, diagrams will be intended here as *cognitive tools* that are meant to spatially display information in order to improve memory and promote inference, and not necessarily to depict mathematical objects. This will have two consequences: first, the focus of the analysis will be on diagrams and not on visualizations; second, lengthy discussions about the implications for the ontology of mathematics will be avoided. For these issues, one can refer among others to *Brown*, who claims that diagrams are not really pictures but rather “windows to Plato’s heaven” [22.1, p. 40], or to *Sherry*, who argues that some particular uses of diagrams make a realist view problematic [22.2].

Diagrammatic reasoning is surely relevant for human reasoning in general. As has been pointed out, human reasoning is *heterogeneous*: humans happen to rely on many different sorts of instruments with the aim of externalizing thought, diagrams being among them [22.3]. A common saying is that we are halfway to finding a solution to a problem when we are able to draw the right diagram for it. Nonetheless, in relation to mathematics, it is necessary to distinguish between mere sketches and diagrams. Sketches are certainly widespread and useful for the mathematician to reason about a problem or to communicate with one’s peers. However, they will not be the topic of this chapter, which will be devoted to diagrams as parts of a system of representation. Such diagrams obey some (more or less explicit) rules and their manipulation is controlled by the particular practice, in terms that will be defined later.

Not surprisingly, most analyses of diagrammatic reasoning in mathematics have dealt with Euclidean geometry, where the recourse to diagrams is so natural

and spontaneous that there is a tendency to take the presence and the effectiveness of diagrams for granted. Moreover, most diagrams in Euclidean geometry become part of our *visual repertoire* from a very early age at school. Think of the *Pythagorean theorem* and the impressive number of so-called *visual proofs* that have been given for it [22.4]. According to this theorem, the square of the hypotenuse ( $c$ ) of a right triangle equals the sum of the squares of its other two sides ( $a$  and  $b$ ). In letters,

$$a^2 + b^2 = c^2. \quad (22.1)$$

One of the possible visualizations for the Pythagorean theorem is offered in Fig. 22.1.

In Fig. 22.1a, four identical right triangles have been arranged into two rectangles. To obtain a square of side  $a + b$ , these two rectangles are added to two squares: one of side  $a$ , and the other of side  $b$ . In Fig. 22.1b, the same four triangles have been rearranged inside the square of side  $a + b$  and they now individuate another square of side  $c$ . By looking at the two diagrams together and by applying subtraction of the same objects – the four right triangles – to the same object – the square of side  $a + b$  – the Pythagorean theorem is obtained.

However, there are cases of diagrammatic reasoning that may be less obvious than in Euclidean geometry, for example, for statements about numerical properties. Consider the following *geometric series*

$$\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \dots = 1 \quad (22.2)$$

and its possible spatial arrangement in Fig. 22.2, in which each new rectangle or square drawn in the diagram – each new element added to the series – brings us closer to the square of area 1 (the example is taken from [22.1, pp. 36–38]).

As Brown points out, this *picture proof* should be contrasted with a traditional proof using  $\varepsilon$ - $\delta$  techniques. In such a proof, we first have to note that an infinite series converges to the sum  $S$  whenever the sequence of

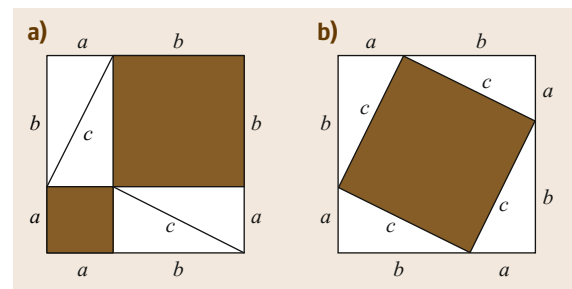
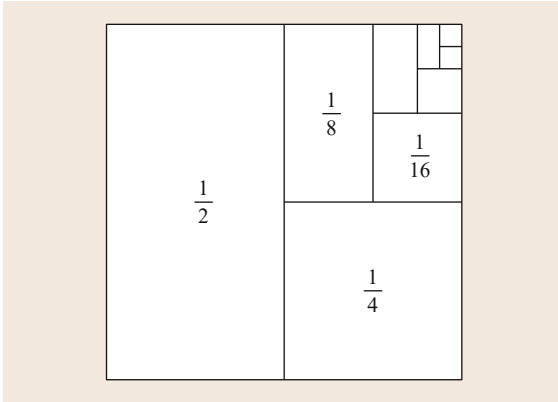


Fig. 22.1a,b Pythagorean theorem



**Fig. 22.2** A geometric series

partial sums  $\{s_n\}$  converges to  $S$ . In this case, we have

$$\begin{aligned} s_1 &= \frac{1}{2}, & s_2 &= \frac{1}{2} + \frac{1}{4}, & s_3 &= \frac{1}{2} + \frac{1}{4} + \frac{1}{8}, \\ s_n &= \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \dots + \frac{1}{2^n}. \end{aligned} \quad (22.3)$$

The values of these partial sums are

$$\frac{1}{2}, \frac{3}{4}, \frac{7}{8}, \dots, \frac{2^n - 1}{2^n}. \quad (22.4)$$

This infinite sequence has the limit 1, provided that for any number  $\varepsilon > 0$ , no matter how small, there is a number  $N(\varepsilon)$ , such that whenever  $n > N$ , the difference between the general term of the sequence  $\frac{2^n - 1}{2^n}$  and 1 is less than  $\varepsilon$ .

In symbols,

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{2^n - 1}{2^n} &= 1 \\ \iff (\forall \varepsilon)(\exists N) n > N &\rightarrow \left| \frac{2^n - 1}{2^n} - 1 \right| < \varepsilon. \end{aligned} \quad (22.5)$$

By applying some algebra, one obtains

$$\left| \frac{2^n - 1}{2^n} - 1 \right| < \varepsilon \iff \left| \frac{-1}{2^n} \right| < \varepsilon \iff 2^n > \frac{1}{\varepsilon}$$

$$\iff \log_2 \frac{1}{\varepsilon} < n. \quad (22.6)$$

Let now  $N(\varepsilon) = \log_2 \frac{1}{\varepsilon}$ . As a consequence,

$$n > \log_2 \frac{1}{\varepsilon} \rightarrow \left| \frac{2^n - 1}{2^n} - 1 \right| < \varepsilon. \quad (22.7)$$

We have thus proven that the sum of the series is 1. Compare now the easiness of forming the belief that the sum of the series is 1 by looking at the diagram in Fig. 22.2 with the resources required to prove the same result in a traditional way. The topic of the chapter will thus not only be Euclidean geometry. Other studies will be presented that analyze the usefulness of diagrammatic reasoning also in other branches of mathematics.

For the sake of completeness, there exists also very interesting work on ancient mathematics other than in Greece, involving, in some cases, also visual tools [22.5, 6]. Nonetheless, for reasons of space and given the specificity of the research, these works will not be among the subjects of the present chapter. It must also be noted that analogous considerations about the importance of diagrammatic reasoning in mathematics can be made to logic. Many scholars have discussed diagrammatic reasoning in logic, in an interdisciplinary fashion. Some studies have focused on the cognitive impact of diagrams in reasoning [22.7] and others on the importance of heterogeneous reasoning in logical proofs [22.8] and on the characteristics of nonsymbolic, in particular diagrammatic, systems of representation [22.9]. Very recently, and coherently with what will be later said about diagrammatic reasoning in mathematics, it was claimed that different forms of representation in logic are complementary to one another, and that future research should look into more accurate road maps among various kinds of representation so that the appropriate one may be chosen for any given purpose [22.10]. However, for reasons of space and despite the numerous parallels with the case of mathematics, the use of diagrammatic reasoning in logic will not be a topic of the present chapter.

## 22.2 Diagrams and (the Philosophy of) Mathematical Practice

The subject of diagrammatic reasoning in mathematics has recently gained new attention in the philosophy of mathematics. By contrast, in the nineteenth and twentieth centuries, this topic was neglected and not considered to be of philosophical interest; the heuris-

tic power of diagrams in mathematics was never denied, but visual mathematical tools were commonly relegated to the domain of psychology or to the context of discovery – by referring to a distinction between the context of discovery and that of justification that was very pop-



ular and that has become more and more precarious in recent years.

Famously, among others, *Russell* criticized Euclidean geometry for not being rigorous enough from a logical point of view [22.11, p. 404ff]. Consider the very first proposition of the *Elements*, which corresponds to the diagram in Fig. 22.3. The proposition invites the reader to construct an equilateral triangle from a segment AB by tracing two circles with centers A and B, respectively, and then connecting the extremes of the segment with the point that is created at the intersection of the two circles. According to *Russell*, “There is no evidence whatever that the circles which we are told to construct intersect, and if they do not, the whole propositions fails” [22.11, p. 404]. The proposition does, in fact, contain an implicit assumption based on the diagram – the assumption that the circles drawn in the proposition will actually meet. From *Russell*’s and analogous points of view, diagrams do not entirely belong to the formal or logical level, and therefore they should be considered as epistemically fragile. If this is assumed, then a proof is valid only when it is shown to be independent from the corresponding diagram or figure. In order to save Euclidean geometry from the potential fallacies derived from the appeal of diagrams, such as the one just shown, some assumptions, sometimes called *Pasch axioms*, were introduced. For example, it is necessary to assume that *A line touching a triangle and passing inside it touches that triangle at two points*, so as to avoid the reference to the corresponding diagram and make it a logical truth. By contrast, prior to the nineteenth century, such assumptions were generally taken to be “diagrammatically obvious” [22.12, p. 46].

There were historical reasons for this kind of scepticism in relation to the use of visual tools in mathematics. At the end of the nineteenth century, due to progress in disciplines such as analysis and algebra on the one hand, and the development of non-Euclidean geometries on the other, the request for a foundation of

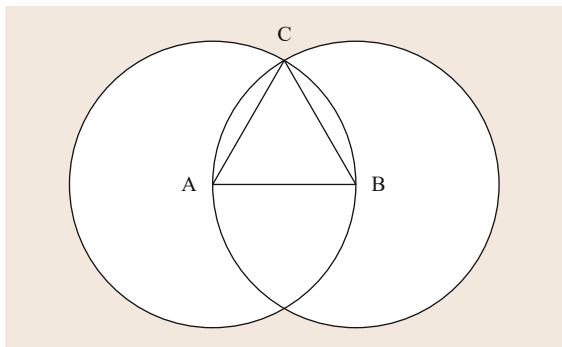


Fig. 22.3 Euclid, Proposition I.1

mathematics expressed a genuine mathematical need. Euclidean geometry was not the only logically possible geometry, and therefore it did not necessarily convey truth about the physical world: perception, motion, and superposition of figures had to be excluded as illegitimate procedures. In the course of the twentieth century, this *search for certainty* – as *Giaquinto* called it – became a sort of philosophical obsession [22.13]. Figures were considered as definitely unreliable, since they did not any more represent our knowledge of physical space. Moreover, they give rise to errors. Famously, *Klein* presented a case of a diagram that is apparently correct, but which in fact induces one to draw the – false – conclusion that all triangles are isosceles triangles [22.14, p. 202]. Paradigmatic in this sense was *Hilbert*’s program, who attempted to rewrite geometry without any unarticulated assumptions [22.15]. For such post-nineteenth century philosophy of mathematics, a proof should be followed, not *seen*.

However, some studies based on the scrutiny of the practice of mathematics have recently challenged this standard point of view. As editors of a book on visualization, explanation, and reasoning styles in mathematics, *Mancosu* et al. explained in 2005 how it was necessary to extend the range of questions to raise about mathematics besides the ones coming from the traditional foundational programs. The focus should be turned toward the consideration of “what mathematicians are actually doing when they produce mathematics” [22.16, p. 1]:

“Questions concerning concept-formation, understanding, heuristics, changes in style of reasoning, the role of analogies and diagrams etc. have become the subject of intense interest. [...] How are mathematical objects and concepts generated? How does the process tie up with justification? What role do visual images and diagrams play in mathematical activity?”

This invitation to widen the topics of philosophical inquiry about mathematics has developed into a sort of movement, the so-called *philosophy of mathematical practice*, which also criticizes the “single-minded focus on the problem of *access* to mathematical objects that has reduced the epistemology of mathematics to a *torso*” [22.16, p. 1]. Epistemology of mathematics can venture beyond the present confines and address epistemological issues that have to do with [22.16, p. 1]

“fruitfulness, evidence, visualization, diagrammatic reasoning, understanding, explanation and other aspects of mathematical epistemology which are

orthogonal to the problem of access to *abstract objects*.”

This approach would be more in line with what at least some of the very practitioners seem to think about the practice of mathematics. As *Jones*, a topologist and former Field medallist, summarizes, it is quite usual among mathematicians to have very little understanding of its philosophical underpinnings; in his view, for a mathematician, it is actually not at all difficult to live with worries such as Russell’s paradox while having complete confidence in one’s mathematics [22.17].

In this perspective, the study of diagrammatic reasoning in mathematics thus resumes its philosophical interest, by taking into account the appropriate areas of mathematics. Before presenting the different analyses that have been provided about diagrammatic reasoning in mathematics, three features of diagrammatic reasoning that will characterize most of the studies reviewed should be pointed out. First, diagrammatic reasoning in mathematics is not only *visual* reasoning. In fact, in most cases, a diagram comes with a *text*, and, as a consequence, any analysis of diagrammatic reasoning cannot disregard the role of the text accompanying diagrams. In two very fascinating volumes, *Nelsen* collected a series of proofs, taken from the *Mathematics Magazine*, that he calls “without words” [22.18, 19]. Nonetheless, these proofs are not exactly “without

words,” since to use a diagram is not only a matter of applying specific perceptual capacities but also of mastering the relevant background knowledge. In *Nelsen*’s proofs, diagrams refer to mathematical statements that can in some way be *found* in them. Diagrams and texts are, in fact, related: each practice will in turn define the terms of this relation. Second, there is another sense in which diagrammatic reasoning is not only visual. In most cases, diagrams are *kinaesthetic* objects, that is, they are intended to be changed and manipulated according to practice. A diagram can be conceived as an experimental ground, where mathematicians are qualified to apply *epistemic actions*, which are – following *Kirsch* and *Maglio*’s definition – “actions that are performed to uncover information that is hidden or hard to compute mentally” [22.20]. Third, as will be discussed in the Conclusions, the philosophical interest in studying diagrammatic reasoning is due to the cognitively hybrid status of diagrams. In fact, diagrams are certainly related to text, but at the same time, they are more than a mere visual translation of it; moreover, they are not only synoptic images, but also tools subject to manipulation; finally, they are not only part of the process of discovery, but in the appropriate context of use they are also able to constitute evidence for justification. The inquiry into diagrammatic reasoning in mathematics will in the end force us to blur the standard boundaries between the various elements of the mathematical practice.

## 22.3 The Euclidean Diagram

A review of the literature on diagrammatic reasoning in mathematics has to start from the research on Euclidean geometry. This section will be thus focused in particular on some of the most influential studies on the role and use of diagrams in the Euclidean system, both from a historical and a cognitive perspective. Given the complexity of such a discussion, details beyond the consideration of diagrammatic reasoning in mathematics will not be treated.

The reason for devoting one whole section of the chapter to Euclidean geometry is that the Euclidean diagram has always been considered as the paradigm of diagrammatic reasoning in mathematics. As *Ferreiros* has proposed, the mathematical practice of Greek geometers summarized in the *Elements* can be considered as a theoretical study of practical geometry [22.21, Chap. 5]. Its theoretical nature comes not only from the new goals and values that are identified as guiding the practice, but also from the idealizations introduced. This picture of Greek geometry contrasts with the ab-

stract tendency of reflections on the subject since Pasch and Hilbert. I have already pointed out in Sect. 22.1 that the post-nineteenth century approach tried to formalize mathematical proofs in such a way that diagrams are not part of them. One of the consequences of such an attitude was to consider diagrams as simple heuristic tools that are possibly useful in illustrating a result, but not constitutive of it. Therefore, there was an interest in translating Euclid’s *Elements* – maybe the most widely read text in the entire history of mathematics – into formal sentences of quantificational logic, so as to show that the reference to implicit assumptions based on the diagram could be avoided. A common feature of the studies that will be presented in this section will be precisely to point out that such a move would not represent Euclidean geometry as was originally conceived. If this is true, then it is necessary to provide a plausible explanation for the way in which information that is relevant for the proof can be read off from an Euclidean diagram. The post-nineteenth century philos-

ophy of mathematics gave foundations of logic for what was implicitly assumed in reference to a particular diagram. But what is *implicit* in a diagram? What cognitive abilities are needed to recognize this information and use it in a proof? Some proposals gave a Kantian reading of the *spatial intuition* that is involved in reasoning with a Euclidean diagram, as, for example, in the works of *Shabel* [22.22] and *Norman* [22.23]. According to these views, in Euclid's time, spatial and visual intuition was considered as mathematically reliable, and tacit assumptions were warranted on *the basis of spatial and visual information*. Nonetheless, these works have a wider scope than that of the present chapter, that is, they aim to give evidence in favor of the plausibility of a Kantian philosophy of mathematics, or at least of part of it. For this reason, they will not be discussed here.

Despite the specificity of the Euclidean case, in the remainder of the chapter, it will become evident how some of the characteristics of diagrammatic reasoning in Euclidean geometry can also be adapted to other mathematical practices involving diagrams. As already mentioned, the literature about diagrammatic reasoning in ancient Greek geometry is vast. The studies presented here are among the most influential ones. For other works, one can refer to the bibliography at the end of the chapter and to the references given in the single studies.

### 22.3.1 The (Greek) Lettered Diagram

The first analysis that will be introduced is the original and fascinating contribution on the shaping of Greek deduction provided by *Netz* [22.12]. *Netz*' aim is to reconstruct a *cognitive history* of the use of diagrams and text in Greek mathematics. According to his definition, cognitive history lies at the intersection of the history of science and cognitive science: it is analogous to the history of science, because it takes into account cultural artifacts, but it is also comparable to cognitive science because it approaches knowledge not through its specific propositional contents but by looking at its forms and practice. In *Netz*' words, such an intersection is "an interesting but dangerous place to be in" [22.12, p. 7]. In fact, his worry is that historians might see his research as over-theoretical and too open to generalization, while cognitive scientists might consider it as too "impressionistic" [22.12, p. 7].

*Netz*'s idea, in line with the philosophical approach described in Sect. 22.1, is to look at specific practices and consider the influence that they might have (or might have had) on the cognitive possibilities of science. His case study is Greek geometry. Note that *Netz*' analysis concerns Greek geometry in general and, differently from the studies that will be presented be-

low, does not focus on Euclid only. He starts from the observation that despite the already discussed post-nineteenth century criticisms, when doing Euclidean geometry, one would find it difficult [22.12, p. 23]

"to *unsee* the diagram, to teach oneself to disregard it and to imagine that the only information there is is that supplied by the text. Visual information is itself compelling in an unobtrusive way."

Euclidean diagrams seem to be part of the visual repertoire of shapes and figures that we are familiar with. If this is the case, then any analysis of Euclidean geometry must take this fact into account. One possible strategy would be to try to reconstruct the geometric practice of the time and focus on what *Netz* believes is the distinctive mark of Greek mathematics, something that has not been developed independently by any other culture: the *lettered diagram*.

Following *Netz*' definition, the lettered diagram is a combination of distinct elements that taken together make it possible to generalize an argument that is given in a single diagram having specific geometrical properties. The lettered diagram can thus be considered at different levels. At the logical level, it is composed, as the name suggests, by a combination of the continuous – the diagram – and the discrete – the letters added to it. At the cognitive level, it is a mixture of the visual resources that are triggered by it, and the finite manageable models that the letters made accessible. By following *Peirce*'s distinction among icons, indexes, and symbols [22.24], the lettered diagram associates, at the semiotic level, an icon – the diagram – with some indices – the letters. As will be shown in the next sections, *Peirce*'s distinction will be a reference also for other studies on diagrammatic reasoning in mathematics. It is interesting to point out from now that the *Peircean* terminology, despite being a common background for many of these authors, is applied in a variety of ways to different elements of diagrammatic reasoning in mathematics. The lettered diagram can be considered also from an historical point of view. Against this background, the same diagram is a combination of two elements. First, it refers to an art related to the construction of the diagram which, in *Netz*' analysis, is most likely a *banausic* art, that is, a practical art serving utilitarian purposes only. Second, it exploits a form of very sophisticated reflexivity, which is related to the use of the letters. The lettered diagram is an effective geometric tool precisely because of the richness of these different aspects characterizing it. In a lettered diagram, we see how almost antagonistic elements are integrated, so as to make it the appropriate instrument to promote and justify deduction [22.12, p. 67].

In Netz' reconstruction, Greek mathematics is constituted by a whole set of procedures for argumentation. These procedures are based on the diagram, which consequently serves as a source of evidence. Thanks to the procedure described in the text accompanying the lettered diagram in Fig. 22.3, one knows that the circles will actually meet at the intersection point. An interesting consequence of this reading is that the lettered diagram supplies a universe of discourse, without referring to any ontological principle. According to Netz, this would be a characteristic feature of Greek mathematics: the proof is done at an object level – the level of the lettered diagram – and no abstract objects corresponding to it need to be assumed. As he explains, in Greek practice [22.12, p. 57]:

“One went directly to diagrams, did the dirty work, and, when asked what the ontology behind it was, one mumbled something about the weather and went back to work. [...] There is a certain single-mindedness about Greek mathematics, a deliberate choice to do mathematics and nothing else. That this was at all possible is partly explicable through the role of the diagram, which acted, effectively, as a substitute for ontology.”

This point on the ontology of the Euclidean diagram is not uncontroversial. Other studies dealing with the Euclidean practice consider it necessary to take into account the abstract objects to which, in a way to define, the diagrams seem to refer. For example, *Azzouni* conjectures that the Greek geometers had to posit an ontology of geometrical objects, even if, in his stipulationist reading, this drive was not motivated by sensitivity to the presence of anything ontologically independent from us that mathematical terms refer to, but rather by geometers' need to prove things in a greater generality and to make applications easier [22.25]. We will see later how Panza introduces quasi-concrete geometrical objects (Sect. 22.3.4).

In this perspective, the paradox that Netz has to solve is how to explain that one proof – done by referring to a particular diagram, inevitably having specific properties – can be considered as a general result. In his interpretation, a proof in the Greek practice is an event occurring on a papyrus or in a given oral communication, and, despite this singularity, is something that is *felt* to be valid. Nonetheless, validity must be intended here in a different sense than the standard one. When looking at Greek mathematics, and contrary to the post-nineteenth century philosophy of mathematics, logic seems to collapse back into cognition.

In order to reply to this challenge, Netz first points out that generality in Greek mathematics exists only on a *global* plane: a theorem is proved having the global

system of Greek mathematics as a background. Thanks to this feature, the proof can be considered as invariant under the variability of the single action of drawing one diagram on the papyrus or of presenting the particular proof orally. Therefore, in Greek mathematics, what counts is the *repeatability* of the proof rather than the *generalizability* of the result (for details, see [22.12, Chap. 5]). According to Netz, to understand Greek geometry, a change of mentality is required: while we are used to generalizing a particular result, Greek mathematicians were used to extending the particular proof to other proofs using other and different objects that are nonetheless characterized by the same invariant elements. A particular construction, given by the lettered diagrams – the diagram plus the text accompanying it – can be repeated, and this is considered as certain.

The lettered diagram was a very powerful tool, because it allowed Greek mathematicians to automatize and elide many of the general cognitive processes that are implied in doing geometry. This was connected to expertise: the more expert a mathematician was, the more immediately he became aware of relations of form and the more readily he read off information from the diagram. Interestingly enough, such a feature of the practice with the Greek diagrams seems to be found in other contemporary mathematical practices as well. As the topologist and former Field medallist *Thurston* has proposed, mathematicians working in the same field and thus familiar with the same practice share the same “mental model” [22.26], which seems to refer precisely to the structure of the particular field and the amount of procedures that can be automatized or elided. To sum up, the diagram is a static object, but it becomes kinaesthetic thanks to the language that refers to it as a constructed and manipulable object: the proof is based on a practical invariance. In Netz' careful analysis, this is the best solution to the problem of generality that could be afforded at the time, given the means of communication at hand. If this is true, then any reconstruction as formalization, such as the one proposed by Hilbert, would not be faithful to the Greek practice. Moreover, Netz argues that Greek mathematics did not deal with philosophical matters. In the sources, nothing like a developed theory supporting this solution can be found.

### 22.3.2 Exact and Co-Exact Properties

Netz' approach is not the only one based on practical invariances. Consider Manders' contribution in an article that has been – in *Mancosu's* words – “an underground classic” [22.27, p. 14] and that was finally published in 2008 (in its original version, which dates back to 1995) [22.28]. In a later introductory paper, *Manders*

presents some of the philosophical issues that emerge from diagrammatic reasoning in geometry [22.29]. For him, Euclidean practice deserves philosophical attention, even only for the simple reason that it has been a stable and fruitful tool of investigation across diverse cultural contexts for over 2000 years. Up to the nineteenth century, no one would have denied that such a practice was rigorous; by contrast, it was rather considered as the most rigorous practice among the various human ways of knowing. Also in Manders' view, the Euclidean practice is based on a distribution of labor between two artifact types – the diagram and the text sequence – that have to be considered together. Note that once again the notion of artifact comes onto the scene as referring to diagrams as well as to text, that is, natural language plus letters linking the text to the diagram. Humans, due to their limited cognitive capabilities, cannot control the production and the interpretation of a diagram so as to avoid any case of alternative responses to it. For this reason, the text is introduced with the aim of tracking equality information. As Manders explains, in practice, the diagram and the text share the responsibility of allowing the practitioners to respond to physical artifacts in a “stable and stably shared fashion” [22.28, p. 83].

In Manders' reconstruction, proofs in traditional geometry have two parts: one verbal – the *discursive text* – and the other graphical – the *diagram*. The very objects of traditional geometry seem to arise in the diagram: in his words, “We enter a diagonal in a rectangle, and presto, two new triangles pop up” [22.28, p. 83]. The text ascribes some features to the diagram, and these features are called *diagram attributions*. Letters are introduced to facilitate cross-references between the text and the diagram – also Manders' Euclidean diagram is *lettered*. Defining diagram attributions, Manders introduces a distinction between *co-exact* and *exact* features of the diagram that has become, as will be shown, very influential. A *co-exact* feature is a directly attributable feature of the diagram, which has certain perceptual cues that are fairly stable across a range of variations. Moreover, such a feature cannot be readily eliminated, thanks to what Manders calls *diagram discipline*, that is, the proper exercise of skill in producing diagrams that is required by the practice. To clarify, if one continuously varies the diagram in Fig. 22.3, its *co-exact* attributes will not be affected. Imagine deforming the two circles no matter how: this would not change the fact that there still is a point at which the two figures intersect. The distinction thus concerns the control that one can have on the diagram and on its possible continuous deformations. This would be in line with the basic general resource of traditional geometrical practice, that, is diagram discipline: the appearance of

diagrams is controlled by standards for their proper production and refinement. Diagram discipline governs the possible constructions.

Consider the features of the diagram of a triangle. Such a diagram would *have to be* a nonempty region bounded by three visible curves, and these curves are straight lines. The first property is *co-exact* and the second is *exact*. Paradigmatic *co-exact* properties are thus features such as a region containing another – unaffected if the boundaries are shifted or deformed – or the existence of an intersection point such as the one required in Euclid I.1, as already discussed. By contrast, *exact* features are affected by deformation, except in some isolated cases. If one varies the diagram of the equilateral triangle, lines might no longer be straight or angles might lose their equality. In this framework, what is typically alleged as *fallacy of diagram use* rests on reading off from a diagram *exact* conditions of this kind – for example, that the lines in a triangle are not straight. However, the practice – the diagram discipline – *never allows* such a situation to happen. As already mentioned, practitioners created the resources to control the recourse to diagrams, so as to allow the resolution of disagreement among alternative judgements that are based on the appearance of diagrams, and therefore to limit the risk of disagreement for *co-exact* attributions. Things become trickier when it comes to *exact* properties, and this is the reason why the text comes in as support. In fact, since *exact* attributes are, by definition, unstable under the perturbation of a diagram, they can be priorly licensed by the *discursive text*. To go back to Euclid I.1, that the curves introduced in the course of the proof are circles is licensed, for example, by Postulate 3; furthermore, it is recorded in the *discursive text* that other subsequent *exact* attributions are to be licensed, such as the equality of *radii* (by Definition 15, again in the *discursive text*).

To sum up, for Manders, the diagram discipline is such that it is able to supervise the use of appropriate diagrams. In the remainder of the chapter, it will be shown how Manders' ideas have influenced other research in diagrammatic reasoning also going beyond traditional Euclidean geometry.

### 22.3.3 Reasoning in the Diagram

*Macbeth* has proposed a reading of the Euclidean diagram that is in line with the ones that have just been presented [22.30]. For the purpose of the chapter, it is interesting to note that her aim in reconstructing the practice of Euclidean geometry is to see whether a clarification of the nature of this practice might ultimately tell us something about the nature of mathematical practice in general. She criticizes the interpretation of the

*Elements* as an axiomatic system and proposes to see it instead as a system of natural deduction. Common notions, postulates, and definitions are not to be intended as premises, but as *rules* or *principles* according to which to reason. Moreover, in her view, a diagram is not an instance of a geometrical figure, but an *icon*. Such a feature of the Euclidean diagram makes the demonstration in the Euclidean system general throughout.

In order to clarify such a claim, Macbeth introduces Grice's distinction between *natural* and *nonnatural* meaning [22.31]. For Grice, natural meaning is exemplified by sentences, such as *These spots mean measles*. By contrast, a sentence, such as *Schnee means snow*, expresses nonnatural meaning. Let us suppose then that a drawing is an instance of a geometrical figure, that is a particular geometrical figure. If this is the case, it would have natural meaning and a semantic counterpart. For example, in Fig. 22.3, one sees a particular triangle ABC that is one instance of some sort of geometrical entity called an *equilateral triangle*. But let us instead hypothesize that the drawing has nonnatural meaning and therefore is not an instance of an equilateral triangle but *is taken for* an equilateral triangle. Then, the crucial step would be to recognize the *intention* that is behind the making of the drawing. This is the reason one can also draw an imprecise diagram – for example, drawing a circle that looks like an ovoid – as long as the intention – the one of drawing a circle – is clear. Such an intention is expressed throughout the course of the demonstration. Also, Azzouni has suggested that the proof-relevant properties are not the actual (physical) properties of singular diagrammatic figures, but conventionally stipulated ones, the recognition of which is *mechanically executable* [22.25].

To sum up, in Macbeth's reconstruction, the Euclidean diagram has nonnatural meaning and is, by intention, general. Moreover, by following Pierce's distinction again, it is an icon because it *resembles* what it signifies. However, resemblance here cannot be intended as resemblance in appearance. The Euclidean diagram resembles what it signifies by displaying the same relations of parts, that is, by being *isomorphic* to it. The circles in Fig. 22.3 are icons of a geometrical circle because there is a likeness in the relationship of the parts of the drawings. Specifically, the resemblance is in the relation of the points on the drawn circumference to the drawn center compared to the relation of the corresponding parts of the geometrical concept. Such a resemblance can be a feature of the diagram because the geometer means or intends to draw a circle, that is, to represent points on the circumference that are equidistant from the center. Given this intention, it is not important whether or not the figure is precise, that is, whether or not the points on the circumference in the

drawn figure really look that way. There is a correspondence between the iconicity of the Euclidean diagram as introduced by Macbeth and co-exact properties in Manders' terms. Also in Macbeth's reading, the diagram is intended to show the relations that are constitutive of the various kinds of geometrical entities involved. As she summarizes, "A Euclidean diagram does not instantiate content but instead formulates it" [22.30, p. 250].

Finally, Macbeth aims to show that the chain of reasoning in Euclidean geometry involving diagrams is not diagram-based but *diagrammatic*. According to her terminology, a reasoning is diagram-based when its moves are licensed or justified by the diagram; by contrast, it is diagrammatic when the mathematician is asked to reason *in* the diagram. Consider again Fig. 22.3. There is a sense in which this figure is analogous to the Wittgensteinian duck–rabbit picture, where one alternates between seeing it as the picture of a duck and seeing it as the picture of a rabbit. In a similar fashion, in order for the demonstration to go through, the mathematician has to alternate between seeing certain lines in the figure as icons of *radii* – and therefore equal in length – and as icons of the sides of a triangle – so as to draw the conclusion that the appropriately constructed triangle is in fact equilateral. The point then is that the physical marks on the page have the potential to be regarded in radically different ways. By pointing at such a feature of the Euclidean diagram, Macbeth aims to make sense of Manders' view, saying that geometrical relations *pop out* of the diagram as lines are added to it. The mathematician uses the diagram to reason *in* it and to make new relations appear.

Moreover, according to Macbeth, the Euclidean diagram has three levels of articulation in the way it can be parsed by the geometer's gaze. At a first level, there are the primitive parts: points, lines, angles, and areas. At the second level, there are geometrical objects that are intended to be represented in the diagram. At the third level, there is the whole diagram, which is not in itself a geometrical figure but, in some sense, contains the objects at the other levels. In the course of the demonstration, the diagram can thus be configured and reconfigured according to different intermediate wholes. Thanks to such a function of diagrams, significant and often surprising geometrical truths can be proved. In Macbeth's account, the site of reasoning is the diagram, and not the accompanying text. Her conclusion is that Euclidean geometry is [22.30, p. 266]

"a mode of mathematical enquiry, a mathematical practice that uses diagrams to explore the myriad discoverable necessary relationships that obtain among geometrical concepts, from the most obvious to the very subtle."

Another more recent study has complemented Manders and Macbeth's account by emphasizing even more strongly how the Euclidean diagram has a role of *practical synthesis*: to draw a figure means to balance multiple desiderata, making it possible to put together insight – that is timeless – and constructions – that are given in time [22.32]. We also mention here that Macbeth has applied similar arguments to the role of Frege's *Begriffsschrift* as exhibiting the inferentially articulated contents of mathematical concepts [22.33]. Despite the interest of this account, for the reasons given in Sect. 22.1, we will not give here the details of such a study.

In Sect. 22.4, we will come back to the notion of iconicity and see how the productive ambiguity to which Macbeth alludes to in talking about the parsing of the Euclidean diagram can also be found in other cases of diagrammatic reasoning.

### 22.3.4 Concrete Diagrams and Quasi-Concrete Geometrical Objects

Another view on the generality of the Euclidean diagrams has recently been proposed by Panza [22.34]. His aim is to analyze the role of diagrams in Euclid's plane geometry, that is, the geometry as expounded by Euclid in the first six books of the *Elements* and in the *Data*, and as largely practiced up to early-modern age (see also [22.35]). In his view, Euclid's propositions are general insofar as they assert that there are some admitted rules that have to be followed in constructing geometric objects. Once again, what matters for generality are construction procedures. These admitted rules allow the geometer to construct an object having certain properties and relations. To put it briefly, it would be impossible for one to follow the rules and end up with constructing an object without the requested properties.

Panza argues that arguments in the Euclidean system are *about* geometrical objects: points, segments of straight lines, circles, plane angles, and polygons. Taking inspiration from Parsons [22.36], he defines such geometrical objects as *quasi-concrete*. Their quasi-concreteness depends precisely on the relation they have with the relevant diagrams, which are instead *concrete* objects: the Euclidean diagram is a configuration of points and lines, or better is what is common to equivalence classes of such configurations. Two claims describe the peculiarity of the relation between quasi-concrete geometrical objects and concrete diagrams. First, the identity conditions of the geometrical objects are provided by the identity conditions of the diagrams that represent them. In his definition, this is the *global* role of diagrams in Euclid's arguments: a diagram is taken as a starting point of licensed procedures for

drawing diagrams and a geometrical object can be given in the Euclidean system when a procedure is stipulated for drawing a diagram representing it. Second, the geometrical objects inherit some properties and relations from these diagrams. This is the *local* role of Euclidean diagrams. Such properties and relations are recognized because a diagram is compositional. So understood, a diagram is a configuration of concrete lines drawn on an appropriate flat material support. According to Panza, Euclid's geometry is, therefore, neither an empirical theory nor a contentual one in Hilbert's sense, that is, a theory of "extra-logical discrete objects, which exist intuitively as immediate experience before all thought" [22.37, p. 202]. In his view, differently from the approaches described so far, it is crucial to define an appropriate ontology for the Euclidean diagram. In fact, his objective is to argue against the view that arguments in Euclid's geometry are not about singular objects, but rather about something like general schemas, or only about concepts. Such a view, according to which Euclidean geometry would deal with purely ideal objects, is often taken to be Platonic in spirit and is supposed to have been suggested by Proclus [22.38, 39]. Panza's proposal is instead closer to an Aristotelian view that geometric objects result by abstraction from physical ones, but the author claims that it is not his intention to argue that Euclid was actually guided by an Aristotelian rather than a Platonic insight.

In the same spirit, also Ferreiros suggests that the objects of Greek geometry are taken to be the diagrams and other similarly shaped objects [22.21, Chap. 5]. Of course, the diagram in this context is not intended to refer to the physically drawn lines that are empirically given, but to the *interpreted* diagram, which is perceived by taking into account the idealizations and the exact conditions conveyed in the text and derived from the theoretical framework in the background. For the geometer, the figure one works with is not intended as an empirical token but as an ideal type. Nonetheless, it is crucial to remark that such an ideal type does not exist outside the mind of the geometer and becomes available only thanks to the diagram. Therefore, on the one hand, the object of geometry is the diagram, and, as a consequence, the diagram *constitutes* the object of geometry; on the other hand, the diagram has to be interpreted in order to make the object emerge, and accordingly it also *represents* the object of geometry. Moreover, quoting Aristotle, Ferreiros points out that Greek geometry remains a form of theoretical and not practical geometry, for the reason that its objects are conceived as *immovable and separable*, without this necessarily leading to the thesis that there exist *immovable and separable* entities [22.40].

## 22.4 The Productive Ambiguity of Diagrams

This brief section will be devoted to the discussion of the role of ambiguity in diagrammatic reasoning. Grosholz has devoted her work to develop a pragmatic approach to mathematical representations, by arguing that the appropriate epistemology for mathematics has to take into account the pragmatic as well as the syntactic and semantic features of the tools that are used in the practice of mathematics. The post-nineteenth century philosophy of mathematics wants all mathematics to be reduced to logic; by contrast, Grosholz claims that philosophy should account for all kinds of mathematical representations, since they are all means to convey mathematical information. Moreover, the powers and limits of each of them should be explored. One format might be chosen among the others for reasons of convenience, depending on the problem to solve in the context of a specific theory or in a particular historical moment. Even the analysis of the use of formal language can thus be framed in terms of its representational role in a historical context of problem-solving. As Grosholz explains [22.41, p. 258],

“Different modes of representation in mathematics bring out different aspects of the items they aim to explain and precipitate with differing degrees of success and accuracy.”

In such a picture, a central cognitive role is played in mathematics by a form of controlled and highly structured ambiguity that potentially involves all representations, and is particularly interesting in the case of diagrams. Grosholz as well adopts the general Peircean terminology and distinguishes between *iconic* and *symbolic* uses of the same representation. These two different uses make representations potentially ambiguous.

To clarify, consider as an example Galileo’s treatment of free fall and projectile motion in the third and fourth days of his *Discourses and Mathematical Demonstrations Concerning Two New Sciences*. Galileo draws a geometrical figure to prove that (Fig. 22.4),

“The spaces described by a body falling from rest with uniformly accelerated motion are to each other as the squares of the time-intervals employed in traversing these distances.”

In the right-hand figure, the line HI stands for the spatial trajectory of the falling body, but is articulated into a sort of ruler, where the intervals representing distances traversed during equal stretches of time, HL, LM, MN, etc., are indicated in terms of unit intervals, which are represented by a short cross-bar, and

in terms of intervals, whose lengths form the sequence of odd numbers (1, 3, 5, 7, . . .), which are represented by a slightly longer cross-bar. The unit intervals are intended to be counted as well as measured. In the left-hand figure, AB represents time, divided into equal intervals AD, DE, EF, etc., with perpendicular instantaneous velocities raised upon it – EP, for example, represents the greatest velocity attained by the falling body in the time interval AE – generating a series of areas which are also a series of similar triangles. Thanks to an already proven result from Th. I, Prop. 1, Galileo builds the first proposition, according to which the distance covered in time AD (or AE) is equal to the distance covered at speed  $1/2$  DO (or  $1/2$  EP) in time AD (or AE). Therefore, the two spaces that we are looking for are to each other as the distance covered at speed  $1/2$  DO in time AD and the distance covered at speed  $1/2$  EP in AE. Th. IV, Prop. IV tells us that “the spaces traversed by two particles in uniform motion bear to one another a ratio which is equal to the product of the ratio of the velocities by the ratio of the times”; in this case, given the similarity of the triangles ADO and AEP, AD and AE are to each other  $1/2$  DO and  $1/2$  EP. Then, the proportion between the two velocities compounded with the time intervals is equal to the proportion of the time intervals compounded with the time intervals, and therefore  $[V_1 : V_2]$  compounded with  $[T_1 : T_2]$  equals  $[T_1 : T_2]^2$ . As a consequence, the spaces described by the falling body are proportional to the squares of the time intervals:  $[D_1 : D_2]^2 = [T_1 : T_2]^2$ . Look now at the left-hand diagram. Consider the sums

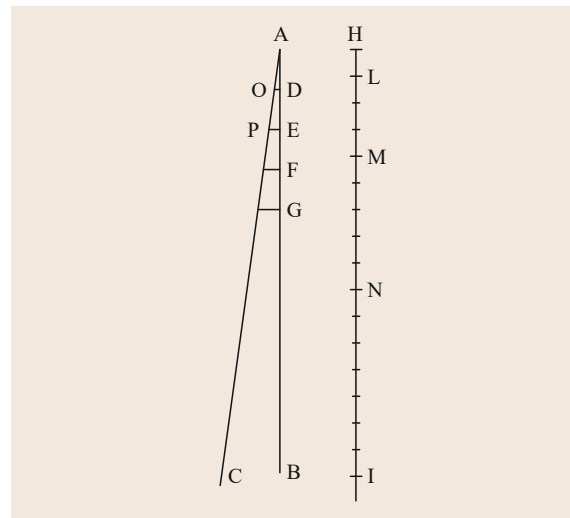


Fig. 22.4 Galileo, *Discorsi*, third day, naturally accelerated motion, Theorem II, Proposition II



$1 + 3 = 2^2$ ,  $1 + 3 + 5 = 3^2$ ,  $1 + 3 + 5 + 7 = 4^2$ , and so forth. These sums represent distances and are proportional to the squares of the intervals. Therefore, the time elapsed is proportional to the final velocity and the distance fallen will be proportional to the square of the final velocity.

Galileo's use of the diagram can be analyzed in relation to the different *modes of representation* that are employed to express his argument to prove the theorem. First, he refers to at least four modes of representation: proportions, geometrical figures, numbers, and natural language. Second, the same geometrical diagram serves as an icon and at the same time as a symbol. As an icon, it is configured in such a way that it can stand for a geometrical figure and exhibit patterns of relations among the data it contains. For example, when proportions are taken as finite, they are represented iconically. When the proportions are taken as infinitesimal (because one may take "any equal interval of time whatsoever" [22.41, p. 14]), the diagram is instead used as a symbol. In this case, the configuration of the diagram changes because it is now intended to represent dynamical, temporal processes. Therefore, despite the fact that an appropriate parsing of the diagram cannot represent iconically something that is dynamical or temporal, it can still do it symbolically. In Grosholz' view, the distinction between iconic and symbolic use of a mode of representation sheds light on the importance of semantics in mathematics. In fact, for a mode of representation to be intended not only iconically but also symbolically, the reference to some background knowledge is necessary. The representation does not have to be intended in its *literal* configuration but from within a more elaborated context of use, which provides a new interpretation and a new meaning for it. Galileo's diagram must thus be interpreted in two ways: intervals have to be seen as finite – so that Euclidean results can be applied – and also as infinitesimals – so as to represent accelerated motion. In the proof, errors are prevented by a careful use of ratios.

Compare this example with Macbeth's discussion of Euclid I.1 in the previous section. Here as well, there is only one set of diagrams, but, in order for the demonstration to go through, it must be read and interpreted in different ways. However, Macbeth and Grosholz employ Peirce's distinction in a different way. Macbeth

talks of Euclidean diagrams as icons for geometrical relations, while Grosholz refers to two possible different uses – iconic or symbolic – of the same diagram. Moreover, Macbeth's Gricean distinction between natural and nonnatural meaning does not coincide with the distinction between the literal and nonliteral – conventional – uses of the representation made here by Grosholz.

Grosholz' approach is not limited to diagrams in mathematics, unless one wants to say that all mathematical representations are diagrammatic. In fact, in her view, another straightforward example of productive ambiguity is Gödel's representation of well-formed formulas through natural numbers, whose efficacy stems from their unique prime decomposition. In her terminology, the peculiarity of Gödel's proof of incompleteness is that the numbers in it must stand iconically for themselves – so as to allow the application of number theoretic results – and symbolically for well-formed formulas – so as to allow transferring those results to the study of completeness and incompleteness of logical systems. Without going into details, it is sufficient to say that Grosholz points out that this particular case shows how much even logicians exploit the constitutive ambiguity of some of the representations they use. In her view, the recourse to ambiguous formats is, in fact, typical of mathematical reasoning in general, and this is precisely the feature of mathematics that has not been recognized by the standard post-nineteenth century approaches, which have focused on the possibility of providing a formal language that would avoid ambiguities. As Grosholz explains [22.41, p. 19]

"the symbolic language of logistics is allegedly an ideal mode of representation that makes all content explicit; it stands in isomorphic relation to the objects it describes, and that one-one correspondence insures that its definitions are 'neither ambiguous nor empty'."

In Grosholz's view, ambiguity and iconicity then seem to be not only a mark of diagrams such as Galileo's one, but also crucial features of mathematical representations – formulas not being an exception.

In the following sections, other examples of productive ambiguity and iconicity in contemporary mathematics will be given.

## 22.5 Diagrams in Contemporary Mathematics

As shown in the previous sections, most examples of diagrams that have been discussed in the literature so far are taken from the history of mathematics; furthermore, the focus has been on geometric diagrams. It is worth

mentioning here an interesting study by Chemla about Carnot's ideas on how to reach generality in geometry, where she analyzes Carnot's treatment of the so-called *theorem of Menelaus* [22.42]. In her reconstruction,

Carnot believes that, at least in the case of the theorem of Menelaus, the diagram must be considered as a configuration, appropriately chosen with the aim of finding the solution to the problem in question. As a consequence, the theorem no longer concerns a specific quadrilateral, but any intersection between a triangle and a straight line. Chemla claims that Carnot’s ideas were nonstandard at his time, because he introduced a way of processing information that relies on individuating what a general diagram is in opposition to a multitude of particular figures.

This section will be devoted to briefly introducing some works on diagrammatic reasoning in present-day mathematics. The studies have been divided into three categories: analysis, algebra, and topology. Differently from the Euclidean or the theory of number case, the examples taken from contemporary mathematics deserve much more technical machinery in order to be understood, that is, even only to introduce the diagram, much mathematics is required. For reasons of space, it is therefore impossible to give here all the mathematical details, and I invite the reader to refer to the original papers.

### 22.5.1 Analysis

In two different articles, *Carter* analyzed a case study of diagrammatic reasoning in free probability theory, an area introduced by Voiculescu during the 1980s [22.43, 44]. The aim of free probability theory was to formulate a noncommutative analog to classical probability theory, with the hope that this would lead to new results in analysis. In particular, Carter discusses a section of a paper written by *Haagerup* and *Thorbjørnsen* [22.45], where a combinatorial expression for the expectation of the trace of the product of so-called *Gaussian random matrices* (GRMs) of the following form is found

$$E \circ \text{Tr}_n[B^* B^p]. \tag{22.8}$$

The authors show that this expression depends on the following

$$E \circ \text{Tr}_n[B_1^* B_{\pi(1)} \dots B_p^* B_{\pi(p)}]. \tag{22.9}$$

The indices  $\pi(i)$  are symbols denoting the values of a permutation  $\pi$  on  $\{1, 2, \dots, p\}$ . Therefore, the value of the expression depends on the existence and properties of the permutation that pairs the matrices off  $2 \times 2$ .

Diagrams can be introduced to represent the permutations, and this is a crucial move, since such diagrams make it possible to study permutations independently from the fact that they were set forth as indices of the GRM. Moreover, the recourse to diagrams makes it easier to evaluate the properties of the permutations. Once

the relevant properties of the permutation are identified, thanks to the diagram, they can then be reintroduced into the original setting.

To give an idea of what the diagrams representing permutations look like, consider two examples of constructing the permutation  $\hat{\pi}$ . Let  $p = 4$ , so that  $\pi : \{1, 2, 3, 4\} \rightarrow \{1, 2, 3, 4\}$ .

Instead of writing

$$B_1^* B_{\pi(1)} B_2^* B_{\pi(2)} B_3^* B_{\pi(3)} B_4^* B_{\pi(4)}, \tag{22.10}$$

we rewrite the expression in the following form

$$C_1^* \cdot C_2 \cdot C_3^* \cdot C_4 \cdot C_5^* \cdot C_6 \cdot C_7^* \cdot C_8. \tag{22.11}$$

Suppose then that  $\pi(1) = 2$  and  $\pi(3) = 4$ , giving

$$B_1^* \cdot B_2 \cdot B_2^* \cdot B_1 \cdot B_3^* \cdot B_4 \cdot B_4^* \cdot B_3. \tag{22.12}$$

What the permutation  $\hat{\pi}$  is supposed to do is to tell us which of the  $C_s$  are identical, in terms of their indices. By comparing the two expressions, we see that  $C_1 = C_4$ ,  $C_2 = C_3$ ,  $C_5 = C_8$ , and  $C_6 = C_7$ . In terms of the permutation  $\hat{\pi}$ , this means that  $\hat{\pi}(1) = 4$  and  $\hat{\pi}(2) = 3$ , and so on. Both permutations can be represented by the diagrams in Fig. 22.5.

Another example could be  $\pi(1) = 3$  and  $\pi(2) = 4$ , giving

$$B_1^* \cdot B_3 \cdot B_2^* \cdot B_4 \cdot B_3^* \cdot B_1 \cdot B_4^* \cdot B_2. \tag{22.13}$$

By rewriting it in terms of  $C_i$ ’s and comparing again, we obtain  $C_1 = C_6$ ,  $C_2 = C_5$ ,  $C_3 = C_8$ , and  $C_4 = C_7$ , as shown in Fig. 22.6.

First, diagrams would suggest definitions and proof strategies. In Carter’s example, the definitions of a pair of neighbors, or of a noncrossing and a crossing permutation as well as of cancellation of pairs – manipulations that are all clearly visible in the diagrams – are *inspired* by them. Moreover, as confirmed by the very authors of the study, also the formal version of at least a part of the proofs is inspired by the proof based on

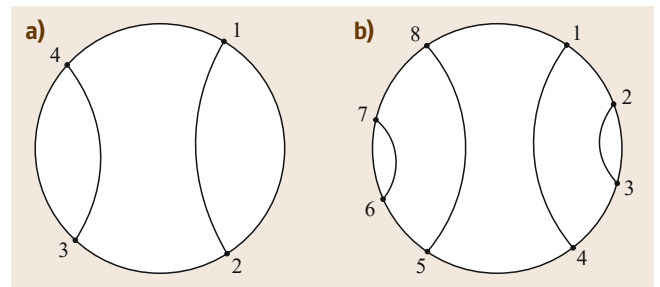


Fig. 22.5 (a)  $\pi$  is the permutation (12)(34); (b) the correspondent  $\hat{\pi}$

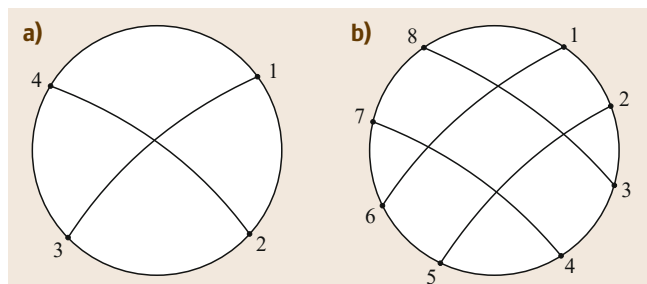


Fig. 22.6 (a)  $\pi$  is the permutation (13)(24); (b) the correspondent  $\hat{\pi}$

diagrams. Second, diagrams function as *frameworks* in parts of proofs: Although they are not used directly to give rigorous proofs, they still play an essential role in the discovery and formulation of both mathematical theorems and proofs, and thus in the practice of the mathematical reasoning.

Carter's idea is that certain properties of the diagrams correspond to formal definitions. In her case study, some diagrams are used to represent permutations and similar diagrams to represent equivalence classes. Diagrams thus make it possible to perform *experiments* on them; for example, the crossings identify the number of the equivalence classes and therefore the definition of a crossing is given an algebraic formulation. Likewise, the concepts of a neighboring pair and of removing pairs (from the diagram) are translated into an algebraic setting. To sum up, the relations used in the proof based on the diagram represent relations that *also* hold in the algebraic setting. As Carter explains, the notions of crossing and neighboring pairs are, in Manders-inspired terminology, examples of co-exact properties of the diagrams. In a Peircean semiotic perspective, the diagram in this case would again be *iconic*, and it is for this reason that one can translate the diagrammatic proof into an algebraic proof. In this example, from contemporary mathematics, we are in a sense certainly far from the Euclidean diagram, but we still see that the proof includes an accompanying text; only when the appropriate text is added, do text and diagram – taken together – constitute a proof. The text is also important to disambiguate diagrams that can be interpreted as representing different things (recall Manders' view on the Euclidean diagram).

In a more recent article, Carter discusses at length her reference to Peirce's terminology. Her reconstruction of Peirce's discussion of the use of representation in mathematics is based on some of the most recent studies about Peirce's mathematical philosophy [22.46]. Note that the central notion for Peirce is the one of *sign*, that is, in his words, "Something that stands for something else" [22.24, 2.228]. A sign can stand for something else not in virtue of some of its

particular features, but thanks to an interpretant that links the sign to the object. For Peirce, signs are then divided into three categories: icons, indices, and symbols; icons are signs in virtue of a relation of likeness with their objects, indices are actually connected to the objects they represent, and symbols represent an object because of a rule stipulating such a relation. Central to Peirce's conception of reasoning in mathematics is that all such reasoning is *diagrammatic* – and therefore iconic. Moreover, Peirce employs the term *diagram* in a much wider sense than usual. In his view, even spoken language can be diagrammatic. Consider a mathematical theorem that contains certain hypotheses. By fixing the reference with certain indices, it is possible to produce a diagram that displays the relations of these referents. In statements concerning basic geometry, the diagram could be a geometric diagram such as the Euclidean diagram. But in other parts of mathematics, it may take a different form. In Carter's view, the diagrams in her case study are iconic because they display properties that can be used to formulate their algebraic analogs. Moreover, the role of indices – the numbers – in the diagram is to allow for reinserting the result into its original setting. Once such a framework is assumed, then diagrams as well as other kinds of representations used in mathematics become an interesting domain of research. As already discussed when presenting Grosholz's work, the objects of inquiry extend from mathematical diagrams to mathematical signs – mathematical representations – in general, including, for example, also linear or two-dimensional notations. In the final section, we will say more about this issue. A further point made by Carter is that the introduced diagrams also enable us to break down proofs into manageable parts, and thus to focus on certain details of a proof. By using diagrams at a particular step of the proof, one needs only to focus on one component, thus getting rid of irrelevant information. In an unpublished paper, Manders makes a similar point by introducing the notions of responsiveness and indifference in order to address the topic of progress in mathematics [22.47]. In the following section, more details about this paper will be given.

It is interesting to note that Carter discusses a potential ambiguity of the term *visualization*, used as (i) representation, as in the example given, and as (ii) mental picture, helping the mathematician see that something is the case. In this second meaning, diagrams would be *fruitful frameworks* to trigger imagination. Carter's claim is that there is not a sharp distinction to be drawn here between concrete pictures and mental ones, but quite the opposite: a material picture may trigger our imagination, producing a mental picture, and

vice versa a mental picture may be reproduced by a concrete drawing. We will come back also to this issue later.

### 22.5.2 Algebra

Another case study from contemporary mathematics is taken from a relatively recent mathematical subject: *geometric group theory*. Starikova has discussed how the representation of groups by using *Cayley graphs* made it possible to discover new geometric properties of groups [22.48, 49]. In this case study, groups are represented as graphs. Thanks to the consideration of the graphs as metric spaces, many geometric properties of groups are revealed. As a result, it is shown that many combinatorial problems can be solved through the application of geometry and topology to the graphs and by their means to groups.

The background behind Starikova's work is the analysis proposed by Manders in the unpublished paper already mentioned in presenting Carter's work [22.47]. In this paper, Manders elaborates more on his study on Euclidean diagrams, this time taking into account the contribution of Descartes' *Géométrie* compared to Euclid's plane geometry. He gives particular stress to the introduction of the algebraic notation. In fact, in mathematical reasoning, we often produce and respond to artifacts that can be of different kinds: natural language expressions, Euclidean diagrams, algebraic or logical formulas. In general, mathematical practice can be defined as the control of the *selective responses* to given information, where response is meant to be *emphasizing* some properties of an object while *neglecting* others. According to Manders, artifacts help to implement and control these selective responses, and therefore their analysis is crucial if the target is the practice of the mathematics in question. Moreover, selective responses are often applied from other domains. Think of the introduction of algebraic notation to apply fast algebraic algorithms. In Descartes' geometry, geometric problems are solved through solving algebraic equations, which represent the geometric curves. Also here, the idea is that by using different representations of the same concepts, new properties might become noticeable. Starikova's study would show a case where a change in representation is a valuable means of finding new properties: drawing the graphs for groups would help discovering new features characterizing them. In this perspective, mathematical problem-solving involves the creation of the right strategies of selection: at each stage of practice, some information is taken into account and some other information is disregarded. It is only by responding to some elements coming from the mathematical context and not paying attention to others that we can control

each step of our reasoning. Of course, this control and coordination may have different levels of *quality* across practices. Manders' conclusion is that mathematical progress is based on this coordinated and systematic use of responsiveness and indifference, and that such a coordination is implemented by the introduction and the use of the various representations. The role of the accompanying text is still crucial, since diagrams are produced according to the specifications in the text. Thanks to the text, the depicted relations become reproducible and therefore stable; diagram and text keep supporting each other.

To give the reader an idea of what a Cayley graph for a group looks like, we consider first the definition of a *generating set*. Let  $G$  be a group. Then, a subset  $S \subseteq G$  is called a generating set for the group  $G$  if every element of  $G$  can be expressed as a product of the elements of  $S$  or the inverses of the elements of  $S$ . There may be several generating sets for the same group. The largest generating set is the set of all group elements. For example, the subsets  $\{1\}$  and  $\{2, 3\}$  generate the group  $(\mathbb{Z}, +)$ .

A group with a specified set of generators  $S$  is called a *generated group* and is designated as  $(G, S)$ . If a group has a finite set of generators, it is called a *finitely generated group*. For example, the group  $\mathbb{Z}$  is a finitely generated group, for it has a finite generating set, for example,  $S = \{1\}$ . The generated group  $\mathbb{Z}$  with respect to the generating set  $\{1\}$  is usually designated as  $(\mathbb{Z}, \{1\})$ . The group  $(\mathbb{Q}, +)$  of rational numbers under addition cannot be finitely generated. Generators provide us with a *compact* representation of finitely generated groups, that is, a finite set of elements, which by the application of the group operation gives us the rest of the group.

We can now define a Cayley graph. Let  $(G, S)$  be a finitely generated group. Then the Cayley graph  $\Gamma(G, S)$  of a group  $G$  with respect to the choice of  $S$  is a directed colored graph, where vertices are identified with the elements of  $G$  and the directed edges of a color  $s$  connect all possible pairs of vertices  $(x, sx)$ ,  $x \in G$ ,  $s \in S$ .

In the following, we can see three examples of Cayley graphs: the Cayley graph for the first given example,  $(\mathbb{Z}, \{1\})$ , that is, an infinite chain (Fig. 22.7), another Cayley graph for the same group  $\mathbb{Z}$  with generators  $\{1, 2\}$ , which can be depicted as an infinite ladder (Fig. 22.8), and finally the Cayley graph for the group  $(\mathbb{Z}\{2, 3\})$  (Fig. 22.9). By *geometric* properties of groups, Starikova intends the properties of groups that can be revealed by thinking of their corresponding Cayley graphs as metric spaces. In other words, the idea is to look at groups *through* their Cayley graphs and try to see new (geometric) properties of groups, and then

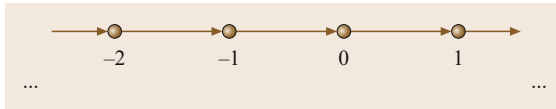


Fig. 22.7 The Cayley graph of the group  $(\mathbb{Z}\{1\})$

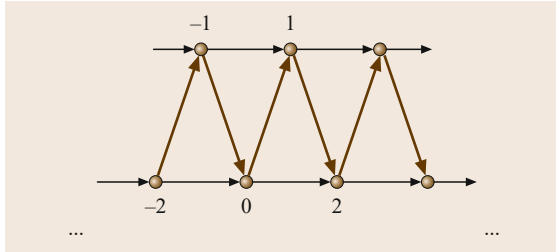


Fig. 22.8 The Cayley graph of the group  $(\mathbb{Z}\{1, 2\})$ , where bold stands for  $\{1\}$

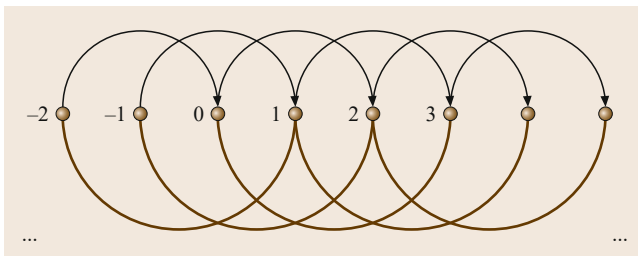


Fig. 22.9 The Cayley graph of the group  $(\mathbb{Z}\{2, 3\})$ , where bold stands for  $\{3\}$

to return to the algebra and check which groups share these properties and under which constraints. Many of these geometric properties turn out, in fact, to be independent from the choice of generators for a Cayley graph. For this reason, they are considered to be the properties of the groups themselves. Such a practice of introducing graphs to represent groups makes it possible to place groups in the same research–object category as classical metric spaces. This can happen because, in Manders’s terminology, we can be indifferent to the discrete structure of the group metric space and at the same time respond to the perceptual similarity of particular Cayley graphs having the same metric space. These responses would be unavailable to the combinatorial approach. Moreover, when responding geometrically to Cayley graphs, we perceive them as objects embedded in a space and having geometric elements. But then the response is *modified*, and some diagrammatic features are neglected in order to highlight more abstract properties. By introducing Cayley graphs, a group theorist thus has the opportunity to use them to define a metric of the group and then exploit its geometry, to define geometric counterparts to some algebraic properties of the group, and finally to clas-

sify groups having these geometric properties. This case study would show how sometimes the right choice of representation of an abstract object might lead to a significant development of a key concept.

### 22.5.3 Topology

Other case studies from contemporary mathematics concern topology.

The first one focuses on the identification and the discussion of the role of diagrammatic reasoning in *knot theory*, a branch of topology dealing with knots. A *knot* is a smooth closed simple curve in the Euclidean three-dimensional space, and a *knot diagram* is a regular projection of a knot with relative height information at the intersection points. *De Toffoli* and *Giardino* have discussed how knot diagrams are *privileged* points of view on knots: they display only a certain number of properties by selecting the relevant ones [22.50]. In fact, a single knot diagram cannot exhaust all the information about the knot type, and, for this reason, it is necessary to look at many diagrams of the same knot in order to *see* its different aspects. For example, both diagrams in Fig. 22.10 represent the unknot – that is, as the name suggests, a not knotted knot type – and we can transform the first into the second by *pulling* down the middle arc. However, this move alone does not allow us to conclude that both diagrams represent the unknot; to see that, we would have to apply further similar moves. In the article, a formalization for these possible modifications is provided.

The general idea behind this work is that diagrams are *kinaesthetic*, that is, their use is related to procedures and possible moves imagined on them. In topology, which is informally referred to as *rubber-band geometry*, a practitioner develops the ability to imagine continuous deformations. Manipulations of topological objects are guided by the consideration of concrete manipulations that would be performed on rubber or other deformable material. Accordingly, experts have acquired a form of imagination that prompt them to re-draw diagrams and calculate with them, performing “epistemic actions” [22.20]. This form of imagination

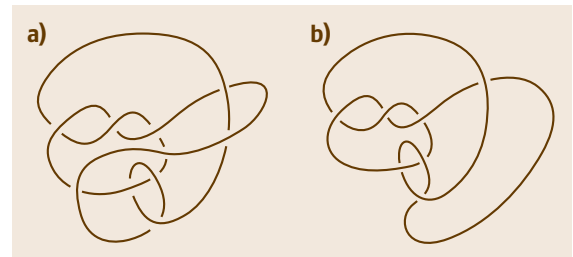


Fig. 22.10a,b Two nontrivial diagrams of the unknot

derives from our interaction with concrete objects and our familiarity with manipulating them. Moreover, the meaning of a knot diagram is fixed by its context of use: diagrams are the results of the interpretation of a figure, depending on the moves that are allowed on them and at the same time on the space in which they are embedded. Once the appropriate moves are established, the ambient space is fixed, thus determining the different equivalence relations. The context of use does not have to be predefined, preserving this kind of ambiguity that is not “damaging” [22.9], but productive. Actually, the indetermination of meaning makes different interpretations co-habit, and, therefore, allows attending to various properties and moves.

The same authors have also analyzed the practice of proving in low-dimensional topology [22.51]. As a case study, they have taken a specific proof: Rolfsen’s demonstration of the equivalence of two presentations of the Poincaré homology sphere. This proof is taken from a popular graduate textbook: *Knots and Links* by Rolfsen [22.52]. The first presentation of Poincaré homology sphere is a *Dehn surgery*, while the second one is a *Heegaard diagram* (Fig. 22.11).

Without going into the details, the aim of the authors is to use this case study to show that, analogously to knot theory, *seeing* in low-dimensional topology means imagining a series of possible manipulations on the representations that are used, and is, of course, modulated by expertise. Moreover, the actual practice of proving in low-dimensional topology cannot be reduced to formal statements without loss of intuition. Several examples of representationally heterogeneous reasoning – that is neither entirely propositional nor entirely visual – are given. Both the very representations introduced and the manipulations allowed on them – what the authors, following a terminology proposed by Larvor [22.53], call *permissible actions* – are epistemologically relevant, since they are integral parts both of the reasoning and the justification provided. To claim that inferences involving visual representations are permissible only within a specific practice is to consider them as context dependent. A consequence would be that it is no longer

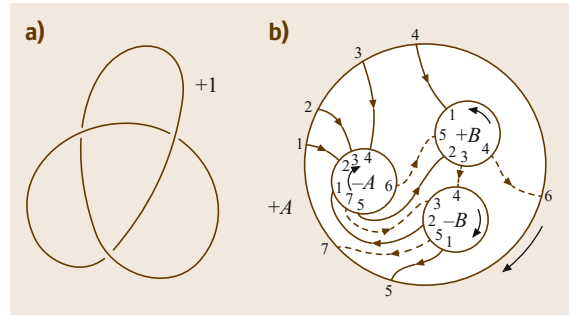


Fig. 22.11a,b The surgery code and the Heegaard diagram for the Poincaré homology sphere

possible to establish general criteria for mathematical validity, since they can only be local. The picture of mathematics emerging from these kinds of studies is thus very different from the one proposed from the post-nineteenth century philosophy of mathematics.

A final remark about representations in topology concerns a point about their materiality, already raised by Carter in a different context. To avoid confusion, it is necessary to keep in mind the distinction between the material pictures and the imagination process, which, especially in the case of trained practitioners, tends to vanish. Actual topological pictures trigger imagination and help see modifications on them, but experts may not find it necessary to actually draw all the physical pictures. The same holds for algebra where experts skip transitions that nontrained practitioners cannot avoid writing down explicitly. This does not mean that experts do not need pictures to grasp the reasoning, but only that, thanks to training and thus to their familiarity with drawing and manipulating pictures, they are sometimes able to determine what these pictures would look like even without actually drawing them. More generally, for each subfield, it would be possible to define a set of *background pictures* that are common to all practitioners, which would determine what Thurston has called the *mental model*. To go back to Netz’ analysis of the Euclidean diagram, here as well diagrams allow for procedures to be automatized or elided.

## 22.6 Computational Approaches

In this section, studies about the possibility of automatizing diagrammatic reasoning in mathematics are briefly introduced. Such attempts are worth being mentioned because they start from the observation that diagrammatic reasoning is crucial, at least in some ar-

eas of mathematics, and furthermore that any possible formalization for it should reflect its straightforwardness and directness. We will introduce the attempts of developing an automated reasoning program for plane geometry and for theory of numbers in turn.

### 22.6.1 (Manders') Euclid Reloaded

The analysis and the definitions provided by Manders about Euclidean geometrical reasoning were used to establish a formalization for diagrams in line with what he calls the diagram discipline. Such a project has brought about the creation of two logical systems, *E* [22.54] and *Eu* [22.55, 56], thanks to the work of Avigad, Dean and Mumma. Both systems produce formal derivations that line up closely with Euclid's proofs, in many cases following them step by step. (Another system that has been created to formalize Euclidean geometry is *FG* [22.57]. For details about *FG* and *Eu* and for a general discussion of the project in relation to model-based reasoning, see Chap. 23.). As summarized in a recent paper [22.58], the proof systems are designed to bring into sharp relief those attributes that are fundamental to Euclid's reasoning as characterized by Manders in his distinction between exact and co-exact properties. Nonetheless, the distinction is made with respect to a more restricted domain.

The Euclidean diagram has some components, which can be simple objects, such as points, lines, segments, and circles, and more complex ones, such as angles, triangles, and quadrilaterals. These components are organized according to some relations, which are the diagram attributes. Exact relations are obtained between objects having the same kind of magnitude: for example, for any two angles, the magnitude of one can be greater than the magnitude of the other or the same. Co-exact relations are instead positional: for example, a point can lie inside a region, outside it, or on its boundary. Co-exact relations concerning one-dimensional objects exclusively, such as line segments or circles, are intersection and nonintersection, while those concerning regions, one-dimensional or two-dimensional, are containment, intersection, and disjointness. Take the diagram in Fig. 22.12, representing the endpoint *A* as lying inside the circle *H* (a co-exact property), along with a certain distance between the point *A* and the circle's center *B* (an exact property). (Consider that in reproducing the diagram from Mumma's original article, the co-exact features were not affected, while the exact ones probably were.) Following Manders, in a proof in Euclid's system, premises and conclusions of diagrammatic inferences are composed of co-exact relations between geometric objects. In Fig. 22.13, an inference is shown (Fig. 22.13a) together with one of its possible associated diagrams (Fig. 22.13b).

In order to develop a formal system for these inferences, the main tasks in developing the programs were two: first, to specify the formal elements repre-

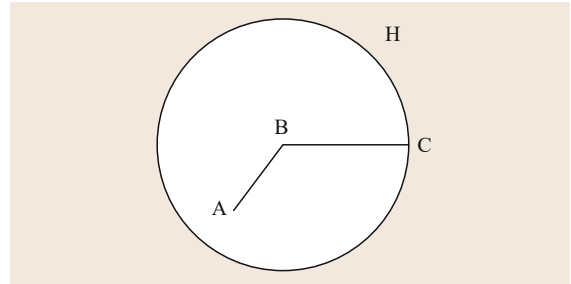


Fig. 22.12 A Euclidean diagram depicting exact and co-exact relations

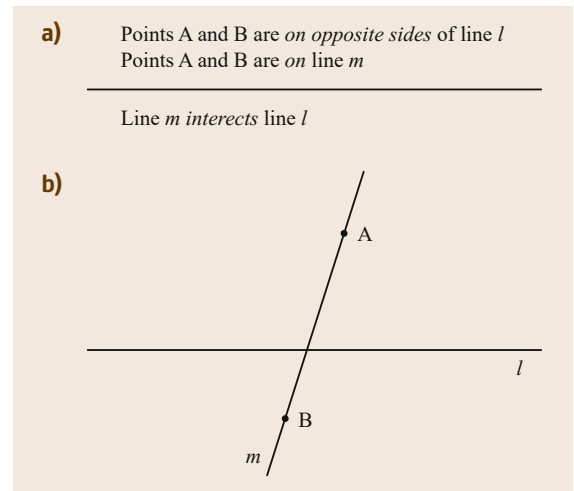


Fig. 22.13a,b An inference in Euclid's system according to Manders' reconstruction

sented co-exact relations; and second, to formulate the rules in terms of the elements whereby diagrammatic inferences can be represented in derivations. The main difference between *Eu* and *E* is how the first task is modulated. *Eu* possesses a diagrammatic symbol type intended to model what is perceived in concrete physical diagrams, while *E* models the information directly extracted from concrete physical diagrams by providing a list of primitive relations recording co-exact information among three object types: points, lines, and circles. In Fig. 22.14, the formalization in *Eu* of the inference in Fig. 22.13a is shown. In Fig. 22.15, the formalization of the same inference in *E* is shown, with the primitive *on(A, l)* meaning *point A is on line l*.

In addition, the formalizations do not only have formal elements corresponding to Euclidean diagrams, but also formal elements corresponding to the Euclidean text, so as to also record exact information. In order to give a proof, the two kinds of representations have to interact.

### 22.6.2 Theorem Provers

Not only formalizations of Euclidean geometry have been provided. Jamnik developed a semi-automatic proof system, called DIAMOND (*Diagrammatic Reasoning and Deduction*), to formalize and mechanize diagrammatic reasoning in mathematics, and in particular to prove theorems of arithmetic using diagrams [22.59]. Interestingly, Jamnik starts by recording a simple cognitive fact, that is that given some basic mathematical training and our familiarity with spatial manipulations – remember the study on knot theory – it suffices to look at the diagram representing a theorem to understand not only what particular theorem it represents, but also that it constitutes a proof for it. As a consequence, one arrives at the belief that the theorem is correct. From here, the question is: Is it possible to simulate and formalize this kind of diagrammatic reasoning on machines? In other words, is this an example of intuitive reasoning that is particular to humans and machines are incapable of?

The first part of Jamnik’s book provides a nice overview of the different diagrammatic reasoning systems that have been developed in the past century, such as, for example, *Gelernter’s Geometry Machine* [22.60] or *Koedinger and Anderson’s Diagram Configuration Model* [22.61]. For reasons of space, these systems will not be discussed here. In order to develop her proof system, she considers many different visual proofs in arithmetic and some of the analyses that have been given for them, by relying on the already mentioned collection edited by *Nelsen* [22.18, 19]. Such an analysis enables her to define a *schematic proof* as “a recursive function which outputs a proof of some proposition  $P(n)$  given some  $n$  as input” [22.59, p. 52].

Consider inductive theorems with a parameter, which, in Jamnik’s proposed taxonomy, are theorems where the diagram that is used to prove them represents one particular instance. An example of a theorem pertaining to this category is the *sum of squares of Fibonacci numbers*. According to this theorem, the sum of  $n$  squares of Fibonacci numbers equals the product of the  $n$ -th and  $(n + 1)$ -th Fibonacci numbers. In symbols,

$$Fib(n + 1) \times Fib(n) = Fib(1)^2 + Fib(2)^2 + \dots + Fib(n)^2. \quad (22.14)$$

The formal recursive definition of the Fibonacci numbers is given as

$$Fib(0) = 0, \quad Fib(1) = 1, \quad Fib(2) = 1, \\ Fib(n + 2) = Fib(n + 1) + Fib(n). \quad (22.15)$$

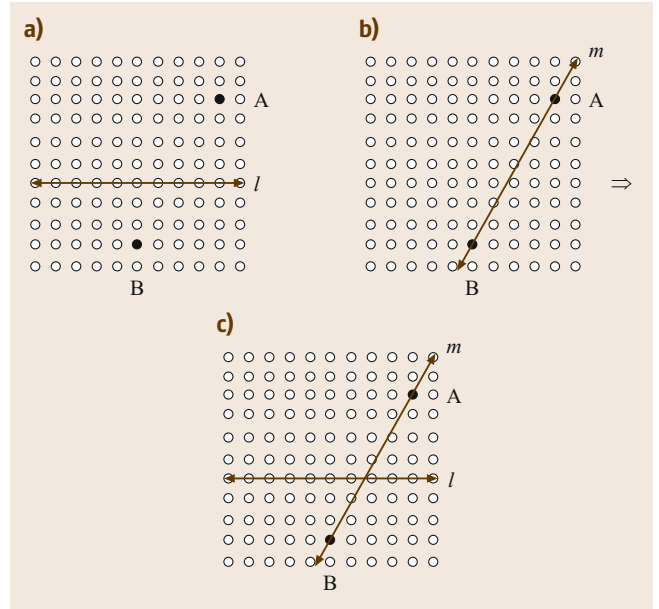


Fig. 22.14a–c The given inference in *EU*

a)	b)	c)
A, B points	on (A, m)	
l line	on (B, m)	intersects (l, m)
Not same side (A, B, l)		

Fig. 22.15a–c The given inference in *E*

Consider now Fig. 22.16. By looking at the spatial arrangement of the dots, we first take the rectangle of length  $Fib(n + 1)$  and height  $Fib(n)$ . Then, we split it in a square of magnitude  $Fib(n)$ , that is, the smaller side of the rectangle. We continue decomposing the remaining rectangle in a similar fashion until it is exhausted, that is, for all  $n$ . The sides of the created squares represent the consecutive Fibonacci numbers, and the longer side of every new rectangle is equal to the sum of the sides of two consecutive squares, which is precisely how the Fibonacci numbers are defined. As noted by Jamnik, the proof can also be carried out inversely, that is, starting from a square of unit magnitude ( $Fib(1)^2$ ) and joining it on one of its sides with another square of unit magni-

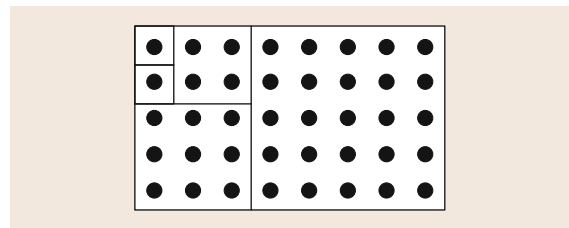


Fig. 22.16 Sum of squares of Fibonacci numbers



tude ( $Fib(2)^2$ ): we have a rectangle. Then we can take the rectangle and join to it a square of the magnitude of its longer side, so as to create another rectangle. The procedure can be repeated for all  $n$ .

The schematic diagrammatic proof for this theorem would then be a sequence of steps that need to be performed on the diagram in Fig. 22.16:

1. Split a square from a rectangle. The square should be of a magnitude that is equal to the smaller side of a rectangle (note that aligning squares of Fibonacci numbers in this way is a method of generating Fibonacci numbers, that is,  $1, 1, 1 + 1 = 2, 1 + 2 = 3, 2 + 3 = 5$ , etc.).
2. Repeat this step on the remaining rectangle until it is exhausted.

These steps are sufficient to transform a rectangle of magnitude  $Fib(n + 1)$  by  $Fib(n)$  to a representation of the right-hand side of the theorem, that is,  $n$  squares of magnitudes that are increasing Fibonacci numbers [22.59, p. 66].

## 22.7 Mathematical Thinking: Beyond Binary Classifications

The reader already acquainted with the topic of diagrammatic reasoning in mathematics might wonder why there has not yet been any explicit reference to the work of *Giaquinto*, who was undeniably one of the first philosophers to revive the attention toward mathematical visualization [22.62]. In one of his papers, we also find a nice overview of the literature concerning the possibility of obtaining rigorous proofs by reasoning diagrammatically [22.63]. The reason for this choice is that *Giaquinto* has not only been a pioneer in the renewed study of diagrammatic reasoning in mathematics, but also and even more interestingly he has given suggestions about the directions that future research should take. In this section, first his ideas on the role of visualization in mathematical discovery will be briefly presented, and then his proposal about how to consider mathematical thinking in general will be discussed (in the course of the final revisions of the present chapter, I discovered that *Giaquinto* recently published an entry on a topic that is related to ours, see for reference [22.64]).

As *Giaquinto* makes it clear, his original motivation for studying visual thinking in mathematics was to provide an epistemology of individual discovery and of actual mathematical thinking, so as to reopen the investigation of early thinkers from Plato to Kant, who indeed had as an objective to explore the nature of the individual's basic mathematical beliefs and skills.

A schematic proof is thus a schematic program which by instantiation at  $n$  gives a proof of every proposition  $P(n)$ . The constructive  $\omega$ -rule justifies that such a recursive program is indeed a proof of a proposition for all  $n$ . This rule is based on the  $\omega$ -rule, that is, an infinitary logical rule that requires an infinite number of premises to be proved in order to conclude a universal statement. The uniformity of this procedure is captured in the recursive program, for example,  $\text{proof}(n)$ . *Jamnik's* attempt is thus to formalize and implement the idea that the generality of a proof is captured in a variable number of applications of geometrical operations on a diagram, and as a consequence to challenge the argument according to which human mathematical reasoning is fundamentally noncomputational, and, therefore, cannot be automatized. Details about *DIAMOND's* functioning cannot be given here. We just point out that also in this case diagrammatic reasoning is interpreted as a series of operations on a particular diagram, which can be repeated on other diagrams displaying the same geometric features.

His strategy is thus first to acknowledge that there is more than one kind of thinking in mathematics, and then to assess the epistemic status of each of these kinds of mathematical thinking. For this reason, and as he himself admits, his view is in some sense much more *traditional* than many of the works produced by the post-nineteenth philosophers of mathematics. Discovery is a very crucial issue for the practice of mathematics and another topic that unfortunately has been neglected by post-nineteenth century approaches, which focused mainly on logic, proof, and justification. *Giaquinto* tries to give an account of the complexity of mathematical thinking, and to this aim he also inquires into fields of research other than philosophy, thus trespassing disciplinary boundaries. His belief is that cognitive science constitutes a new tool that can be helpful for understanding mathematical thinking: though cognition has always been the object of philosophy, the development of cognitive science surely represents an advantage for the philosophers of our century over the scholars of earlier times. Another discipline that could be an ally in disclosing mathematical thinking is mathematical education, traditionally categorized as an applied field unable to provide conclusive hints for theoretical research. Moreover, *Giaquinto* assigns an important role to history, both the history of mathematics and the history of philosophy.

The main epistemological thesis of the book is that there is no reason to assume a uniform evaluation that would fit all cases of visual thinking in mathematics, since visual operations are diverse depending on the mathematical context. Moreover, in order to assess this thesis, we do not need to refer to advanced mathematics: basic mathematics is already enough to account for the process of mathematical discovery by an individual who reasons visually. In fact, only the final part of the book goes beyond very elementary mathematics.

It should be mentioned that also Giaquinto defends a neo-Kantian view according to which in geometry we can find cases of synthetic a priori knowledge, that is cases that do not involve either analysis of meanings or deduction from definitions. In fact, he refers to the already mentioned study by *Norman*, which is neo-Kantian in spirit, as a strong case showing that following Euclid's proof of the proposition that the internal angles of a triangle sum to two right angles require visual thinking, and that visual thinking is not replaceable by nonvisual thinking [22.23]. Nonetheless, the focus in this section will be mostly on the last chapter of the book, where Giaquinto discusses how the traditional twofold division between *algebraic* thinking versus *geometric* thinking is not appropriate for accounting for mathematical reasoning. His conclusion, which can be borrowed also as a conclusion of the present chapter, is that there is a need for a much more comprehensive taxonomy for spatial reasoning in mathematics, which that would include operations such as visualizing motion, noticing reflection symmetry, and shifting aspects. In fact, if one considers thinking in mathematics as a whole, then there arises a sense of dissatisfaction with any of the common binary distinctions that have been proposed between algebraic thinking on the one hand and geometric thinking on the other; the philosopher's aim should be to move toward a much more discriminating taxonomy of kinds of mathematical reasoning.

Consider, for example, *aspect shifting* as precisely one form of mathematical thinking that seems to elude standard distinctions. Aspect shifting is the same cognitive ability that Macbeth describes in discussing the way in which the Greek geometer – and every one of us today who practices Euclidean geometry – reasoned in the Euclidean diagram. Take again the visual proof given in Sect. 22.1 for the Pythagorean theorem. As *Giaquinto* explains, it is possible to look at the square in Fig. 22.1b – that has letters in it, and, therefore, is a kind of *lettered diagram* – and see that the area of the larger square is equal to the area of the smaller square plus the area of the four right-angled triangles [22.62, pp. 240–241]. How do we acquire this belief? Giaquinto's reply is that first we have to reason *geometrically* and shift

between aspects, so as to recognize that the area of the square is both  $(a+b)^2$  and  $2ab+c^2$ . From here, we then have to proceed *algebraically* as follows

$$a^2 + 2ab + b^2 = 2ab + c^2, \quad a^2 + b^2 = c^2. \quad (22.16)$$

At this point, by looking back at the figure, we realize – *geometrically* again – that the smaller square is also the square of the hypotenuse of the right-angled triangle. Finally, from the formula, we conclude that the area of the square of the hypotenuse is equal to the sum of the squares of its other two sides. Then the question is: Is this argument as a whole to be considered as primary algebraic or geometric? It seems that neither of these two categories would be fully appropriate to capture it.

This is an interesting point also relative to other kinds of mathematical reasoning by means of some particular representation. Consider a notation that is used in topology and take as an example the *torus* that can be defined as a square with its sides identified. In order to obtain the torus from a square, we identify all its four sides in pairs. The square in Fig. 22.17a has arrows in it indicating the gluings, that is, the identifications. First, we identify two sides in the same direction, so as to obtain the *cylinder* (Fig. 22.17b); then, we identify the other two, again in the same direction: in Fig. 22.17c, one can see the torus with two marked curves, where the gluings, that is, the identifications, were made.

In discussing the role of notation in mathematics, *Colyvan* takes into consideration diagrams such as the one in Fig. 22.17a, and points out that this notation is “something of a halfway house between pure algebra and pure geometry” [22.65, p. 163]. In *Colyvan*'s view these diagrams are, on the one hand, a piece of notation, but, on the other, also an indication on how to construct the object in question. The first feature seems to belong to algebra, while the second to geometry. Moreover, note that if we identify two sides of the square

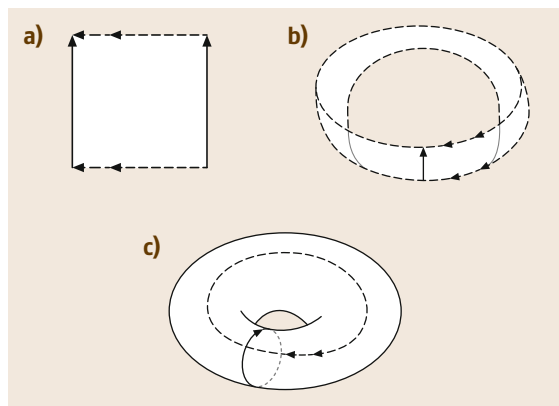


Fig. 22.17a–c Constructing the torus

in the same direction and the other two in the opposite direction, we obtain the *Klein bottle*, which is a very peculiar object, since it is three-dimensional but needs four spatial dimensions for its construction, and, even more interestingly, has no inside or outside. The Klein bottle demonstrates how powerful such a notation is: it leads to objects that would be otherwise considered as nonsense, and it also allows us to deduce their properties. As *Colyvan* summarizes, “Whichever way you look at it, we have a powerful piece of notation here that does *some genuine mathematical work for us*” ([22.65, p. 163], emphasis added). Diagrams, as well as other *powerful* notations, operate *at our place*. Moreover, at least some of them seem to be some kinds of *hybrid* objects, trespassing boundaries. They are geometric and algebraic at the same time.

Consider again the relations between the algebra of combinatorial groups and their geometry (Sect. 22.5.2). As *Starikova* tells us, first the combinatorial group theory was amplified with a geometric element – a graph –

where *geometric* refers mostly to geometric constructions as methods of geometry rather than algebra. But eventually this geometric element was significantly expanded and groups became geometric objects in virtue of their revealed geometric properties. The introduction of graphs thus provided mathematicians with a powerful instrument for facilitating their intuitive capacities and furthermore gave a good start for further intuitions which finally lead to advanced conceptual links with geometry and the definition of a broader geometric arsenal to algebra. Also in the knot theory example (Sect. 22.5.3), knot diagrams are shown to have at the same time diagrammatic and symbolic elements, and, as a consequence, their nature cannot be captured by the traditional dichotomy between geometric and algebraic reasoning. All this is to show that *Giaquinto’s* invitation to define a “more discriminating and more comprehensive” [22.13, p. 260] taxonomy for mathematical thinking going beyond twofold divisions is still valid, and that more on this topic needs to be done.

## 22.8 Conclusions

The objective of the present chapter was to introduce the different studies that have recently been devoted to diagrammatic reasoning in mathematics. The first topic discussed was the role of diagrams in Euclidean and Greek geometry in general (Sect. 22.3); then, the productive ambiguity of diagrams was defined (Sect. 22.4) and case studies in contemporary mathematics were briefly reviewed (Sect. 22.5). It has been shown how some attempts have tried to automatize diagrammatic reasoning in mathematics, in particular to formalize arguments in Euclidean geometry and proofs in theory of numbers (Sect. 22.6); finally, it has been argued that the attention to diagrammatic reasoning in mathematics can shed light on the fact that mathematics makes use of different kinds of representations that are so intertwined that it is difficult to draw sharp distinctions between the different subpractices and the corresponding reasoning (Sect. 22.7). We started from the study of diagrammatic reasoning and we arrived at the consideration of mathematical thinking as a whole, and of the role of notations and representations in it. Mathematicians use a vast range of cognitive tools to reason and communicate mathematical information; some of these tools are material, and, therefore, they can easily be shared, inspected, and reproduced. Specific representations are introduced in a specific practice and, once they enter into the set of the available tools, they may have an influence on the very same practice. This process plays a significant role in mathematics.

There is a last remark to make at the end of this survey, that is, that in diagrammatic reasoning, we have seen the continuity and the discreteness of space operating. Continuous was the space of the Euclidean diagram, discrete (at least in part) the space of the diagrams for Galileo’s theorem and for the sum of the Fibonacci numbers. Diagrammatic reasoning thus seems to have fundamentally a geometric nature, since it organizes space. Nonetheless, we have also shown that a diagram never comes alone, but always with some form of text giving indications for its construction or stipulating its correct interpretation. The relation between the diagram and text is defined each time by the specific practice. As a consequence, diagrams appear to be very interesting hybrid objects, whose nature cannot be totally captured by standard oppositions. They are cognitive tools available for thought, whose effectiveness depends on both our spatial and our linguistic cognitive nature.

**Acknowledgments.** My thanks go to the people quoted in the text, for their work and for the fruitful discussions in which I have taken part at recent conferences and workshops. I am particularly indebted to Mario Piazza and Silvia De Toffoli, with whom I have extensively reflected upon the topic of diagrammatic reasoning in mathematics. I am also grateful to Albrecht Heffer, who gave me the occasion of working on this chapter and to the *Université de Lorraine* and the *Région Lorraine* for having sustained my research.

## References

- 22.1 J.R. Brown: *Philosophy of Mathematics: An Introduction to the World of Proofs and Pictures* (Routledge, New York 1999)
- 22.2 D. Sherry: The role of diagrams in mathematical arguments, *Found. Sci.* **14**, 59–74 (2009)
- 22.3 S.-J. Shin: Heterogeneous reasoning and its logic, *Bull. Symb. Log.* **10**(1), 86–106 (2004)
- 22.4 E. Maor: *The Pythagorean Theorem. A 4000-Year History* (Princeton Univ. Press, Princeton 2007)
- 22.5 J. Høyrup: Tertium non datur: On reasoning styles in early mathematics. In: *Visualization, Explanation and Reasoning Styles in Mathematics, Synthese Library*, Vol. 327, ed. by P. Mancosu, K.F. Jørgensen, S.A. Pedersen (Springer, Dordrecht 2005) pp. 91–121
- 22.6 K. Chemla: The interplay between proof and algorithm in 3rd century China: The operation as prescription of computation and the operation as argument. In: *Visualization, Explanation and Reasoning Styles in Mathematics*, ed. by P. Mancosu, K.F. Jørgensen, S.A. Pedersen (Springer, Berlin 2005) pp. 123–145
- 22.7 K. Stenning, O. Lemon: Aligning logical and psychological perspectives on diagrammatic reasoning, *Artif. Intell. Rev.* **15**, 29–62 (2001)
- 22.8 J. Barwise, J. Etchemendy: Visual information and valid reasoning. In: *Logical Reasoning with Diagrams*, ed. by G. Allwein, J. Barwise (Oxford Univ. Press, Oxford 1996) pp. 3–25
- 22.9 S.-J. Shin, O. Lemon, J. Mumma: Diagrams. In: *The Stanford Encyclopedia of Philosophy*, ed. by E. Zalta, Fall 2013 Edition, <http://plato.stanford.edu/archives/fall2013/entries/diagrams/>
- 22.10 S.-J. Shin: The mystery of deduction and diagrammatic aspects of representation, *Rev. Philos. Psychol.* **6**, 49–67 (2015)
- 22.11 B. Russell: *The Principles of Mathematics* (W.W. Norton, London 1903/1937)
- 22.12 R. Netz: *The Shaping of Deduction in Greek Mathematics: A Study of Cognitive History* (Cambridge Univ. Press, Cambridge 1999)
- 22.13 M. Giaquinto: *The Search for Certainty* (Oxford Univ. Press, Oxford 2002)
- 22.14 F. Klein: *Elementary Mathematics from an Advanced Standpoint* (Dover, Mineola 2004), the first German edition is 1908
- 22.15 D. Hilbert: *The Foundations of Geometry* (K. Paul, Trench, Trübner, London 1899/1902)
- 22.16 P. Mancosu, K.F. Jørgensen, S.A. Pedersen (Eds.): *Visualization, Explanation and Reasoning Styles in Mathematics* (Springer, Berlin 2005)
- 22.17 V.F.R. Jones: A credo of sorts. In: *Truth in Mathematics*, ed. by H.G. Dales, G. Oliveri (Clarendon, Oxford 1998)
- 22.18 R. Nelsen: *Proofs without Words II: More Exercises in Visual Thinking*, Classroom Resource Materials (The Mathematical Association of America, Washington 2001)
- 22.19 R. Nelsen: *Proofs without Words: Exercises in Visual Thinking*, Classroom Resource Materials (The Mathematical Association of America, Washington 1997)
- 22.20 D. Kirsh, P. Maglio: On distinguishing epistemic from pragmatic action, *Cogn. Sci.* **18**, 513–549 (1994)
- 22.21 J. Ferreiros: *Mathematical Knowledge and the Interplay of Practices* (Princeton Univ. Press, Princeton 2015)
- 22.22 L.A. Shabel: *Mathematics in Kant's Critical Philosophy: Reflections on Mathematical Practice* (Routledge, New York 2003)
- 22.23 J. Norman: *After Euclid* (CSLI Publications, Univ. Chicago Press, Chicago 2006)
- 22.24 C.S. Peirce: *Collected Papers* (The Belknap Press of Harvard Univ. Press, Cambridge 1965)
- 22.25 J. Azzouni: Proof and ontology in Euclidean mathematics. In: *New Trends in the History and Philosophy of Mathematics*, ed. by T.H. Kjeldsen, S.A. Pedersen, L.M. Sonne-Hansen (Univ. Press of Southern Denmark, Odense, Denmark 2004) pp. 117–133
- 22.26 W.P. Thurston: On proof and progress in mathematics, *Bull. Am. Math. Soc.* **30**(2), 161–177 (1994)
- 22.27 P. Mancosu (Ed.): *The Philosophy of Mathematical Practice* (Oxford Univ. Press, Oxford 2008)
- 22.28 K. Manders: The Euclidean diagram. In: *The Philosophy of Mathematical Practice*, ed. by P. Mancosu (Oxford Univ. Press, Oxford 2008) pp. 80–133
- 22.29 K. Manders: Diagram-based geometric practice. In: *The Philosophy of Mathematical Practice*, ed. by P. Mancosu (Oxford Univ. Press, Oxford 2008) pp. 65–79
- 22.30 D. Macbeth: Diagrammatic reasoning in Euclid's elements. In: *Philosophical Perspectives on Mathematical Practice*, Vol. 12, ed. by B. Van Kerkhove, J. De Vuyst, J.P. Van Bendegem (College Publications, London 2010)
- 22.31 H.P. Grice: Meaning, *Philos. Rev.* **66**, 377–388 (1957)
- 22.32 P. Catton, C. Montelle: To diagram, to demonstrate: To do, to see, and to judge in Greek geometry, *Philos. Math.* **20**(1), 25–57 (2012)
- 22.33 D. Macbeth: Diagrammatic reasoning in Frege's *Begriffsschrift*, *Synthese* **186**, 289–314 (2012)
- 22.34 M. Panza: The twofold role of diagrams in Euclid's plane geometry, *Synthese* **186**(1), 55–102 (2012)
- 22.35 M. Panza: Rethinking geometrical exactness, *Hist. Math.* **38**, 42–95 (2011)
- 22.36 C. Parsons: *Mathematical Thought and Its Objects* (Cambridge Univ. Press, Cambridge 2008)
- 22.37 P. Mancosu (Ed.): *From Brouwer to Hilbert. The Debate on the Foundations of Mathematics in the 1920s* (Oxford Univ. Press, Oxford 1998)
- 22.38 Proclus: *In primum Euclidis Elementorum librum commentarii* (B.G. Teubner, Leipzig 1873), ex recognitione G. Friedlein, in Latin
- 22.39 Proclus: *A Commentary on the First Book of Euclid's Elements* (Princeton Univ. Press, Princeton 1992), Translated with introduction and notes by G.R. Morrow
- 22.40 Aristotle: *Metaphysics*, Book E, 1026a, 6–10
- 22.41 E. Grosholz: *Representation and Productive Ambiguity in Mathematics and the Sciences* (Oxford

- Univ. Press, Oxford 2007)
- 22.42 K. Chemla: Lazare Carnot et la Généralité en Géométrie. Variations sure le Théorème dit de Menelaus, *Rev. Hist. Math.* **4**, 163–190 (1998), in French
- 22.43 J. Carter: Diagrams and proofs in analysis, *Int. Stud. Philos. Sci.* **24**(1), 1–14 (2010)
- 22.44 J. Carter: The role of representations in mathematical reasoning, *Philos. Sci.* **16**(1), 55–70 (2012)
- 22.45 U. Haagerup, S. Thorbjørnsen: Random matrices and  $K$ -theory for exact  $C^*$ -algebras, *Doc. Math.* **4**, 341–450 (1999)
- 22.46 M.E. Moore: *New Essays on Peirce's Mathematical Philosophy* (Open Court, Chicago and La Salle 2010)
- 22.47 K. Manders: Euclid or Descartes: Representation and responsiveness, (1999), unpublished
- 22.48 I. Starikova: Why do mathematicians need different ways of presenting mathematical objects? The case of Cayley graphs, *Topoi* **29**, 41–51 (2010)
- 22.49 I. Starikova: From practice to new concepts: Geometric properties of groups, *Philos. Sci.* **16**(1), 129–151 (2012)
- 22.50 S. De Toffoli, V. Giardino: Forms and roles of diagrams in knot theory, *Erkenntnis* **79**(4), 829–842 (2014)
- 22.51 S. De Toffoli, V. Giardino: An inquiry into the practice of proving in low-dimensional topology. In: *From Logic to Practia*, (Springer, Cham 2015) pp. 315–336
- 22.52 D. Rolfsen: *Knots and Links* (Publish or Perish, Berkeley 1976)
- 22.53 B. Larvor: How to think about informal proofs, *Synthese* **187**(2), 715–730 (2012)
- 22.54 J. Avigad, E. Dean, J. Mumma: A formal system for Euclid's elements, *Rev. Symb. Log.* **2**(4), 700–768 (2009)
- 22.55 J. Mumma: Proofs, pictures, and Euclid, *Synthese* **175**(2), 255–287 (2010)
- 22.56 J. Mumma: Intuition formalized: Ancient and modern methods of proof in elementary geometry, Ph.D. Thesis (Carnegie Mellon University, Pittsburgh 2006)
- 22.57 N. Miller: *Euclid and His Twentieth Century Rivals: Diagrams in the Logic of Euclidean Geometry* (CSLI Publications, Stanford 2007)
- 22.58 Y. Hamami, J. Mumma: Prolegomena to a cognitive investigation of Euclidean diagrammatic reasoning, *J. Log. Lang. Inf.* **22**, 421–448 (2013)
- 22.59 M. Jamnik: *Mathematical Reasoning with Diagrams* (Univ. Chicago Press, Chicago 2002)
- 22.60 H. Gelernter: Realization of a geometry theorem-proving machine. In: *Computers and Thought*, ed. by E. Feigenbaum, J. Feldman (Mac Graw Hill, New York 1963) pp. 134–152
- 22.61 K.R. Koedinger, J.R. Anderson: Abstract planning and perceptual chunks, *Cogn. Sci.* **14**, 511–550 (1990)
- 22.62 M. Giaquinto: *Visual Thinking in Mathematics* (Oxford Univ. Press, Oxford 2007)
- 22.63 M. Giaquinto: Visualizing in mathematics. In: *The Philosophy of Mathematical Practice*, ed. by P. Mancosu (Oxford Univ. Press, Oxford 2008) pp. 22–42
- 22.64 M. Giaquinto: The epistemology of visual thinking in mathematics. In: *The Stanford Encyclopedia of Philosophy*, ed. by E.N. Zalta, Winter 2015 Edition, <http://plato.stanford.edu/archives/win2015/entries/epistemology-visual-thinking/>
- 22.65 M. Colyvan: *An Introduction to the Philosophy of Mathematics* (Cambridge Univ. Press, Cambridge 2012)

## 23. Deduction, Diagrams and Model-Based Reasoning

John Mumma

A key piece of data in understanding mathematics from the perspective of model-based reasoning is the use of diagrams to discover and to convey mathematical concepts and proofs. A paradigmatic example of such use is found in the classical demonstrations of elementary Euclidean geometry. These are invariably presented with accompanying geometric diagrams. Great progress has been made recently with respect to the precise role the diagrams plays in the demonstrations, so much so that diagrammatic formalizations of elementary Euclidean geometry have been developed. The purpose of this chapter is to introduce these formalizations to those who seek to understand mathematics from the perspective of model-based reasoning.

The formalizations are named **FG** and **Eu**. Both are based on insights articulated in Ken Manders'

23.1	<b>Euclid's Systematic Use of Geometric Diagrams</b> .....	524
23.2	<b>Formalizing Euclid's Diagrammatic Proof Method</b> .....	525
23.2.1	The Formal System <b>FG</b> .....	526
23.2.2	The Formal System <b>Eu</b> .....	528
23.3	<b>Formal Geometric Diagrams as Models</b> .....	532
	<b>References</b> .....	534

seminal analysis of Euclid's diagrammatic proofs. The chapter presents these insights, the challenges involved in realizing them in a formalization, and the way **FG** and **Eu** each meet these challenges. The chapter closes with a discussion of how the formalizations can each be thought to prespecify a species of model-based reasoning.

The formalization of mathematical knowledge has been a mainstay of the philosophy of mathematics since the end of the nineteenth century. The goals and assumptions characteristic of the enterprise are foundationalist. A piece of mathematics is formalized to obtain a clear view of it within the context of justification. The various lemmas, theorems and corollaries of the mathematics are translated into sentences of a fixed formal language, whereby it becomes possible to ascertain with precision the logical relationships of the lemmas, theorems and corollaries to one another and to a group of sentences distinguished as axioms. The end result is a picture of the how the mathematics is – or at least can be – grounded on a collection of its basic truths.

Formalization would thus seem to be of little use to those who seek to understand mathematics from the perspective of model-based reasoning. The goal from this perspective is to obtain a clear view of the mathematics within the context of discovery. What is of interest is how the lemmas, theorems, and corollaries of the mathematics came to be known in the first place. A fundamental premise is that the process is a *reasoning* process, where the reasoning involved is

different in kind from the strictly regulated procedures of inference prescribed by a formalization. Inference concerning some mathematical subject is driven by the reasoner's engagement with representations modeling  $X$ , rather than the logical form of sentences expressing claims about  $X$ . For a paradigmatic example of such an  $X$  consider elementary geometry. From the perspective of the tradition that investigates mathematical knowledge via formalization, what is fundamental are *sentences* expressing the axioms and theorems of elementary geometry in a fixed formal language. Geometric reasoning is depicted as a progression of sentences laid out along the rigid pathways defined by the formalization's rules. From the perspective of model-based reasoning, what is fundamental are the *diagrams* that model the geometric situations that the axioms and theorems concern. Geometric reasoning is an open-ended process in which mind and diagram interact.

A curious recent development has been the appearance of formalizations advanced to show that diagrams of elementary geometry can be understood as part of the formal syntax of the subject's proofs. These specifically

are the proof systems **FG** [23.1] and **Eu** [23.2]. Since the target of the formalizations is the use of diagrams in proving geometric theorems, one may think that they provide a model-based reasoning picture of mathematical proof. At the same time, by the very fact that they are formalizations, one may take them to miss what is important about geometric diagrams from the perspective of model-based reasoning. Perhaps the formal objects identified as geometric diagrams within them are best understood as sentences formulated with an unconventional notation.

In this chapter I present **FG** and **Eu** with the aim of illuminating their potential relevance to researchers

in model-based reasoning. For the presentation of a formal proof system closely related to **Eu** – termed *E* – see [23.3]. For a discussion of the utility of the analyses of **Eu** and *E* in understanding Euclidean diagrammatic reasoning from a cognitive perspective, see [23.4]. The formalizations are based on principles formulated in [23.5], *Ken Manders’* seminal analysis of Euclid’s diagrammatic proof method in the *Elements*. In the chapter’s first section I present this analysis. In the second, I sketch **FG** and **Eu** as formal proof systems. Finally, in the chapter’s third section, I advance an interpretation of the systems where each characterizes a kind of model-based reasoning.

## 23.1 Euclid’s Systematic Use of Geometric Diagrams

For most of its long history, Euclid’s *Elements* was the paradigm for careful and exact mathematical reasoning. In the past century, however, it has been just the opposite. Its proofs are often invoked to illustrate what rigor in mathematics does *not* consist of. Though some steps of Euclid’s proofs are respectable as logical inferences, a good many are not. With these, one cannot look only at the logical form of the claims in the proof and understand what underlies them. One is forced, rather, to look at the accompanying diagram. The modern opinion is that Euclid’s proofs exhibit deductive gaps at such places.

*Ken Manders* set out to explode this story in [23.5]. His analysis of Euclid’s diagrammatic proof method reveals that Euclid employs diagrams in a controlled, systematic way. It thus calls into question the common, negative assessment of the rigor of the *Elements*. Moreover, the specifics of Manders’ analysis suggest that Euclid’s proof can be understood to adhere to a formal diagrammatic logic.

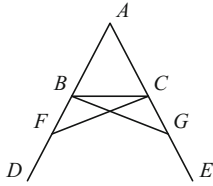
The first step in understanding the role of diagrams in Euclid’s *Elements* is simply to recognize them as components of proofs, rather than as mere illustrations of the proofs’ sentences. Accordingly, Manders characterizes the proofs as proceeding along two tracks: a discursive one resulting in a sequence of assertions, and a graphic one resulting in a geometric diagram. A proof step within the discursive, or sentential, track consists of the addition of a new assertion to the sequence, and a proof step within the diagrammatic track consists of the addition of a new graphic object to the diagram. The graphic objects in the diagram are linked to assertions in the text via labels. Crucially, it is not just previous assertions that license the addition of a new assertion within the sentential track; the diagram can directly license such an addition as well.

Central to Manders’ account of how diagrams do this is his distinction between the *exact* and *co-exact* features of diagrams. Being drawn by human hands, Euclidean diagrams cannot fully realize the properties geometric lines and circles are conceived to possess. There are bound to be small variations from the ideal of perfectly straight lines and perfectly circular circles in the diagrams produced by ruler and compass. Manders defines as *co-exact* any attribute of a diagram that is stable under such variations or perturbations. Exact attributes are then whatever is not so stable. In a diagram in which a segment is extended from a vertex of a triangle to the side opposite the vertex, for example, the triangle’s containment of the segment is a *co-exact* attribute of the diagram. The precise ratio between the two angles induced by the segment is not.

Manders’ key observation is that Euclid’s diagrams contribute to proofs only through their *co-exact* attributes. Euclid never infers an exact attribute from a diagram unless it follows directly from a *co-exact* attribute. Geometric claims concerning other exact attributes either are assumed from the outset or are proved via a chain of inferences in the text. It is not difficult to hypothesize why Euclid would have restricted himself in such a way. It is only in their capacity to exhibit *co-exact* relations that diagrams seem capable of functioning effectively as symbols of proof. The exact relations depicted by diagrams are too refined to be easily reproducible and to support determinate judgments [23.5, Sect. 4.2.2].

For an example of how diagrams carry *co-exact* information for Euclid, consider the proof of proposition 5 of the *Elements*. The proposition states that all isosceles triangles have equal base angles. The proof as given in [23.6] is as follows:

*Proof:*



Let  $ABC$  be an isosceles triangle having the side  $AB$  equal to the side  $AC$ ; and let the straight lines  $BD$ ,  $CE$  be produced further in a straight line with  $AB$ ,  $AC$  (postulate 2).

I say that the angle  $\angle ABC$  is equal to the  $\angle ACB$ , and the angle  $\angle CBD$  to the angle  $\angle BCE$ .

Let a point  $F$  be taken at random on  $BD$ ; from  $AE$  the greater let  $AG$  be cut off equal to  $AF$  the less (proposition I, 3); and let the straight lines  $FC$ ,  $GB$  be joined (postulate 1).

Then, since  $AF$  is equal to  $AG$  and  $AB$  to  $AC$ , the two sides  $FA$ ,  $AC$  are equal to the two sides  $GA$ ,  $AB$ , respectively; and they contain a common angle, the angle  $\angle FAG$ . Therefore, the base  $FC$  is equal to the base  $GB$ , and the triangle  $AFC$  is equal to the triangle  $AGB$ , and the remaining angles will be equal to the remaining angles respectively, namely those which the equal sides subtend, that is, the angle  $\angle ACF$  to the angle  $\angle ABG$ , and the angle  $\angle AFC$  to the angle  $\angle AGB$  (proposition I, 4).

And, since the whole  $AF$  is equal to the whole  $AG$ , and in these  $AB$  is equal to  $AC$ , the remainder  $BF$  is equal to the remainder  $CG$ .

But  $FC$  was also proved equal to  $GB$ ; therefore the two sides  $BF$ ,  $FC$  are equal to the two sides  $CG$ ,  $GB$  respectively; and the angle  $\angle BFC$  is equal to the angle  $\angle CGB$ , while the base  $BC$  is common to them;

therefore the triangle  $BFC$  is also equal to the triangle  $CGB$ , and the remaining angles will be equal to the remaining angles respectively, namely those which the equal sides subtend; therefore angle  $\angle FBC$  is equal to the angle  $\angle GCB$ , and the angle  $\angle BCF$  to the angle  $\angle CBG$ .

Accordingly, since the whole angle  $\angle ABG$  was proved equal to the angle  $\angle ACF$ , and in these the angle  $\angle CBG$  is equal to the angle  $\angle BCF$ , the remaining angle  $\angle ABC$  is equal to the remaining angle  $\angle ACB$ ; and they are at the base of the triangle  $ABC$ . But the angle  $\angle FBC$  was also proved equal to the angle  $\angle GCB$ ; and they are under the base. ■

Two steps of the proof rely on the diagram. The first is the application of the equals-subtracted-from-equals rule (common notion 3 in the *Elements*) to infer the equality of lengths  $BF$  and  $CG$ , the second is the application of the same rule to infer the equality of angles  $\angle ABC$  and  $\angle ACB$ . A requirement for the correct application of the common notion is that certain co-exact containment relations hold. In order to apply equals-subtracted-from-equals in the last step for instance, angle  $\angle ABG$  is required to contain  $\angle ABC$  and  $\angle CBG$ , and angle  $\angle ACF$  is required to contain  $\angle ACB$  and  $\angle BCF$ . On Manders' account the diagram of the proof licenses the inference that these co-exact conditions are satisfied.

Generally, the results of elementary geometry depend on nonmetric positional relations holding between the components of a configuration. A method for proving the results, then, must provide a means for recording such information about a configuration, and grounding inferences with respect to it. According to Manders' account of Euclid's method, diagrams fulfill this function, and do so in a mathematically legitimate way – i. e., they do not compromise the rigor of the method.

## 23.2 Formalizing Euclid's Diagrammatic Proof Method

Both **FG** and **Eu** were created to flesh out Manders' account in formal terms by characterizing Euclid's diagrams as syntactic objects in a formal proof system. Carrying the project out amounts to three tasks:

1. The definition of a class of syntactical objects that serve to represent Euclid's diagrams in the formal system
2. The specification of the geometric information expressible with the formal diagrams of the system
3. The specification of a method for regulating the information expressible by formal diagrams of the system in geometric proofs.

The structure of the concrete diagrams accompanying presentations of Euclid's arguments defines the first task. The abstract formal diagrams of the formal proof system ought somehow to embody this structure. How the formal system accomplishes task 2 constitutes the system's analysis of Manders' notion of co-exactness. With such an analysis in place, the third task becomes possible. The method furnished by the proof system for employing diagrams in proofs must be geometrically sound relative to the geometric information expressed by the system's diagrams.

Below I describe how **FG** and **Eu** fulfill tasks 1)–3). Beforehand it is worth discussing in more detail what

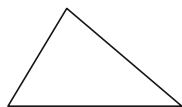


task 3 involves, as its successful completion amounts to a solution of the *generality* problem surrounding Euclid's proofs.

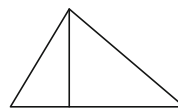
The generality problem arises with Euclid's proofs because the diagram used for a proof is always a *particular* diagram. Euclid clearly did not intend his propositions to concern just the figure on display beside the proposition. They are applied in subsequent proofs to other figures which are not exact duplicates of the original. And so, for Euclid, consultation of the original diagram, with all its particular features, is somehow supposed to license a generalization. But Euclid leaves the process by which this is done obscure. And so we are left with some doubt as to whether the jump from the particular to general is justified. Even before the nineteenth century, when the legitimacy of Euclid's methods was taken for granted, philosophers recognized that there was something to be explained with this jump.

Manders' exact/co-exact distinction provides the basis for a partial explanation. The co-exact properties of a diagram can be shared by all geometric configurations in the range of a proof, and so in such cases one is justified in reading off co-exact properties from the diagram. In a proof about triangles for instance, variation among the configurations in the range of the proof is variation of exact properties – e.g., the measure of the triangles' angles or the ratios between their sides. They all share the same co-exact properties – i.e., they all consist of three bounded linear regions which together define an area.

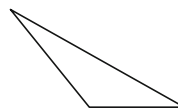
This is not a full answer because Euclid's proofs typically involve constructions on an initial configuration type. With the proof of proposition 5, for example, a construction on a triangle is specified. In such cases, a diagram may adequately represent the co-exact properties of an initial configuration. But the result of applying a proof's construction to the diagram cannot be assumed to represent the co-exact properties of all configurations resulting from the construction. One does not need to consider complex geometric constructions to see this. Suppose for instance the initial configuration type of a proof is a triangle. Then the diagram



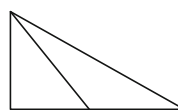
serves to represent the co-exact properties of this type. Suppose further that the first step of a proof's construction is to drop the perpendicular from a vertex of the triangle to the line containing the side opposite the vertex. Then the result of carrying this step out on the diagram, i.e.,



ceases to be representative. That the perpendicular falls within the triangle in the diagram is a co-exact feature of it. But there are triangles with exact properties different from the initial diagram where applying the construction step results in a perpendicular lying outside the triangle. For example, with the triangle



the result of applying the construction step is



And so, carrying out a Euclidean construction on a representative diagram can result in an unrepresentative diagram. If a formal system is to provide a compelling analysis of Euclid's diagrammatic proofs it must account for this in carrying out task 3.

### 23.2.1 The Formal System FG

#### Task 1 in FG: FG Diagrams

The four fundamental syntactical notions of **FG** are *frame*, *dot*, *solid segment*, and *dotted segment*. Every **FG** diagram possesses a frame, characterized as “a rectangular box drawn in the plane” [23.1, p. 22]). Within it dots, solid segments, and dotted segments can lie. The dots of an **FG** diagram are point-like graphic objects. Solid segments and dotted segments are one-dimensional graphic objects that do not intersect any other objects of the diagram and terminate either in dots or the diagram's frame. Solid segments serve to represent line segments, and dotted segments serve to represent arcs of circles. Accordingly, an **FG** diagram comes equipped with a partition on its set of solid segments and a partition on its set of its dotted segments. The *dlines* of the diagram are the components of the former partition, and the *dcircles* of the diagram are the components of the latter. See Fig. 23.1 for an example of an **FG** diagram.

Aside from the requirement that solid and dotted segments do not intersect anything (including themselves), there are no constraints imposed upon them. They are free to bend and curve any which way between the dots that bound them. Consequently, there is no upper bound on the number of times sets of such objects can intersect one another at dots within an **FG**

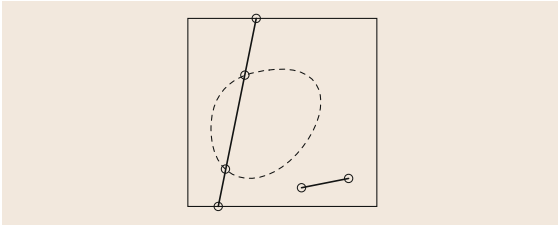


Fig. 23.1 Example FG diagram

frame. There are upper bounds, however, on the number of times Euclidean lines and circles can intersect one another in points. Two distinct lines, for instance, intersect in at most one point. Thus, so that dlines and dcircles intersect one another like Euclidean lines and circles, they are required to satisfy a variety of conditions. One of these conditions, for instance, ensures that two dlines do not intersect at more than one dot in an FG diagram. For the details see [23.1, Sect. 2.1]. As illustrated in the discussion of FG's completion of task 3 below, such conditions play an essential role in FG's formalization of Euclid's proofs.

### Task 2 in FG: Corresponding Graph Structures

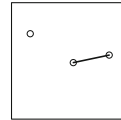
These conditions notwithstanding, the FG definition of a diagram leaves room for a great deal of variation. Consider those FG diagrams with a fixed number of dots, dlines and dcircles – say  $k$ ,  $m$  and  $n$ . The range of ways such objects can differ from one another is vast. Task 2 within FG amounts to specifying which differences matter in proofs. The task, in particular, is to specify which differences among FG diagrams with  $k$  dots,  $m$  dlines, and  $n$  dcircles express geometric differences among planar configurations of  $k$  points,  $m$  lines and  $n$  circles.

This is done in topological terms. What matters about a FG diagram in a proof, roughly, are the way its dlines and dcircles divide the region defined by the frame into smaller connected regions, and the position (inside or outside) of its dots and dsegments relative to these regions. The notion making this precise is that of a FG diagram's *corresponding graph structure* [23.1, Sect. 2.2]. It is via this notion that the representational link is made between FG diagrams as proof symbols and Euclidean configurations as mathematical objects. Any Euclidean configuration of points, lines and circles can be understood as an FG diagram by enclosing the configuration within a suitably large rectangle. Consequently, any Euclidean configuration has its own corresponding graph structure (as any suitably large rectangle is decomposed into regions by the configuration in the same way that the frame of an FG diagram is decomposed by its dlines and dcircles). What an FG diagram represents, then, are all Euclidean

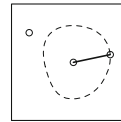
configurations that have the diagram's corresponding graph structure.

### Task 3 in FG: FG Case Analysis

The generality problem arises in FG as follows. The basic Euclidean operations of constructing a segment, extending a segment, and constructing a circle are paralleled within FG by the syntactic operations of adding a dline, extending a dline, and adding a dcircle. And so for any Euclidean construction there is a parallel FG construction. Yet Euclidean constructions, in general, do not yield unique corresponding graphs structures. Consequently, if we are given a Euclidean construction and produce an FG diagram  $D$  according to the parallel FG construction, we cannot assume that the corresponding graph structure of  $D$  is shared by all configurations produced by the construction. For a simple example, suppose that  $D'$  is the diagram



and a circle is constructed on the segment as radius to obtain  $D$

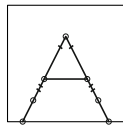


The Euclidean configurations that have  $D'$ 's corresponding graph structure are simply those consisting of a segment and a point off of it. Constructing a circle from the segment in each such configuration does not always result in a configuration with the corresponding graph structure of  $D$ . The point may not lie outside the circle.

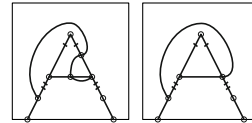
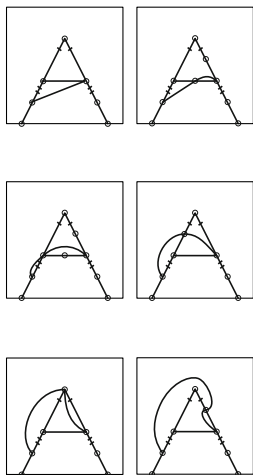
The problem is resolved in FG via the implementation of a uniform case-branching method. Suppose that within an FG derivation a construction step is to be applied to a configuration represented by the diagram  $D'$ . In performing the construction step, one must not only add to  $D'$  the object of the construction step to obtain a diagram  $D$ . One must also produce diagrams for all FG cases that could result from the construction step, where the notion of 'FG case' is specified in terms of the notions of a corresponding graph structure and an FG diagram. Specifically, if  $E'$  is a diagram with the same corresponding graph structure as  $D'$  and  $E$  is an FG diagram obtained by adding the object of the construction step to  $E'$ , then the corresponding graph structure of  $E$  is an FG case of the construction step. Since the definition of an FG diagram does not require

dlines to be straight or dcircles to be circular, there may be **FG** cases that are not Euclidean cases – i. e., some corresponding graph structures may not be realized by any Euclidean configuration. But all Euclidean cases will be an **FG** case. And so a relation that appears in all **FG** cases of the construction step holds in all Euclidean cases of the construction step.

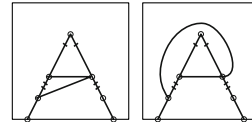
Accordingly, the **FG** method for isolating the invariant co-exact relations of a construction is to produce all **FG** cases of the construction, and determine which co-exact relations are obtained in all cases. Such a method is satisfactory, of course, only if there is a procedure for producing all the **FG** cases of the construction. *Miller* has implemented such a procedure in a computer program **CDEG**, the general principles behind which he describes in section 3.5 of [23.1]. To understand how the procedure works, consider the step in the construction of proposition I, 5 in which the segment from *F* to *C* is added. The configuration on which the construction is performed can be represented by the following **FG** diagram (the hashmarks in the diagram represent equality of lengths according to the standard convention).



The parallel construction step in **FG** is to add a dline connecting the points representing *F* and *G*. The cases that result from this step are individuated by the ways the new dline can snake through the regions of the corresponding graph structure of *D'*. If the dline were conceived simply as a one-dimensional curve, some of these cases would be



A dline, however, is a one-dimensional curve required to satisfy certain constraints. One of these, as pointed out above, is that a dline cannot intersect another dline in more than one dot. This eliminates all but the first and last case



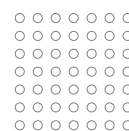
Generally, **FG** cases can arise when a dline is added to a diagram, when a dcircle is added to a diagram, or when a dline is extended. For each possibility, the conditions on dlines and dcircles are such that the possible routes of the added element through the corresponding graph structure of the diagram are sufficiently restricted – i. e., the resulting **FG** cases are finite in number and can be systematically enumerated.

A side note: **FG** is a purely diagrammatic formal system. Thus, the techniques whereby one recognizes parts of an **FG** diagram (e.g., its dots) in terms of Euclid’s verbal presentation of a construction (e.g., *the point F*, *the point G*) are taken from the beginning to be external to it. As discussed in the next section, **Eu** is a heterogeneous system – i. e., it possesses both a diagrammatic and a sentential syntax. Sentential symbols label its diagrams, and formalize (to a certain extent) a means for relating sentential and diagrammatic representations. The labels also provide a means for classifying two diagrams as equivalent with respect to the geometric information they express. Nothing in the definition of **FG** diagrams prevents the development of a heterogeneous version of **FG** with these features.

### 23.2.2 The Formal System Eu

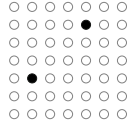
#### Task 1 in Eu: Eu Diagrams

Common to all **Eu** diagrams is a discrete two-dimensional array structure that serves to model the spatial background within which diagrammatic points, lines and circles are constructed. The arrays are square and of arbitrary finite dimension. An example is the  $7 \times 7$  array

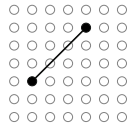


The array elements are identified by their coordinates, with the lowest, left-most element having coordinates  $(0, 0)$  and the highest, right-most element in an  $n \times n$  array having coordinates  $(n - 1, n - 1)$ .

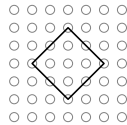
A point of an **Eu** diagram is a distinguished array element. An example of an **Eu** diagram with two points is



A linear element of an **Eu** diagram is a linear subset of its array elements – i. e., a subset of array elements whose coordinates satisfy a linear equation. The elements of a linear element can be further constrained by inequalities on its first or second coordinate. If there is one inequality to be satisfied the linear element is a ray; if there are two it is a segment. The linear element of the diagram below is a segment defined by the conditions:  $y = x + 1, 1 \leq x \leq 5$



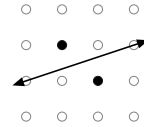
Finally, a circle of an **Eu** diagram is a subset of array elements that form the perimeter of a convex polygon.



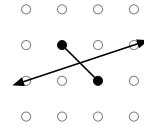
The correspondence between the abstract, formal diagrams of **Eu** and the concrete diagrams they model is not one-to-one, but many-to-one. Specifically, a single concrete diagram is modeled in **Eu** by a set of **Eu** diagrams with the same syntactic type. What is and is not possible with the concrete diagram modeled is determined by all **Eu** diagrams with the concrete diagram's syntactic type.

Roughly, having the same syntactic type means differing only with respect to the number of underlying array entries. Given a diagram  $\delta$  we can increase the number of array entries it contains while leaving the relative position of its objects within the array fixed. Since the resulting **Eu** diagram has the same objects with the same relative positions, it is taken to model the same concrete diagram  $\delta$  does.

The procedure of refinement addresses a worry one may have about the suitability of **Eu**'s diagrams. As discrete objects, they will fail in general to produce the intersection points that appear in Euclid's diagrams. For instance, in the diagram



we can join the points above and below the line to obtain the diagram

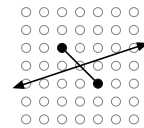


Given what this diagram is intended to represent, we ought to be able to produce an intersection point between the segment and the line. But the underlying array of the diagram is too coarse. An array entry does not exist where a point ought to be.

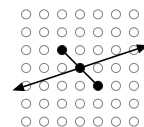
This can always be dealt with by refining an **Eu** diagram into one of the same syntactic type. The equation that characterizes a line (and the circumference of a circle) is linear, expressed in terms of the coordinates of the array entries. Since the arrays are discrete, the coefficients of the equation are always integers. Thus, the solution for two equations characterizing geometric elements of a diagram will always be rational.

This means that if two geometric elements ought to intersect but don't in a diagram, we can always find a diagram of the same syntactic type where they do. It will just be the original diagram with a more refined underlying array. In particular, if the original diagram has dimension  $n$  and the solution between the two equations is a rational with an  $m$  in its denominator, the new diagram will have dimension  $m(n - 1) + 1$ .

For the diagram above, then, adding the desired intersection point is a two-step process. First, the diagram is refined to an equivalent diagram of dimension 7.



Then the intersection point is added.



Another natural worry has to do with the circles of diagrams. The circles that appear in Euclid's diagrams actually appear circular. The circles of diagrams, however, are rectilinear. If Euclid exploits the circularity of his circles in his proofs, then the diagrams of **Eu** would fail to capture this aspect of Euclid's mathematics. Eu-

clid, however, never does this. All he seems to assume about circles is that they have an interior. Thus, with respect to the project of formalizing Euclid's proofs, **Eu** circles suffice. If on the other hand the circular appearance of Euclid's circles were deemed for other reasons to be important, the **Eu** syntax could be modified accordingly. **Eu** circles could be defined as regular polygons with, say, at least 1000 sides.

### Task 2: Semantic Equivalence

The relation *syntactic equivalence*, discussed above, serves to abstract from the features of **Eu** diagrams irrelevant to them as representations of concrete symbols. The relation of *semantic equivalence* serves to abstract from those features irrelevant to them as proof symbols. Two **Eu** diagrams carry the same information in a proof if and only if they satisfy the relation of semantic equivalence. It is an equivalence relation analogous to that of two **FG** having the same corresponding graph structure.

The relation is based ultimately on certain positional relations that can hold between pairs of objects in a diagram. For any two kinds of objects of an **Eu** diagram, we can stipulate a disjunctive range of mutually exclusive qualitative relations. For instance, with point  $p$  and circle  $c$  the range of relations stipulated are

$p$  lies inside  $c$ ,  $p$  lies on  $c$ ,  $p$  lies outside  $c$

Similarly, the relation range of a point and a linear element contains all the possible positions a point can have to a linear element. For a point  $p$  and a line  $l$  these are

$p$  on one side of  $l$ ,  $p$  on the other side of  $l$ ,  
 $p$  lies on  $l$

where sidedness is determined by labels on the end-arrows of  $l$  fixing the line's orientation. When the linear element  $l$  is a ray (or segment), the relation range contains the additional possibility (or possibilities) of  $p$  lying on an extension of  $l$ . The relation range for a line and circle lists the possible relations of tangency, intersection or nonintersection between them, and the relation range for two circles lists the possible relations of containment, tangency or intersection between them. Finally, the relation range between two linear elements is defined in terms of the positions of their endpoints and/or end arrows to one another.

The rough idea is that two diagrams are semantically equivalent if pairs of corresponding objects realize the same qualitative relation in the relation range stipulated for object types of the pair. More precisely,

semantic equivalence is a relation between *labeled Eu* diagrams. A labeling of an **Eu** diagram assigns variables to the points, circles and end arrows of the diagram. If the same variables label the same object types in two diagrams, then the labeling induces a one-to-one correspondence between the objects of the two diagrams. This is a precondition of semantic equivalence. The two diagrams are then semantically equivalent if corresponding pairs of objects realize the same relation in the appropriate relation range. For an example of two semantically equivalent **Eu** diagrams, see Fig. 23.2.

### Task 3: Single Diagram Proofs

The generality problem arises in **Eu** similarly to the way it does in **FG**. The syntactic operations that parallel the basic Euclidean operations do not in general preserve semantic equivalence. That is, applying the same syntactic operation to two semantically equivalent **Eu** diagrams does not in general result in two semantically equivalent **Eu** diagrams. In contrast to **FG**, however, **Eu** does not demand that every case be listed in the course of a geometric construction. A guiding ideal behind the design of **Eu** derivations is what could be called the *one proof-one diagram* conception of geometric proof. (The name is taken from [23.5]. In the second footnote of the paper *Manders* comments "Euclid, by and large, lives by *one proof, one diagram*" [23.5, p. 85]). According to it, a single diagram ought to be enough to establish a geometric proposition. Not all **Eu** derivations can be understood to correspond to proofs that rely on a single diagram. The formal system provides a framework, however, in which many of Euclid's proofs can be understood as such.

Within the framework, the task of establishing a geometric proposition with a single diagram divides into two subtasks. The first is the task of *producing* a geometric diagram as a concrete graphic object satisfying certain formal conditions. The second is the task of *reasoning* with the object produced. The reasoning consists, specifically, of verifying that certain qualitative positional relations exhibited by the diagram – such as the position of a segment within an angle – are representative of all the configurations within the scope of the proposition.

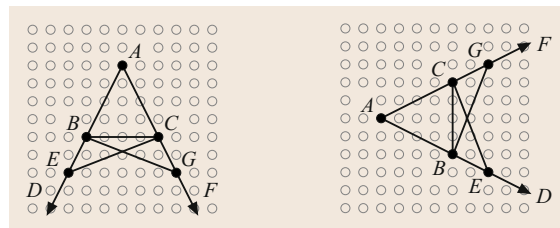


Fig. 23.2 Two equivalent **Eu** diagrams

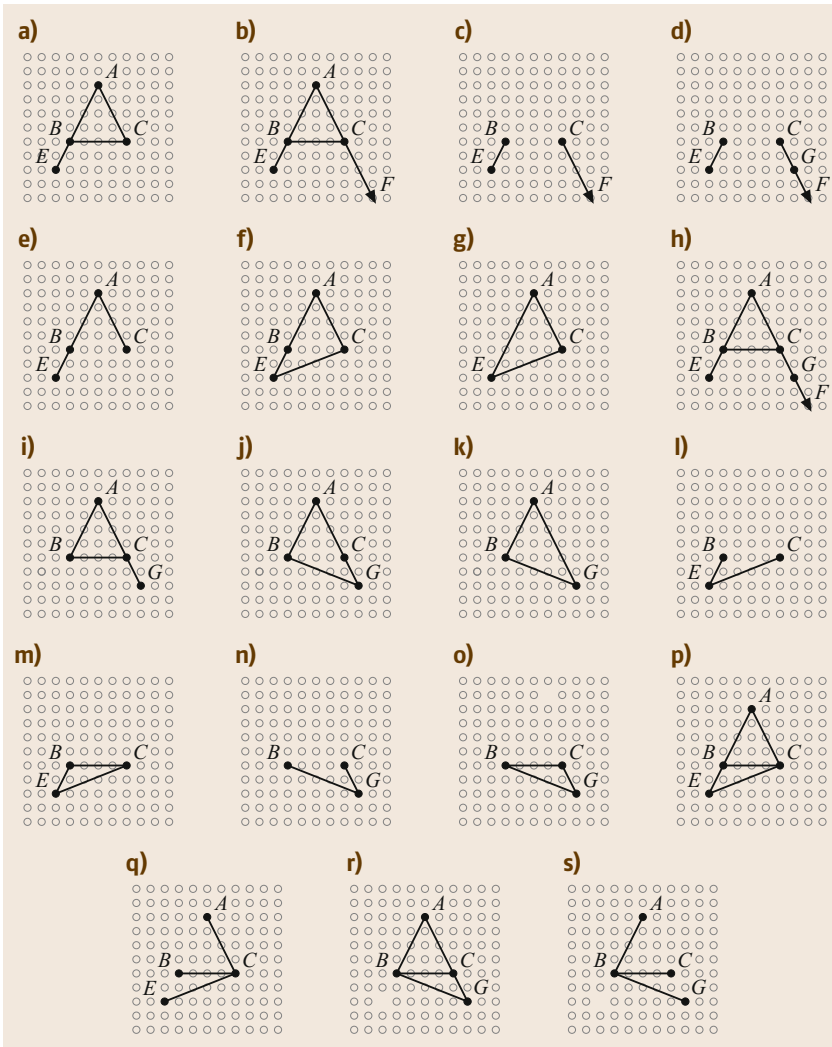


Fig. 23.3 **Eu** demonstration sequence for proposition I, 5

An **Eu** derivation thus splits into two stages, a *construction* stage and a *demonstration* stage. The construction stage is intended to correspond to the production of a geometric diagram as a concrete object, while the demonstration stage is intended to correspond, in part, to the reasoning carried out with the concrete object. (The demonstration stage also serves to record reasoning carried out with sentences representing relations between geometric magnitudes.) Both the construction and the demonstration stage in an **Eu** derivation can (and in most cases of interest do) contain many distinct **Eu** diagrams, even in the canonical case of a derivation that is intended to model a single diagram proof. And so, how the distinct, abstract **Eu** diagrams of such a derivation are to be understood in relation to the concrete diagram of single diagram proof requires some explanation. The general idea is

as follows: the different **Eu** diagrams of the construction correspond to different stages in the construction of a single concrete diagram  $\mathcal{D}$ ; each of the different **Eu** diagrams of the demonstration correspond to the product of an act of attention directed at  $\mathcal{D}$ .

Consider for instance the **Eu** diagrams in Figs. 23.3 and 23.4. These are the diagrams that appear in an **Eu** derivation modeling a single diagram proof of proposition 5, book I of the *Elements*. The final **Eu** diagram of the construction sequence (Fig. 23.4) corresponds to the proof's concrete diagram  $\mathcal{D}$ . The **Eu** diagrams preceding it correspond to stages in the construction of  $\mathcal{D}$ . All the **Eu** diagrams of the demonstration sequence (Fig. 23.3) are subdiagrams of the final diagram of the construction sequence. They correspond to acts of attention whereby certain relationships present in  $\mathcal{D}$  are verified to hold in general. The sequence thus repre-

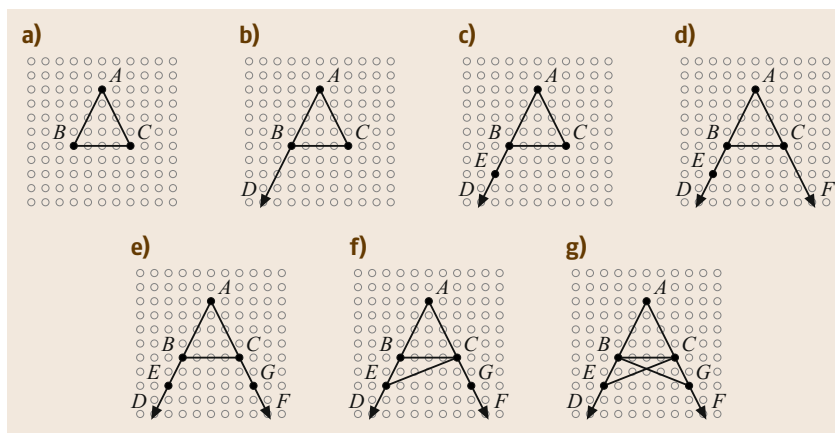


Fig. 23.4 Eu construction sequence for proposition I, 5

sents a reasoning process verifying that the position of  $CB$  within angle  $\angle ACE$  and the position of  $BC$  within  $\angle ABG$  hold in general.

As it is only the demonstration sequence that is intended to correspond to a reasoning process, it is only

the demonstration stage that is governed by the relation of semantic equivalence in **Eu**. Specifically, the rules that license the addition of a diagram to the demonstration sequence given previous diagrams in the sequence must preserve semantic equivalence.

### 23.3 Formal Geometric Diagrams as Models

In itself, a formalization is just a system of rules for producing formal objects. They must be interpreted in order to have any epistemological significance. For instance Peano arithmetic, understood purely as a formal system, simply prescribes a method for producing formal symbols from a fixed formal vocabulary. It is only after one systematically relates the formal structure of these sequences to logical and arithmetical concepts that we arrive at a foundational analysis of arithmetic. The sequences then come to be seen as sentences asserting facts about natural numbers.

Similarly, when understood purely as formalizations, **FG** and **Eu** simply provide rules for producing certain kinds of formal structures. These structures are, to be sure, different in kind from an axiomatization in first-order logic like Peano arithmetic. They are nevertheless formal and have no epistemological significance with respect to elementary geometry without some kind of interpretation.

One option for such an interpretation is furnished by a first-order axiomatization of elementary geometry (e.g., the axiomatization given in [23.7]). The formal structures of **FG** and **Eu** can be systematically related to the formal sentences of the axiomatization. Accordingly, the formal structures of **FG** and **Eu** would then amount to an idiosyncratic, and very indirect, notation for sentences asserting facts about the Euclidean plane. In this concluding section I explore the prospect of an

alternative interpretation of the proof systems where each formalizes a species of model-based reasoning.

Elementary geometry is a mathematical subject, and proofs in mathematical subjects are deductive. And so, an interpretation of **FG** and **Eu** in line with the model-based reasoning perspective ought to proceed from a conception of deduction in line with the model-based reasoning perspective. *Lorenzo Magnani* articulates such a conception in [23.8]. Interestingly, he does so by considering proofs in elementary geometry.

Magnani specifically refers to *Hintikka* and *Remes*' investigations into the logic of geometric proof in [23.9]. (For *Hintikka*'s most recent discussion of geometric proof see [23.10].) One of *Hintikka*'s central points is that the logic of proof in elementary geometry is best understood in terms of the method of semantic tableau. Take the proof of proposition 5 from book 1 of the *Elements* as an example. The proposition asserts that for any collection of points and segments forming an isosceles triangle, the angles opposite the equal sides of the triangle are equal. Thus the logical form of the proposition is

$$\forall x_1, x_2, \dots, x_n [\varphi_1(x_1, \dots, x_n) \rightarrow \varphi_2(x_1, \dots, x_n)],$$

where  $\varphi_1$  corresponds to the conditions defining an isosceles triangle, and  $\varphi_2$  to the condition that the angles opposite the equal sides of the triangle are equal.

Now, at the heart of the proof of a statement of the above form within the semantic tableau setting is the proof of a conditional

$$\varphi_1(a_1, \dots, a_n) \rightarrow \varphi_2(a_1, \dots, a_n)$$

in which  $a_1, \dots, a_n$  are understood as arbitrary, and the formulas  $\varphi_1(a_1, \dots, a_n)$  and  $\varphi_2(a_1, \dots, a_n)$  come to be linked via logical operations, axioms and previously proven theorems. With respect to proposition 5, the  $a_1, \dots, a_n$  represent an *instantiation* of the theorem. Thus, according to the framework of semantic tableau, at the heart of the reasoning establishing proposition 5 is the consideration of representations understood to instantiate the theorem. The  $a_1, \dots, a_n$  serve, in other words, *to model* the type of configuration the proposition concerns, and it is by interacting with this modeling that the proposition is established.

Hintikka's work leads us thus to the following abstract characterization of deduction from a model-based reasoning perspective. Deductive inference concerns a complex of interrelated objects. The complex is assumed to satisfy certain conditions  $\varphi_1$ , and is inferred to satisfy further conditions  $\varphi_2$ . The inference proceeds via consideration of a representation modeling the complex. An important aspect of deduction understood in this way, emphasized by both Hintikka and Magnani but passed over in the above discussion of proposition 5, is that the representation modeling the complex can be enriched. One need not restrict oneself to the objects the deduction explicitly concerns in constructing a model for it. One may add to the representation additional objects to facilitate the reasoning. With respect to a proof of elementary geometry, this simply amounts to performing a construction on the initial configuration of the proof.

Call this the model-based reasoning, or MBR, conception of deduction. The conception differs from the standardly accepted one in that what is front and center are representations of objects and their relations, rather than sentences asserting relations between objects. It is such representations that drive the deductive inference that any collection of objects satisfying conditions  $\varphi_1$  also satisfy the conditions  $\varphi_2$ . To perform such an inference, one must recognize that the conditions  $\varphi_1$  impose constraints upon a collection of objects with respect to the relations in  $\varphi_2$ . This act is accomplished by representing an instantiation of  $\varphi_1$  and  $\varphi_2$ , augmented perhaps with additional objects. The representation serves to reveal the constraints the conditions  $\varphi_1$  impose upon objects with respect to the relations in  $\varphi_2$  directly.

How do representations of instantiations do this? It is not immediately clear from the general logical perspective Hintikka assumes. From this perspective, the only way to represent an instantiation is sententially – i. e., via predicates and singular terms. If the singular terms are understood simply to denote objects in the broadest logical sense, a listing of predicates that the singular terms satisfy reveals on its own only trivial constraints. Suppose we have a three-place predicate  $B$  and singular terms  $a_1, a_2, a_3$  and  $a_4$ . Then

$$B(a_1a_2a_3) \quad B(a_2a_3a_4)$$

qualifies as a sentential representation of an instantiation. But the only constraint on  $a_1, a_2, a_3$  and  $a_4$  that the representation reveals, if our conception of the objects is the broadly logical one, is that the triples  $\langle a_1, a_2, a_3 \rangle$  and  $\langle a_2, a_3, a_4 \rangle$  must satisfy  $B$ . Aside from the two sentential expressions that negate  $B(a_1a_2a_3)$  and  $B(a_2a_3a_4)$  – i. e.,  $\neg B(a_1a_2a_3)$  and  $\neg B(a_2a_3a_4)$  – we are free to add to the representation any sentential expression with the singular terms  $a_1, a_2, a_3$  and  $a_4$ .

This observation shows, at the very least, that if the MBR conception of deduction is to be of any interest, the operative conception of object in a deductive inference has to be richer than the austere one furnished by logic. There has to be, in other words, background knowledge with respect to the objects and their combination in complexes – e.g., what relations can and cannot obtain among the objects of a complex, what additions can be made to a given complex, and so on. If this is accepted, the question then becomes: how does this background knowledge exert itself when a representation of an instantiation is considered in the course of a deduction?

Here is where proven theorems and/or axioms come into play in a semantic tableau formalization. At the initial stage, before any theorems are proven, all background knowledge about the objects under consideration is encoded in unproven axioms. We look to these for what can and cannot be done with representations of instantiations. These then allow us to use such representations to deduce nontrivial theorems via semantic tableau, which then can be used in future deductions. For an example of how this works, consider the three-place predicate  $B$  again, and suppose that it denotes the relation of betweenness for points on a geometric line. Suppose further we are at a point where the basic fact about betweenness given by

$$\forall x, y, z, w [(B(xyz) \wedge B(yzw)) \rightarrow B(xyw)]$$



is known, either as an axiom or a theorem. Then in any instantiation

$$B(a_1a_2a_3) \quad B(a_2a_3a_4)$$

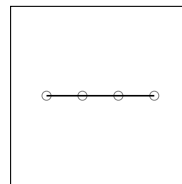
within a semantic tableau, we can add

$$(B(a_1a_2a_3) \wedge B(a_2a_3a_4)) \rightarrow B(a_1a_2a_4)$$

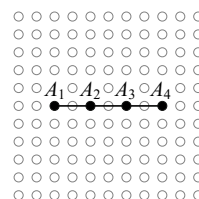
Application of the rules of the semantic tableau to these formal representations eventually shows that the conditions  $B(a_1a_2a_3)$  and  $B(a_2a_3a_4)$  impose a constraint on betweenness with respect to  $a_1$ ,  $a_2$  and  $a_4$  – i. e., the relation  $B(a_1a_2a_4)$  in fact holds.

The virtue of semantic tableau is that it provides a topic-neutral formal framework for investigating the MBR conception of deduction. It does not follow from this, however, that it is the *only* formal framework that functions in accordance with the MBR conception. It does not follow, that is, if the conception's central principle is that that deduction is accomplished via representations of instantiations. A formal framework could very well be designed for representing instantiations of object complexes of a particular kind. Within such a framework, the formal structures for representing instantiations would do (at least some of) the work done by the axioms in the semantic tableau framework. Background knowledge of what is and is not possible within a complex would be embedded in the constraints on producing these formal structures.

This is what is in fact happening, I maintain, with the **FG** and **Eu** formalizations. The function of its diagrammatic syntax is not to convey conjunctions of geometric conditions with an idiosyncratic notation, but to present instantiations of geometric configurations for the sake of facilitating deductions. Consider again a configuration made up of points  $a_1, a_2, a_3$  and  $a_4$  on a geometric line, where  $a_2$  is between  $a_1$  and  $a_3$ , and  $a_3$  is between  $a_2$  and  $a_4$ . Within **FG**, such a configuration is represented as



while in **Eu** it is represented as:



With both, the constraint that  $B(a_1a_2a_3)$  and  $B(a_2a_3a_4)$  imposes on the relative positions of  $a_1, a_2$  and  $a_4$  is immediately evident. Generally, the diagrammatic syntax of **Eu** and **FG** embodies what must be laid down as axioms in an axiomatization. As *Miller* puts it with respect to his system [23.1, p. 40]

“[...] many of the facts that Hilbert adopts as his axioms of incidence and order [in [23.11]] are consequences of the diagrammatic machinery built into the definitions of **FG**.”

We thus have, I maintain, with **FG** and **Eu** two formal characterizations of deduction carried out according to the principles of model-based reasoning. With **FG**, the diagrammatic representation of instantiations serve to give the reasoner direct access to a range of geometrical possibilities. The diagrams of **Eu**, in contrast, allow the reasoner to consider all the components of a geometric configuration in one place and to focus on those components relevant to a proof. The precise formal pictures they provide can provide a basis for further investigations into the relation between instantiation and deduction in mathematics.

## References

- |  |  |
|--|--|
| <p>23.1 N. Miller: <i>Euclid and His Twentieth Century Rivals: Diagrams in the Logic of Euclidean geometry</i> (CSLI, Stanford 2007)</p> <p>23.2 J. Mumma: Proofs, pictures, and Euclid, <i>Synthese</i> <b>175</b>, 255–287 (2010)</p> <p>23.3 J. Avigad, E. Dean, J. Mumma: A formal system for Euclid's <i>Elements</i>, <i>Rev. Symb. Log.</i> <b>2</b>, 700–768 (2009)</p> <p>23.4 Y. Hamani, J. Mumma: Prolegomena to a cognitive investigation of Euclidean diagrammatic reasoning, <i>J. Log. Lang. Inf.</i> <b>22</b>, 421–448 (2014)</p> | <p>23.5 K. Manders: The Euclidean diagram. In: <i>Philosophy of Mathematical Practice</i>, ed. by P. Mancosu (Clarendon Press, Oxford, 2008) pp. 112–183</p> <p>23.6 Euclid: <i>The Thirteen Books of the Elements</i>, Vol. I–III, 2nd edn. (Dover, New York 1956), transl. by T.L. Heath</p> <p>23.7 A. Tarski: What is elementary geometry? In: <i>The Axiomatic Method, with Special Reference to Geometry and Physics</i>, ed. by L. Henkin, P. Suppes, A. Tarski (North Holland, Amsterdam 1959) pp. 16–29</p> |
|--|--|

- 23.8 L. Magnani: Logic and abduction: Cognitive externalizations in demonstrative environments, *Theoria* **60**, 275–284 (2007)
- 23.9 J. Hintikka, U. Remes: *The Method of Analysis: Its Geometrical Origin and General Significance* (Reidel, Dordrecht 1974)
- 23.10 J. Hintikka: Method of analysis: A paradigm of mathematical reasoning?, *Hist. Philos. Log.* **33**, 49–67 (2012)
- 23.11 D. Hilbert: *Foundations of Geometry* (Open Court, La Salle 1971)

## 24. Model-Based Reasoning in Mathematical Practice

Joachim Frans, Isar Goyvaerts, Bart Van Kerkhove

The nature of mathematical reasoning has been the scope of many discussions in philosophy of mathematics. This chapter addresses how mathematicians engage in specific modeling practices. We show, by making only minor alterations to accounts of scientific modeling, that these are also suitable for analyzing mathematical reasoning. In order to defend such a claim, we take a closer look at three specific cases from diverse mathematical subdisciplines, namely Euclidean geometry, approximation theory, and category theory. These examples also display various levels of abstraction, which makes it possible to show that the use of models occurs at different points in mathematical reasoning. Next, we reflect on how certain steps in our model-based approach could be achieved, connecting it with other philosophical reflections on the nature of mathematical reasoning. In the final part, we discuss a number of specific purposes for which mathematical models can be used in this context. The goal of this chapter is, accordingly, to show that embracing modeling processes

24.1	<b>Preliminaries</b> .....	537
24.2	<b>Model-Based Reasoning: Examples</b> ...	538
24.2.1	First Example: From Euclidean Geometry .....	538
24.2.2	Second Example: From Approximation Theory .....	539
24.2.3	Third Example: From Category Theory .....	540
24.3	<b>The Power of Heuristics and Plausible Reasoning</b> .....	540
24.4	<b>Mathematical Fruits of Model-Based Reasoning</b> .....	542
24.5	<b>Conclusion</b> .....	546
24.A	<b>Appendix</b> .....	546
	<b>References</b> .....	548

as an important part of mathematical practice enables us to gain new insights in the nature of mathematical reasoning.

In this chapter, we explore the significance of model-based reasoning for mathematical research. In Sect. 24.1, we start by outlining an account of the nature of scientific modeling, and how it could be applied to mathematics. This becomes more clear in Sect. 24.2, where this account will be briefly applied to three specific examples coming from different mathematical subdisciplines, and also exhibiting a different level of abstraction, namely Euclidean geometry, approximation theory, and category theory re-

spectively. Section 24.3 reflects on how specific transitional steps in the model-based argument schemes presented are to be achieved, and more particularly on what are commonly called types of plausible reasoning that thus arguably play an important role in mathematical discovery. In Sect. 24.4, some of the alleged epistemic merits or purposes of model-based reasoning as presented in the context of mathematical practice are considered. Section 24.5 concludes the chapter.

### 24.1 Preliminaries

*Aris* [24.1] has proposed the following definition (as quoted in *Davis* and *Hersh* [24.2, p. 78]):

“A mathematical model is any complete and consistent set of mathematical equations which are designed to correspond to some other entity, its pro-

totype. The prototype may be a physical, biological, social, psychological or conceptual entity, perhaps even another mathematical model.”

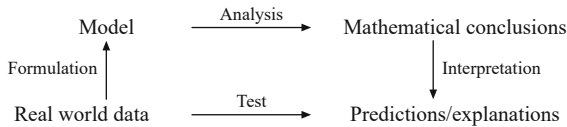
*Davis* and *Hersh* [24.2, p. 78–79] have commented on this:

“One might substitute the word *structure* for *equations* [in the above quote], for one does not always work with a numerical model. Some of the purposes for which models are constructed are:

1. To obtain answers about what will happen in the physical world.
2. To influence further experimentation or observation.
3. To foster conceptual progress and understanding.
4. To assist the axiomatization of the physical situation.
5. To foster mathematics and the art of making mathematical models.

The realization that physical theories may change or may be modified [...], that there may be competing theories, that the available mathematics may be inadequate to deal with a theory in the fullest sense, all this has led to a pragmatic acceptance of a model as a *sometime thing*, a convenient approximation to a state of affairs rather than an expression of eternal truth.”

*De Vries* [24.3], for another, has depicted the process of mathematical modeling as follows:



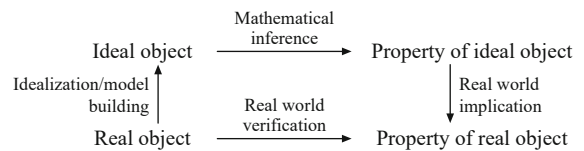
We from our part shall argue that, *making abstraction* from mathematical reality, in the sense explained below, aspects of mathematical practice can be seen as a sort of modeling, in case the prototypes referred to in

the Aris quote already are certain mathematical structures.

First of all, let us specify what can be meant by *making abstraction* from mathematical reality in our present context. We intimately follow the treatment of this subject by *Davis* and *Hersh* [24.2, pp. 126–36]. The term *abstraction* is used in different but related senses in mathematics; *Davis* and *Hersh* distinguish *abstraction as idealization* and *abstraction as extraction*. The idealizations in this context proceed from the world of spatial experience to the mathematical world. Aristotle is referred to in this respect, pointing out that [24.2, p. 127]:

“the mathematician strips away everything that is sensible, for example, weight, hardness, heat, and leaves only quantity and spatial continuity.”

It is then said that the development of contemporary mathematical models exhibits updated versions of this (Aristotle’s) process. The visualization of this type of abstraction process provided by *Davis* and *Hersh* [24.2, pp. 129], has quite some similarities with the sketch of the process of mathematical modeling of *de Vries* quoted above (To serve comparison though, we have rotated *Davis* and *Hersh*’s diagram 90° counter clockwise.):



Therefore, as already mentioned earlier, we dare claim that model-based reasoning can and does occur within mathematical practice.

## 24.2 Model-Based Reasoning: Examples

In this section, we briefly elaborate three examples; a very *basic problem-solving* one, one at an intermediate level of abstraction, and finally an utmost conceptual one, calling on some notions from category theory. In Sect. 24.4, we shall argue that this model-based way of reasoning exhibits features that are similar to the five purposes models are designed for in other sciences (as described earlier by *Davis* and *Hersh*).

### 24.2.1 First Example: From Euclidean Geometry

Imagine one is given a right-angled triangle  $T$ , whose legs  $a$  and  $b$  measure 4 and 3 cm respectively. A second

right-angled triangle  $T'$  is given as well, its legs  $a'$  and  $b'$  measuring 3 and 2 cm, respectively. Now suppose we care about knowing the length of the hypotenuses of  $T$  and  $T'$ , denoted by  $c$  and  $c'$ . Clearly, a first and most evident technique consists in just measuring these lengths with a marked ruler. One finds that this is something like 5 cm for  $c$  and approximately 3.6 cm for  $c'$ .

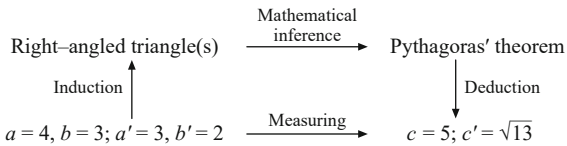
One satisfied with these results can carry on without asking the question of *why*  $c$  and  $c'$  measure exactly as they do. If however one does not suffice with these mere numerical results, he or she might notice at some point, after further inquiry, that indeed something deeper is going on. Note that we largely ignore at this point *how* this is realized, as we also disregard for the sake of

the present discussion how the following observations are (to be) made. We indeed suppose that all steps in the cycle described *can* actually be carried out at some point. In Sect. 24.3, we shall explore the alleged importance of informal reasoning (including analogies and visualizations) in the process of inductive and deductive mathematical inference (see however also the other contributions in the current part of this volume.).

Thus, in whichever way, inductive reasoning, by reading and *taking in* the given information described above, makes one realize that  $a$  and  $b$ ,  $a'$  and  $b'$  are legs of two right-angled triangles. Mathematical inference teaches us that for such triangles, the universal regularity called Pythagoras' theorem holds, which assures us that the square of the length of the hypotenuse equals the sum of the squares of the lengths of the legs. This means that by deduction, plugging in the values of  $a$  and  $b$  in this result, we find that

$$c^2 = a^2 + b^2 ;$$

hence  $c = 5$ , as lengths are positive real numbers. Similarly, plugging in  $a'$  and  $b'$  in Pythagoras' theorem, we find that  $c' = \sqrt{13}$  (observing that 3.6 is relatively close to this). Schematically depicted, we arrive at the following *more or less* commutative diagram (by commutative diagram we mean that the diagram has the property that all directed paths with the same start and endpoint lead to the same result by following the arrows.):



Some brief remarks with respect to this diagram are appropriate. First of all, we added *more or less* when talking about the above diagram's being commutative. This is because we indeed are disregarding the issue of actually establishing the link between *measuring* the hypotenuses on the one hand, and *theoretically deducing* their *real* lengths from Pythagoras' theorem on the other. Given the discussion of ever *perfectly* measuring length, 5 as well as  $\sqrt{13}$  in this case, one might indeed wonder if it could *at all* be feasible to render a *down-to-earth* diagram as the one given (involving empirical verification) commutative. We do hold that for both pedagogical and conceptual reasons, arriving at *the* measures of  $c$  and  $c'$  by following the induction–mathematical inference–deduction route of the diagram, is way more satisfactory than just measuring their *approximate* values.

### 24.2.2 Second Example: From Approximation Theory

Our second case comes from the mathematical field called approximation theory, one of the central goals of which it is to “represent an arbitrary function in terms of other functions which are nicer or simpler or both” (*Hrushikesh and Devidas* [24.4, p. 1]). This area of research is thus mostly concerned with how functions can be better approximated with easier functions, and with how the errors occurring in this process can be characterized. The point is that, in many cases, it is difficult or even impossible to extract exact analytical information from an arbitrary function  $f$ . In such cases, it is nevertheless useful and therefore important to be able to approximate  $f$  with a simpler function. Intuitively speaking, in cases like these, mathematicians sometimes look for a function  $g$ , such that the relevant calculation can be performed on the function  $g$  while  $g$  is *close enough to*  $f$  in the sense that the outcome of the calculation performed on  $g$  gives us meaningful information about  $f$ .

Let us give a concrete and simple example to clarify what the role of approximation theory can be. The example is inspired by *Christensen and Christensen* [24.5]. Assume that we want to compute the following integral

$$\int_0^1 e^{-\frac{x^2}{2}} dx .$$

Now, a primitive function of the function

$$f(x) = e^{-\frac{x^2}{2}}$$

cannot be expressed as a combination (sum, composition, multiplication, quotient) of elementary functions (polynomials, trigonometric, logarithmic functions, and their inverses). So in order to obtain numerical values of the above integral, other means are called for. This is where approximation theory enters the picture. One of the goals is to search a function  $g$  for which (i)  $\int_0^1 g(x)dx$  can be calculated, and (ii)  $g(x)$  is *close to*

$$e^{-\frac{x^2}{2}}$$

for  $x \in [0, 1]$ , in the sense that we can keep under control how much  $\int_0^1 g(x)dx$  deviates from

$$\int_0^1 e^{-\frac{x^2}{2}} dx .$$

A possible way of doing so, is to find a positive integrable function  $g$  for which, for some  $\epsilon > 0$ ,

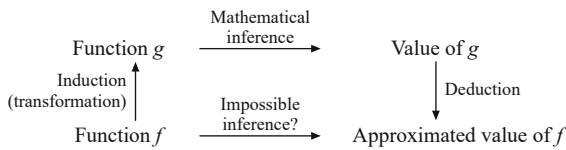
$$-\epsilon \leq e^{-\frac{x^2}{2}} - g(x) \leq \epsilon, \quad \forall x \in [0, 1],$$

$$-\epsilon + g(x) \leq e^{-\frac{x^2}{2}} \leq \epsilon + g(x), \quad \forall x \in [0, 1].$$

Consequently, one can obtain that

$$-\epsilon + \int_0^1 g(x) dx \leq \int_0^1 e^{-\frac{x^2}{2}} dx \leq \epsilon + \int_0^1 g(x) dx.$$

As a result,  $\int_0^1 g(x) dx$  gives us an approximate value for the desired integral. Now the question remains, of course, of how to determine or *choose*  $g$ . Approximation theory is the field where precisely this type of questions are further discussed and developed. This quest has led to several powerful and useful mathematical techniques, such as the theory of Chebyshev polynomials or Fourier analysis. From the point of view of concrete numerical computations, such a transformation is often preferable, because it gives a simple way of obtaining information about the function which would otherwise be difficult to collect, or could not even be traced at all. Following our take on model-based reasoning in mathematics, we can present this practice in the following, by now familiar scheme:



### 24.2.3 Third Example: From Category Theory

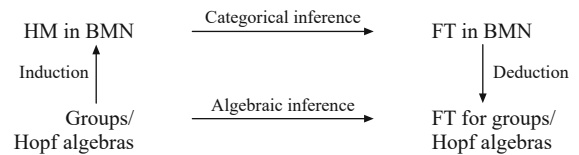
Finally, we would also like to provide an illustration from *higher level* mathematical practice, sketching some of the features of research in category theory. In order to appreciate the results and techniques from this

field, it might be helpful to recall some of the mathematical notions involved. Therefore, without going into full detail, we briefly develop some of those ideas, or at least an interpretation of them which is relevant to our setting, in the appendix.

Groups are algebraic objects intimately related to the notion of symmetry. Hopf algebras (over a field  $k$ ) are – slightly more complicated – algebraic objects, group algebras being an important class of examples of such structure. As explained in the appendix, groups can be seen as Hopf monoids (HM) in the braided monoidal category of sets, denoted  $\widehat{\text{Sets}}$ . Similarly, Hopf algebras are simply Hopf monoids in  $\widehat{\text{Vect}}_k$ .

As remarked in *Verduyn* [24.6] (§5.1), most of the theory of (classical) Hopf algebras can be lifted to the setting of Hopf monoids in arbitrary braided monoidal categories (BMN), sometimes under additional assumptions (such as the existence and preservation of certain (co)limits); one such a result being the so-called Fundamental theorem (FT) for Hopf modules for instance (cf. *Takeuchi* [24.7]). Plugging in your favorite braided monoidal category (satisfying the necessary assumptions) will then give you a version of this theorem for Hopf monoids in that particular category.

This result can also be obtained through *direct manipulation* of particular objects in the chosen category as well. For Hopf algebras for instance, FT can be obtained directly by applying techniques from linear algebra (i. e., manipulation of  $k$ -vector spaces), that is, in case you choose to plug in  $\widehat{\text{Vect}}_k$  (actually, both  $\widehat{\text{Sets}}$  and  $\widehat{\text{Vect}}_k$  satisfy the necessary conditions for FT to hold). It should be noted that in many cases, the categorical proof is inspired by a *classical* (often linear algebraic) proof of the statement for some particular algebraic object. Again, this whole practice can schematically be recapitulated in the following way:



## 24.3 The Power of Heuristics and Plausible Reasoning

Before considering the epistemic merits of the reasoning model just presented, we want to briefly pause to reflect on how the appropriate steps or *transitions* in the above model-based diagrams are to be achieved, particularly in the induction and mathematical inference phases. Indeed, in the previous section, we have

remained utterly silent on how mathematicians safely arrive at the nodes in the upper half of the diagrams, that is, what particular intellectual processes are required in order to reach such a higher abstraction level. While it would take an in-depth study of these aspects as applied to the specific cases presented, in order to fully substan-

tiate why this is in fact possible, we nevertheless want to briefly explore this *context of discovery*-dimension here in somewhat more general terms.

An interesting starting point can be found in the work of Pólya. In *Induction and analogy in mathematics*, he states that all our knowledge outside mathematics and demonstrative logic consists of conjectures. It is certainly the case that some of these conjectures, such as those expressed in general laws of physical science, are highly reliable and commonly accepted. Other conjectures are neither reliable or respectable.

The support for conjectures is obtained by plausible reasoning, while our mathematical knowledge is secured by demonstrative reasoning. Mathematical proofs are part of demonstrative reasoning, while the inductive, circumstantial, or statistical evidence for a scientist belongs to plausible reasoning. One of the main differences between these two kinds of reasoning is that demonstrative reasoning leads to safe and final knowledge that is beyond revision. Plausible reasoning, on the other hand, leads to provisional and controversial knowledge.

What Pólya argues for is that, while mathematics is regarded as a demonstrative science, plausible reasoning also plays an important role in mathematics. He clarifies this by referring to finished mathematics and mathematics in the making (by way of a very nice metaphor, Reuben Hersh later called this the distinction between the front and the back in mathematics Hersh [24.8]). Finished mathematics appears to be purely demonstrative, consisting of only proofs. Yet mathematics in the making is similar to other human knowledge in the making. The following passage clarifies this (Pólya [24.9, p. 100]):

“You have to guess a mathematical theorem before you prove it, you have to guess the idea of the proof before you carry through the details. You have to combine observations and follow analogies, you have to try and try again. The results of the mathematician’s creative work is demonstrative reasoning, a proof; but the proof is discovered by plausible reasoning, by guessing.”

From this observation, Pólya concludes that students of mathematics should learn both kinds of reasoning.

A very similar story is told by Lakatos in his landmark study *Proofs and refutations* (Lakatos [24.10]), where he identifies three rough stages in mathematical reasoning. First, mathematicians use induction (in the sense of generalization on the basis of particular instances) to discover conjectures worth trying to prove. Then they develop and criticize highly *informal* proofs

of these conjectures. Only in a last phase of mathematical labor, they formalize these informal theories, establishing the deduction of the (by then) theorems by means of formal transformations on an axiomatic basis. Lakatos famously illustrates this practice with a (*rationaly*) reconstructed history of how the proof of Euler’s polyhedron formula  $V - E + F = 2$  came about,  $V$  being the number of vertices,  $E$  the number of edges, and  $F$  the number of faces of any given polyhedron.

Let us take a somewhat closer look at some of Lakatos’ terminology, which he used to outline various methods by which mathematical discovery (and subsequent justification) can occur. These methods describe ways in which mathematical concepts, conjectures, and proofs gradually evolve through interaction between mathematicians. Central to these practices, so Lakatos claims, are counterexamples, and he discusses several ways in which mathematicians or students can react to these: by surrender, monster-barring, exception-barring, monster-adjusting, or lemma-incorporation. In what follows, we briefly sketch the essence.

First of all, surrender amounts to abandoning a conjecture in the light of a counterexample. This is however not done lightly, so more frequent are other reactions. Monster-barring, for instance, which consists in ignoring or excluding an alleged counterexample. This implies that one has to show why it is not within the relevant concept definition. One can claim, for example, that a hollow cube, that is a cube with a cube-shaped hole in it, is not a counterexample to Euler’s conjecture, by arguing that the hollow cube is not in fact a polyhedron, and thus cannot threaten the conjecture. This means that the concept polyhedron is under discussion, soliciting a further explication of its definition.

As for exception-barring, Lakatos argues that exceptions, rather than simply being problematic for cases and thus dismissed as monsters, can lead to new knowledge. Two ways to deal with exceptions are discussed. One is piecemeal exclusion, for example, by excluding one type of polyhedron from the conjecture in order to set aside a whole class of counterexamples. The other is strategic withdrawal, which does not directly rely on counterexamples. Instead, positive examples of a conjecture are used in order to generalize to a class of objects, and consequently limit the domain of the conjecture to this class.

Yet another way of responding to counterexamples is termed monster adjusting. It is intended to meet the possible criticism that both monster-barring and exception-barring are not taking counterexamples serious enough. Here, the mathematicians reinterpret the counterexamples so that they indeed fall within the scope of the original formulation of the conjecture, and thus show how the anomalies are in fact unproblematic.

The method of lemma-incorporation differs from all the above methods in that it uses properties of the proof itself. The idea is to examine the proof in order to determine exactly which lemma has been refuted by the counterexample. The guilty lemma is then added as a condition to the conjecture, and is consequently no longer refuted by the counterexample.

The most important or general method for Lakatos, as the title of his work suggests, is that of proofs and refutations, which in a certain sense amounts to a dialectic form of the method of lemma incorporation. Lemma incorporation enables one to make a distinction between global and local counterexamples. Global counterexamples refute the main conjecture, while local counterexamples are counterexamples to specific proof steps only. If a counterexample is both global and local, and thus constitutes a problem for both argument and conclusion, one should modify the conjecture by incorporating the problematic step as a condition. If the counterexample is not global but just local, which means the conclusion can still be correct while one of the reasons for believing it is flawed, one should leave the conjecture unchanged and modify the proof. Finally, if the counterexample is global but not local, which

means that there is a problem with the conclusion without obvious problems for any of the reasoning steps, then one should look for a possible hidden assumption in one of the proof steps and modify the proof by making this assumption explicit.

Summarized, Lakatos' proposed method consists in exploiting proof steps to suggest counterexamples. By looking for objects violating an argumentative step, one can identify possible such candidates. Whenever a counterexample is actually found, one needs to determine of its kind and accordingly modify the proof or conjecture. Note that other modes of model-based reasoning, such as several ones touched upon elsewhere in this collection (metaphorical, analogical, and/or visual reasoning) can very much be at play in these particular stages of mathematical inquiry. Indeed are these not your exemplary instances of heuristic or plausible reasoning in the context of discovery? We shall not further explore this issue ourselves, and after this interlude return to our central topic, in order to consider how and why the reasoning model introduced in Sect. 24.2 should work, that is what its original (if perhaps not essential) contributions to mathematical research might be.

## 24.4 Mathematical Fruits of Model-Based Reasoning

Let us recall the purposes of model-based reasoning as listed by *Davis* and *Hersh* [24.2, pp. 78–9] in Sect. 24.2, and consider their translation to a mathematical context:

- “(1) To obtain answers about what will happen in the physical world.
- (2) To influence further experimentation or observation.
- (3) To foster conceptual progress and understanding.
- (4) To assist the axiomatization of the physical situation.
- (5) To foster mathematics and the art of making mathematical models.”

As this entire collection may testify, mathematical models are most important and powerful tools for the empirical sciences in order to gain fresh insights about aspects of the physical world. These insights can come in different forms, for example, as a prediction about what will probably happen, an explanation of something that has already happened, or a deeper understanding of (part of) the physical situation under investigation. A satisfactory account of how indeed a model can help answer such questions about reality

is the mapping account (*Pincock* [24.11]), where one adopts the view that there is an appropriate mapping between the target system and the model. The knowledge obtained in the model can be translated into knowledge about the target system, because the model's various aspects correspond (at least to a large enough extent) to elements in the real world.

Now it would be strange to argue that all the purposes of models exhibited in empirical science have their exact counterparts in mathematical practices. Take prediction (*purpose 1*), which obviously plays a crucial role in the sciences. Mathematics, on the other hand, is not exactly your kind of discipline that has room for the same kind of prediction. Surely, mathematicians make guesses about the nature of mathematical objects (*facts*) or about how a certain proof will (or should) look like, but this does not warrant the drawing of too a strong connection between mathematical and scientific prediction. Be that as it may, the central purpose of using models does remain in place, namely the conversion of knowledge obtained in the model into knowledge about the target system. Also, the reason for making this possible is similar: it is due to the existence of a mapping relation between the two domains. This, we think, has been nicely illustrated by the various examples we have



developed above on the basis of the diagram we had first introduced to visualize the process of model-based reasoning in mathematical practice. Mathematicians do not simply use abstraction in order to find a result at that particular level. For after obtaining results there, they often translate them back to the original *target* system, and consequently, via an abstraction detour, obtain their answers about this more basic mathematical domain.

Indeed, in our example of Euclidean geometry, one theoretically predicts the (approximated) length of the triangles in the real world or one verifies the theoretical result in the real world. Category theory tells us a similar story. Suppose one has some algebraic structure  $A$  of which one can prove that it is a (Hopf) monoid in a certain (braided) monoidal category  $\underline{\mathcal{C}}$ . At the level of monoidal category theory, some general theorem  $T(-)$  that holds for all (Hopf) monoids in any (braided) monoidal category exists (sometimes assuming some small extra conditions on  $\underline{\mathcal{C}}$ ), such as the Fundamental theorem for instance. If you then *plug in* the category  $\underline{\mathcal{C}}$  you get a version of the theorem  $T(A)$  for the algebraic structure  $A$  you started from. Sometimes this theorem  $T(A)$  is known and has been shown to be true before. Then we can speak of *algebraic world verification* in some sense. This verification is not always a priori possible in every field of mathematical research. For instance, in the example of approximation theory, we are given answers to questions that would be unsolvable in certain cases. However, it is always the case that the information gained within the model can be and often is translated back to the mathematical structure it represents.

Does this also imply (*purpose 2*) that modeling allows for further experimentation or observation in mathematics? To answer this, one may first have to wonder what a mathematical experiment is or can be. First of all, the *experimental mood* of the mathematician might be referred to, as a way of personal exploration in the mathematical field. However, this will not do here. A genuine experiment should at least have an element of systematic data-generating or testing. Notice that a field called *experimental mathematics* does in fact exist, and also has its own journal of that name. Let us quote from its editorial policy statement [24.12]:

“While we value the theorem-proof method of exposition, and do not depart from the established view that a result can only become part of mathematical knowledge once it is supported by a logical proof, we consider it anomalous that an important component of the process of mathematical creation is hidden from public discussion. It is to our loss that most of us in the mathematical community are

almost always unaware of how new results have been discovered. [...]

*Experimental Mathematics* was founded in the belief that theory and experiment feed on each other, and that the mathematical community stands to benefit from a more complete exposure to the experimental process. The early sharing of insights increases the possibility that they will lead to theorems [...]. Even when the person who had the initial insight goes on to find a proof, a discussion of the heuristic process can be of help, or at least of interest, to other researchers. There is value not only in the discovery itself, but also in the road that leads to it. [...]

The word *experimental* is conceived broadly: Many mathematical experiments these days are carried out on computers, but others are still the result of pencil-and-paper work, and there are other experimental techniques, like building physical models.”

Obviously, we particularly have to pick up this last sentence here. Next to number crunching (checking as many cases as possible) or probabilistic reasoning techniques, which – either or not aided by computers – have a distinct inductive and thus experimental ring to them, clearly also model building enters the picture.

This issue has been touched upon by *Van Bendegem* [24.13], characterizing an experiment as involving certain actions such as the manipulation of objects, setting up processes in the real world and observing possible outcomes of these processes. An example from mathematics that he discusses is the work of nineteenth-century Belgian physicist Plateau on minimal surface area problems. Plateau build several geometrical shapes of wire, and by dipping these into a soap solution he was able to investigate specific aspects of the minimum surface bounding various particular shapes. Here, we see how a physical experiment leads to relevant information of a mathematical problem. In such cases, we see how both the model and the physical prototype can influence further experiments and observations. On one hand, the physical experiments help one to formulate some general principles about a connected mathematical domain. On the other hand, the mathematician will set up his experiment in such a way to answer specific mathematical questions. However, such experiments are extremely rare in mathematical practice.

Another starting point can be the notion of a mathematical thought experiment, which *Van Bendegem* [24.14, pp. 9–10] characterizes as follows:

“If it is so that what mathematicians are searching for are proofs within the framework of a mathemati-

cal theory, then any consideration that (a) in the case where the proof is not yet available, can lead to an insight to what the proof could possibly look like, and, (b) in the case where the proof is available, can lead to a better understanding of that proof, can be considered to be a mathematical thought experiment.”

A specific example is a description of the octonions by means of (monoidal) category theory (rudimentary background information for this case is presented in the appendix.). It was shown in *Bulacu* [24.15] that octonions are in fact a weak Hopf algebra in the (braided) monoidal category constructed in *Albuquerque* and *Majid* [24.16], revealing thus more details of their algebraic structure. So in this sense, the work executed by Albuquerque and Majid influenced further *experimentation/observation*, leading eventually to a result which might have not been easily deduced by *algebraic world verification* alone, that is without using the results from [24.16]. The similarity with approximation theory, where the model provides information that remains hidden in the target system, should be clear. Since models can give us new information, this can lead to further experimentation or observation.

The notion of understanding (*purpose 3*), in its turn, is closely linked to that of explanation. Indeed, most of the traditional accounts of explanation state that understanding is centrally involved in it. *Achinstein* [24.17, p. 16] writes that there is a “fundamental relation between explanation and understanding”. *Kitcher* [24.18] argues that a “a theory of explanation shows us how scientific explanation advances our understanding” (p. 330). *Woodward* [24.19, p. 249] similarly says that any theory of explanation should “identify the structural features of such explanation which function so as to produce understanding in the ordinary user”.

Recently, we have seen an increasing interest in the topic of mathematical explanation as well (See *Mancosu* [24.20] for a useful overview of the literature.). Philosophical work on it can be divided into two main strands, namely focussing on extra-mathematical and intra-mathematical explanation, respectively. Extra-mathematical explanation is essentially about the role mathematics plays in the natural or social sciences, more precisely whether mathematics is or can provide explanations for physical phenomena. When considering intra-mathematical explanation, on the other hand, one looks into the role of explanation within mathematics itself, for example by distinguishing between explanatory and nonexplanatory proofs. The underlying idea is that all proofs tell us *that* a theorem is true, but only some proofs go further and tell us *why* a theorem is true. Steiner and Kitcher provided the

two best-known and the most discussed approaches to intra-mathematical explanation.

*Steiner* [24.21] uses the concept of *characterizing property* to draw a distinction between explanatory and nonexplanatory proofs. A characterizing property is a property unique to a given entity or structure within a *family* or domain of such entities or structures. The concept of a family is left undefined. According to Steiner, an explanatory proof always makes reference to a characterizing property of an entity or structure mentioned in the theorem. Furthermore, it must be evident that the result depends on the property (if we substitute the entity for another entity in the family which does not have the property, the proof fails to go through) and that by suitably *deforming* the proof while holding the *proof-idea* constant, we can get a proof of a related theorem. Though many of Steiner’s concepts (family, deformation, proof-idea) remain vague, he discusses several examples to clarify his account. He presents, for example, a proof of the irrationality of the square root of 2 as an explanatory proof since it depends on the unique prime factorization of 2 and since similar proofs for the irrationality of the square roots of other numbers can be given. Following this approach to explanations, models can foster understanding if the model produces proofs that depend on characterizing properties, or where it is easier for the mathematicians to identify these characterizing properties.

*Kitcher* [24.22, p. 437] also argues that his account covers mathematical explanations as well:

“The fact that the unification approach provides an account of explanation, and explanatory asymmetries, in mathematics stands to its credit.”

Let us briefly go over the model of unification that Kitcher proposes. Take a consistent and deductively closed set  $K$  of beliefs. A systematization of  $K$  is any set of arguments that derive some sentences of  $K$  from other sentences of  $K$ . The explanatory story, called  $E(K)$ , corresponds to the systematization with the highest degree of unification. The degree of unification is determined by the number of argument patterns, the stringency of patterns and the set of consequences derivable. Finally, an argument pattern is an argument that consists of schematic sentences, filling in instructions and classification of the sentences. Following Kitcher, and contrary to Steiner, there are no criteria that help us to analyze the explanatory power of a singular proof. Rather, explanation is presented as a value of a unified theory or systematization. Within this view, models can foster understanding if the model shows how mathematical results that were considered unrelated are in fact related. We can see, for example, how category the-

ory can advance such understanding. Category theory allows us to see the universal components of a family of structures of a given kind, and show how structures of different kinds are interrelated. Mathematical models can thus advance our understanding of mathematical results. But there are often different models that address the same mathematical result. Furthermore, the background knowledge and skills of a certain mathematician will play a role in determining whether a model grants understanding for this mathematician. The explanatory value of a mathematical model is, in this sense, a contextual notion.

Axiomatization (*purpose 4*) is undoubtedly another important aspect of mathematical practice. How can models assist it? A first observation is that axiomatization can appear quite arbitrary. Although axioms were once seen as self-evident truths about the constitution of the physical world, the emphasis nowadays mostly seems to be on deducing as much as possible from a minimum number of axioms, while the exact nature of these axioms is of secondary importance. Nevertheless, several mathematicians have argued against this so-called arbitrariness (Weyl [24.23], pp. 523–524), (Nevanlinna [24.24, p. 457]):

“One very conspicuous aspect of twentieth century mathematics is the enormously increased role which the axiomatic approach plays. Whereas the axiomatic method was formerly used merely for the purpose of elucidating the foundations on which we build, it has now become a tool for concrete mathematical research. [...] [However] without inventing new constructive processes no mathematician will get very far. It is perhaps proper to say that the strength of modern mathematics lies in the interaction between axiomatics and construction.

The setting up of entirely arbitrary axiom systems as a starting point for logical research has never led to significant results. [...] The awareness of this truth seems to have been dulled in the last few decades, particularly among younger mathematicians.”

Schlimm [24.25] (Sect. 3) has identified four nonarbitrary sources of (new or adapted) axioms from within mathematical practice:

1. Reasoning from accepted theorems, that is backward so to say, by wondering what axioms would be in need in order to substantiate current theories
2. Manipulation of existing axioms, as a way of (game-like) exploration
3. Conceptual analysis of a mathematical domain, such as e.g., number or set theory; and

4. Proofs and refutations, or the combination of the previous origins through the “various applications of initial conjectures, deductive arguments, semantic considerations, and different kinds of refinements” [24.25, p. 62].

It should be rather easily appreciated that model-based reasoning as it has been proposed by us here has an obvious role to play in processes like these.

The following example about structures called Hopf algebroids may be a good illustration [24.26]:

“A Hopf algebroid is a (possibly noncommutative) generalization of a structure which is dual to a groupoid (equipped with atlas) in the sense of space-algebra duality. This is the concept that generalizes Hopf algebras with their relation to groups from groups to groupoids.”

In Vercruyse [24.6], it is remarked that due to the asymmetry in this notion, several different notions of Hopf algebroid were introduced in the literature. Some of these were shown to be equivalent, although this was far from being trivial. The now seemingly overall accepted notion of a Hopf algebroid was introduced by Böhm and Szlachányi [24.27]. Lu [24.28] introduced a nonsymmetric version over a noncommutative base ring, hereby being able to include quite some examples. The definition by Schauenburg [24.29] allows one to recover a version of FT (Sect. 24.2) in this noncommutative setting, amongst other things. Only recently, Bruguières et al. [24.30] provided an interpretation of Schauenburg’s notion by means of so-called Hopf monads (Vercruyse [24.6, §5.2.2.1]):

“It took quite a long time to establish the correct Hopf-algebraic notion over a noncommutative base. The reasons for the difficulties are quite clear. First of all, if  $R$  is a noncommutative ring then the category of right  $R$ -modules is no longer monoidal (in general). Therefore we have to look instead to the category of  $R$ -bimodules, which is monoidal, but in general still not braided. So Hopf monoids cannot be computed *inside* this category. However, we can compute Hopf monads *on* this category (...) Historically, Hopf algebroids were constructed first in a more direct way, and the interpretation via Hopf monads is only very recent.”

The latter approach shows that for certain applications, it is preferable to use Schauenburg’s definition as being *conceptually* the most interesting one. The price to pay, however, is that it cannot include examples that are included in the slightly weaker notions

of Böhm and Szlachányi [24.27] and Lu [24.28], the definition from the latter source in its turn not being adapted to prove *categorically flavored* theorems (such as e.g., FT). As it seems, all depends what flavor one prefers.

Finally, that model-based reasoning should be able to foster mathematics and the art of making mathematical models (*purpose 5*), is of course self-evident in the context of mathematical inquiry itself, at least given all that has been elaborated above. The essence of mathematics resides in inventing methods, tools, strategies, and concepts for solving problems. That is the very answer to the question of why mathematicians prove theorems. From this view, Rav [24.31] concludes that proofs are the primary focus of mathematical interests, because these particular end products embody the methods, tools, strategies and concepts mentioned, and are therefore the true bearers of mathematical knowledge. Note that this goes against the received view that mathematicians are only interested in the mere truth of a theorem. Dawson [24.32] has in this respect discussed no less than eight reasons for mathematicians to look for multiple proofs of the same theorem:

1. To remedy perceived gaps or deficiencies in earlier arguments
2. To employ reasoning that is simpler, or more perspicuous, than earlier proofs
3. To demonstrate the power of different methodologies
4. To provide a rational reconstruction (or justification) of historical practices
5. To extend a result, or to generalize it to other contexts
6. To discover a new route
7. Concerns for methodological purity
8. Role analogous to the role of confirmation in the natural sciences.

Note that several of these reasons have been discussed in previous paragraphs. Indeed it is not hard to see that model-based reasoning plays an important role in fostering mathematics in the sense that its practitioners are interested in the values of *specific* (and not just of *any*) proofs of both conjectures and existing theorems, which enables the mathematician to discover new routes, demonstrate the power of different methodologies, search for a simple argument, etc.

## 24.5 Conclusion

As announced at the outset of this chapter, we have been focusing here on the philosophical significance of one specific aspect of mathematical practice, namely model-based reasoning as a general methodological framework. By presenting three cases, taken from different mathematical subdisciplines and with varying levels of abstraction, we showed how mathematicians engage in model-based reasoning. We are well aware that the general account of such reasoning remains silent on how mathematicians go from one level to another level. Philosophers such as Pólya and Lakatos discuss the richness of different intellectual heuris-

tics that mathematicians use, and further research into these processes is certainly welcome. Nevertheless, the discussion of the mathematical fruits of model-based reasoning should convince the reader of the significance of the general framework of model-based reasoning, as it shows us how mathematical modeling is linked with several specific purposes of mathematical practice such as experimentation, understanding, or axiomatization. Hence, future reflections on the specifications of model-based reasoning in mathematics can provide crucial insights in several interesting questions about mathematical practice.

## 24.A Appendix

In this appendix, we briefly recall some notions from (monoidal) category theory. Classical references for category theoretical notions and constructions are Borceux [24.33] and Mac Lane [24.34]. We start with some basic notions from set theory. Let us consider two nonempty sets  $A$  and  $B$  and a set-theoretical map (or function)  $f$  between them. This situation can be depicted as follows:

$$A \xrightarrow{f} B$$

that is, this *process*  $f$  can be visualized by an *arrow*; the only requirement being that one must be able to tell for every element of the departure set  $A$  where it is going to. Now, let  $A, B, C$  be three sets and consider two

functions  $f$  and  $g$  as follows:

$$A \xrightarrow{f} B \xrightarrow{g} C$$

We can now consider the composition, denoted by  $g \circ f$ :

$$A \xrightarrow{f} B \xrightarrow{g} C$$

$$\quad \quad \quad \overset{g \circ f}{\curvearrowright}$$

The composition of functions has the associative property: whenever  $f, g, h$  are functions (that can be composed), one has  $(h \circ g) \circ f = h \circ (g \circ f)$ . Remark also that for any set  $A$ , we can consider the function that maps any element of  $A$  onto itself:

$$A \xrightarrow{1_A} A$$

This function  $1_A$  is called the identity function on  $A$ . It has the property that for any function  $f : A \rightarrow B$ , the following holds:  $f \circ 1_A = 1_B \circ f = f$ .

Let us now consider a more general scenario, not necessarily set-theoretic. Let  $A$  and  $B$  be *objects*:

$$A \qquad B$$

and replace the set-theoretical notion of map by just an arrow between these objects:

$$A \xrightarrow{f} B$$

We can now give an idea of the notion of *category*:

Roughly speaking, a *category*  $\underline{C}$  consists of objects and arrows (between objects) such that there is a *composition*  $\circ$  for the arrows and an identity arrow  $1_A$  for any object  $A$  of  $\underline{C}$ . These ingredients have to satisfy some conditions that mimic the associative behavior of composition of functions between sets and the above-mentioned property of the identity function on any set.

In this sense, following *Awodey* [24.35], category theory might be called *abstract function theory*.

We give some basic examples of categories:

- $C = \mathbf{Sets}$   
objects: sets  
arrows: functions between sets
- $C = \mathbf{Vect}_k$  ( $k$  being a field)  
objects:  $k$ -vector spaces  
arrows:  $k$ -linear maps.

Now we would like to illustrate the adjective *monoidal* in the term *monoidal category*. Therefore, let us introduce monoids.

A *monoid*  $(M, *, 1_M)$  consists of a set  $M$ , a function  $*$  :  $M \times M \rightarrow M$  and an element  $1_M \in M$  such that:

- We have  $1_M * m = m * 1_M = m$ , for any  $m \in M$
- For any three  $m, n, p \in M$  the following holds
 
$$(m * n) * p = m * (n * p).$$

We are ready to sketch what a monoidal category looks like. Very roughly speaking, a *monoidal category* is a category  $\underline{C}$ , in which we can *multiply* objects and arrows (this *multiplication* is denoted by  $\otimes$ ) and in which a *unit object* exists (denoted by  $I$ ) such that this  $\otimes$  (resp.  $I$ ) *imitate the behavior* of the operation  $*$  (resp. the element  $1_M$ ) from the monoid structure  $(M, *, 1_M)$ .

We will denote such a monoidal category  $(\underline{C}, \otimes, I)$  briefly by  $\underline{C}$  in the sequel.

Actually, in technical terms, the definition of monoidal category is precisely the *categorification* of the definition of monoid (here categorification aims at the name for the process as it was coined by *Crane* and *Yetter* in [24.36]).

Here are some examples of monoidal categorical structures:

- $(\underline{C}, \otimes, I) = (\mathbf{Sets}, \times, \{*\})$ , where:
  - $\times$  is the Cartesian product of sets.
  - $\{*\}$  is any singleton.
 We will briefly denote this monoidal category by  $\mathbf{Sets}$ .
- $(\underline{C}, \otimes, I) = (\mathbf{Vect}_k, \otimes_k, k)$ , where:
  - $\otimes_k$  is the tensor product over  $k$
  - $k$  is any field.

This example will be briefly denoted by  $\mathbf{Vect}_k$ .

Now we have a vague idea of what it means to be a monoidal category, in order to illustrate an example, we wish to glance at certain objects in such categories. More precisely, we start with considering *monoids* (sometimes called *algebras* in literature) in a monoidal category  $\underline{C}$ . The idea is that these objects mimic the behavior of *classical* monoids (i.e., sets with an associative, unital binary operation), the language of monoidal categories offering a natural setting to do so; this is an instance of the so-called microcosm principle of *Baez* and *Dolan* [24.37], affirming that “certain algebraic structures can be defined in any category equipped with a categorified version of the same structure.”

A *monoid* in  $\underline{C}$  is a triple  $A = (A, m, \eta)$ , where  $A \in \underline{C}$  and  $m : A \otimes A \rightarrow A$  and  $\eta : I \rightarrow A$  are arrows in  $\underline{C}$  (such that two diagrams – respectively mimicing the associativity and unitality condition – commute; we refer the reader to [24.6, Sect. 5.3.1] for instance).

Many algebraic structures can be seen as monoids in an appropriate monoidal category; we present some examples here, for details and more examples we refer the reader again to [24.6, Sect. 5.3.1] e.g.:

- Taking  $C$  to be the monoidal category  $\underline{\text{Sets}}$ , one can easily verify that, taking a monoid in  $C$ , one recovers exactly the definition of a classical monoid, as one expects.
- Similarly, a monoid in  $\underline{\text{Vect}}_k$  gives precisely the classical notion of (an associative, unital)  $k$ -algebra.
- A more surprising example is given by the octonions. The octonions  $\mathbb{O}$  are a normed division algebra over the real numbers. There are only four such algebras, the other three being the real numbers, the complex numbers, and the quaternions. Although not as well known as the quaternions or the complex numbers, the octonions are related to a number of exceptional structures in mathematics, among them the exceptional Lie groups. For more details, we refer to the excellent paper by Baez on this subject [24.38]. One of the properties of the octonions is that they are nonassociative (that is, considered as monoid in  $\underline{\text{Vect}}_k$ ). They can be seen, however, as an (associative) monoid in the monoidal category constructed by *Albuquerque* and *Majid* in [24.16].

In case a monoidal category  $\underline{C}$  exhibits moreover a *braided* structure (whatever this means), we denote  $\underline{C}$  equipped with this braided structure as  $\widetilde{C}$ . In this case, one can not only consider monoids in  $\widetilde{C}$ , one can impose more structure on the definition of monoid, obtaining such notion as *Hopf monoid in  $\widetilde{C}$* . To be a bit more precise, a Hopf monoid in  $\widetilde{C}$  is a bimonoid (which is a monoid also having a so-called comonoid structure, both structures being compatible), having an antipode. The reader is referred to [24.6, Sect. 5.3.2] for more details.

The categories  $\underline{\text{Sets}}$  and  $\underline{\text{Vect}}_k$  can be given a braided structure, which we denote by  $\widetilde{\text{Sets}}$  and  $\widetilde{\text{Vect}}_k$  respectively, such that – without going into the details – the notions of *group* and *Hopf algebra* can be recovered as being Hopf monoids in  $\widetilde{\text{Sets}}$  and  $\widetilde{\text{Vect}}_k$ , respectively.

**Acknowledgments.** The authors are mentioned in alphabetical order. The first author is a doctoral research assistant of the Fund for Scientific Research – Flanders. The second author would like to thank Joost Vercruyse for fruitful discussion. The third author is indebted to research project SRP22 of Vrije Universiteit Brussel.

## References

- 24.1 R. Aris: *Mathematical Modelling Techniques* (Pitman, San Francisco 1978)
- 24.2 P.J. Davis, R. Hersh: *The Mathematical Experience* (Penguin Books, London 1983)
- 24.3 G. de Vries: Slides from Workshop on Mathematical Modelling, June 2001 Mathematics Symposium: Focus on Applied and Pure Mathematics, Edmonton Regional Consortium (2001)
- 24.4 N.M. Hrushikesh, V.P. Devidas: *Fundamentals of Approximation Theory* (Narosa Publishing House, New Dehli 2000)
- 24.5 O. Christensen, K.L. Christensen: *Approximation Theory: From Taylor Polynomials to Wavelets* (Springer, New York 2005)
- 24.6 J. Vercruyse: Hopf algebras. Variant notions and reconstruction theorems. In: *Quantum Physics and Linguistics: A Compositional, Diagrammatic Discourse*, ed. by E. Grefenstette, C. Heunen, M. Sadrzadeh (Oxford Univ. Press, Oxford 2013) pp. 115–146
- 24.7 M. Takeuchi: Finite Hopf algebras in braided tensor categories, *J. Pure Appl. Algebr.* **138**, 59–82 (1999)
- 24.8 R. Hersh: Mathematics has a front and a back, *Synthese* **88**, 127–133 (1991)
- 24.9 G. Pólya: From the preface of induction and analogy in mathematics. In: *New Directions in the Philosophy of Mathematics, Revised and expanded edn*, ed. by T. Tymoczko (Princeton Univ. Press, Princeton 1998) pp. 99–101
- 24.10 I. Lakatos: *Proofs and Refutations* (Cambridge Univ. Press, Cambridge 1976)
- 24.11 C. Pincock: A revealing flaw in Colyvan's indispensability argument, *Philos. Sci.* **71**, 61–79 (2004)
- 24.12 D. Epstein, S. Levy, R. de la Llave: Statement of philosophy and publishing criteria, *Exp. Math.* **1**, 1–3 (1992)
- 24.13 J.P. Van Bendegem: What, if anything, is an experiment in mathematics? In: *Philosophy and the Many Faces of Science*, ed. by D. Anapolitanos, A. Baltas, S. Tsinoema (Rowman and Littlefield, London 1998) pp. 172–182
- 24.14 J.P. Van Bendegem: Thought experiments in mathematics: Anything but proof, *Philosophica* **72**, 9–33 (2003)
- 24.15 D. Bulacu: The weak braided Hopf algebra structure of some Cayley–Dickson algebras, *J. Algebr.* **322**, 2404–2427 (2009)
- 24.16 H. Albuquerque, S. Majid: Quasialgebra structure of the octonions, *J. Algebr.* **220**, 188–224 (1999)
- 24.17 P. Achinstein: *The Nature of Explanation* (Oxford Univ. Press, Oxford 1983)
- 24.18 P. Kitcher: Explanatory unification. In: *The Philosophy of Science*, ed. by R. Boyd, P. Gasper, J.D. Trout (MIT Press, Cambridge 1988) pp. 329–347
- 24.19 J. Woodward: A theory of singular causal explanation. In: *Explanation*, ed. by R. David–Hillel (Oxford Univ. Press, New York 1993) pp. 246–274
- 24.20 P. Mancosu: Mathematical explanation: Why it matters. In: *The Philosophy of Mathematical Prac-*

- tice*, ed. by P. Mancosu (Oxford Univ. Press, Oxford 2008) pp. 134–150
- 24.21 M. Steiner: Mathematical explanation, *Philos. Stud.* **34**, 135–151 (1978)
- 24.22 P. Kitcher: Explanatory unification and the causal structure of the world. In: *Scientific Explanation*, ed. by P. Kitcher, W. Salmon (Univ. Minnesota Press, Minneapolis 1989) pp. 410–505
- 24.23 H. Weyl: A half-century of mathematics, *Am. Math. Mon.* **58**, 523–533 (1951)
- 24.24 R. Nevanlinna: Reform in teaching mathematics, *Am. Math. Mon.* **73**, 451–464 (1966)
- 24.25 D. Schlimm: Axioms in mathematical practice, *Philos. Math.* **21**(1), 37–92 (2013)
- 24.26 Zoran Škoda: Hopf Algebroid <http://ncatlab.org/nlab/show/Hopf+algebroid> (2014)
- 24.27 G. Böhm, K. Szlachányi: Hopf algebroids with bijective antipodes: Axioms, integrals and duals, *Commun. Algebr.* **32**, 4433–4464 (2004)
- 24.28 J.-H. Lu: Hopf algebroids and quantum groupoids, *Int. J. Math.* **7**, 47–70 (1996)
- 24.29 P. Schauenburg: Bialgebras over noncommutative rings and a structure theorem for Hopf bimodules, *Appl. Categ. Struct.* **6**, 193–222 (1998)
- 24.30 A. Bruguières, S. Lack, A. Virelizier: Hopf monads on monoidal categories, *Adv. Math.* **227**, 745–800 (2011)
- 24.31 Y. Rav: Why do we prove theorems?, *Philos. Math.* **7**, 5–41 (1999)
- 24.32 J. Dawson: Why do mathematicians re-prove theorems?, *Philos. Math.* **14**, 269–286 (2006)
- 24.33 F. Borceux: *Handbook of Categorical Algebra I. Encyclopedia of Mathematics and Its Applications*, Vol. 50 (Cambridge Univ. Press, Cambridge 1994)
- 24.34 S. Mac Lane: *Categories for the Working Mathematician, Graduate Texts in Mathematics Ser., Vol. 5, second edn* (Springer, Berlin 1998)
- 24.35 S. Awodey: *Category Theory, Oxford Logic Guides Ser., second edn* (Oxford Univ. Press, Oxford 2010)
- 24.36 L. Crane, D.N. Yetter: Examples of categorification, *Cah. Topol. Géom. Différ. Catég.* **39**, 325 (1998)
- 24.37 J.C. Baez, J. Dolan: Higher-dimensional algebra III.  $n$ -categories and the algebra of opetopes, *Adv. Math.* **135**, 145–206 (1998)
- 24.38 J.C. Baez: The octonions, *Bull. Am. Math. Soc.* **39**, 145–205 (2001)

## 25. Abduction and the Emergence of Necessary Mathematical Knowledge

Ferdinand Rivera

The prevailing epistemological perspective on school mathematical knowledge values the central role of induction and deduction in the development of necessary mathematical knowledge with a rather taken-for-granted view of abduction. This chapter will present empirical evidence that illustrates the relationship between abductive action and the emergence of necessary mathematical knowledge.

Recent empirical studies on abduction and mathematical knowledge construction have begun to explore ways in which abduction could be implemented in more systematic terms. In this chapter four types of inferences that students develop in mathematical activity are presented and compared followed by a presentation of key findings from current research on abduction in mathematics and science education. The chapter closes with an exploration of ways in which students can effectively enact meaningful and purposeful abductive thinking processes through activities that enable them to focus on relational or orientation understandings. Four suggestions are provided, which convey the need for meaningful, structured, and productive abduction actions. Together the suggestions target central features in

25.1	<b>An Example from the Classroom</b> .....	551
25.2	<b>Inference Types</b> .....	555
25.2.1	Abduction .....	557
25.2.2	Induction .....	558
25.2.3	Deduction and Deductive Closure.....	559
25.3	<b>Abduction in Math and Science Education</b> .....	561
25.3.1	Different Kinds of Abduction.....	561
25.3.2	Abduction in Mathematical Relationships .....	562
25.4	<b>Enacting Abductive Action in Mathematical Contexts</b> .....	564
25.4.1	Cultivate Abductively-Infused Guesses with Deduction .....	564
25.4.2	Support Logically-Good Abductive Reasoning.....	565
25.4.3	Foster the Development of Strategic Rules in Abductive Processing.....	565
25.4.4	Encourage an Abductive Knowledge-Seeking Disposition .....	565
	<b>References</b> .....	566

abductive cognition, that is, thinking, reasoning, processing, and disposition.

### 25.1 An Example from the Classroom

Table 25.1 provides a short transcript of a very interesting classroom episode on counting by six that happened in a US first-grade class. The task, which was about determining the total number of faces for four separate cubes, was given to the students to help them apply and practice the arithmetical strategy of *counting on*. Anna, Betsy, and all the students together in a chorus-like manner in lines 9, 13, and 17, respectively, eagerly modeled the same process of *putting the last known number in their head and counting six more*. The episode became interesting when Ian started to employ counting by five, an arithmetical skill that the class already knew, to help him count by six in a systematic

way. As conveyed in line 20, Ian initially *saw* multiples of five in the sequence (6, 12, 18, 24). In line 21, when he *added the ones* and saw that the numbers in his head matched the same numbers he saw on the teacher's board, the feeling of having *discovered* a wonderful idea caused him to exclaim *I was right!* and encouraged him to share his abduction with his classmates (lines 22–26).

*Shotter* [25.1] captures the following sense in which first-grade student Ian has embodied abductive thinking in relation to the number sequence (6, 12, 18, and 24): Ian was “carried away unexpectedly by an other or otherness to a place not previously familiar to him” [25.1,



**Table 25.1** Ian’s counting-by-six rule

Ms. Marla [M] presented the following task below during a board math session with her first-grade class	
Number of cubes	Number of faces
1	–
2	–
3	–
4	–
How many faces do four cubes have in all?	
1	M: Let’s say I have four cubes. I want to know how many faces four cubes have in all. So let’s count how many faces one cube would have.
2	
3	Students [Ss]: M points to the faces one by one. One, two, three, four, five, six.
4	M: Okay so how many faces does one cube have?
5	Ss: Six!
6	M: Now I want to know how many faces two cubes would have.
7	Ss: Twelve!
8	M: Let’s see. How would I figure that out?
9	Anna: Put six in your head and count six more.
10	M: Okay so?
11	Ss: 6, 7, 8, 9, 10, 11, 12.
12	M: Okay next.
13	Betsy: You put 12 in your head and count six more.
14	M: Okay everybody!
15	Ss: 12, 13, 14, 15, 16, 17, 18.
16	M: Then?
17	Ss: You put 18 in your head and count. 18, 19, 20, 21, 22, 23, 24.
18	M: So how many faces are there in all?
19	Ss: 24.
As the students began to count by six, Ian [I] decided to count by five using his right hand to indicate one set of 5.	
20	I: 5, 10, 15, 20.
He then used his right thumb and continued to count by one.	
21	I: And then you add the ones. 21, 22, 23, 24. I was right!
Ian eagerly raised his hand and shared his strategy with Ms. Marla and his classmates.	
22	I: Ms. M, I was thinking that in my head. . . Ms. M I know another idea . . .
23	because you have all those sixes and you count by fives and there’s only ones
24	left.
25	M: So you went 5, 10, 15, 20. [Ian nods].
26	I: 21, 22, 23, 24.
27	M: Excellent!

p. 225]. He was pleasantly surprised about how easy it was to count “all the sixes” by “counting by fives and adding the ones left”, which generated in him an intense feeling of discovering something new through a guess that made sense and that he was able to verify to be correct. The following passage below from *El Khachab* [25.2] provides another, and yet deeper, way of thinking about Ian’s experience. El Khachab foregrounds the significance of having a *purpose* as a way of motivating the emergence of new ideas, which is one way of explaining how learners sometimes find themselves being carried away during the process of discovery. The second sentence in the passage articulates in very clear terms the primary purpose of abduction and its central and unique role in the establishment of new knowledge [25.2, p. 172]:

“Before asking where new ideas come from, we need to ask what new ideas are for, and knowing what they are for, we can attune their newness to their purpose. And their purpose is, in the case of abduction, to provide true explanations following experimental verification.”

Ian saw purpose in counting by *five plus one* that encouraged him to further pursue his *new* idea. After verifying that his strategy actually worked on the available cases, he then articulated an explanation that matched what he was *thinking in his head*. The nature of what counts as a *true explanation* in abduction is explored in some detail in the succeeding sections. For now, it makes sense to think of abductive explanations as modeling instances of “relational or orientational

way of knowing”, which is a type of “embodied coping” that attends to [25.2, p. 172]

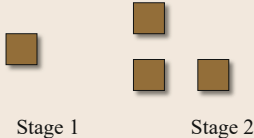
“the possible relations – what we might call the *relational dimensions* – that exist as a dynamical outcome of the interacting of objectively observable phenomena which are not in themselves objectively observable.”

Ian’s abductive thinking about counting by six is worth noting early in this chapter in light of recent findings on children’s algebraic thinking that show how many of them tend to use their knowledge of the multiplication table to help them generate and establish mathematical relationships and support their ability to construct explicit or function-based formulas involving linear patterns [25.3].

US eighth-grade student Dung’s figural processing of the two pattern generalization tasks shown in Figs. 25.1 and 25.2 illustrates another characterization of abductive thinking that “carries over a deeper similarity to a number of seemingly rather different sit-

uations” [25.1, p. 225]. Dung’s processing illustrates a kind of *double description* (i. e., in Bateson’s [25.5, p. 31] sense of “cases in which two or more information sources come together to give information of a sort different from what was in either source separately”) that is a necessary condition when students are engaged in mathematical thinking and learning. When Dung was presented with the ambiguous Fig. 25.1 task consisting of two beginning stages in a growing pattern, he constructed a growing sequence of L-shaped figures (Fig. 25.3). When he was asked to generate explicit rules for his pattern, he suggested  $s = n + n - 1$  and  $s = 2n - 1$ . When he was asked to justify them, Dung saw the pattern stages in terms of groups of squares. In the case of his first rule, each stage in his growing pattern consisted of the union of two variable units having cardinalities  $n$  and  $(n - 1)$  corresponding to the column and row of squares, respectively (see Fig. 25.3 stage 3 for an illustration). In the case of his second rule, two composite sides of squares that had the same number of squares on each side overlapped along the corner square (see Fig. 25.3 stage 5 for an example).

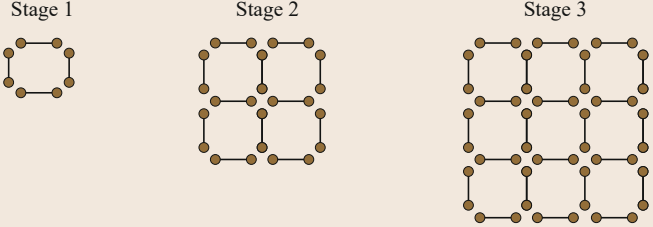
Below are the first two stages in a growing pattern of squares



1. Continue the pattern until stage 5.
2. Find a direct formula in two different ways. Justify each formula.
3. If none of your formulas above involve taking into account overlaps, find a direct formula that takes into account overlaps. Justify your formula.
4. How do you know for sure that your pattern will continue that way and not some other way?
5. Find a different way of continuing the pattern and obtain a direct formula for this pattern.

Fig. 25.1 Ambiguous patterning task in compressed form (after [25.4])

Consider the following array of sticks below



- A. Find a direct formula for the total number of sticks at any stage in the pattern. Justify your formula.
- B. Find a direct formula for the total number of points at any stage in the pattern. Justify your formula.

Fig. 25.2 Square array pattern (after [25.4])

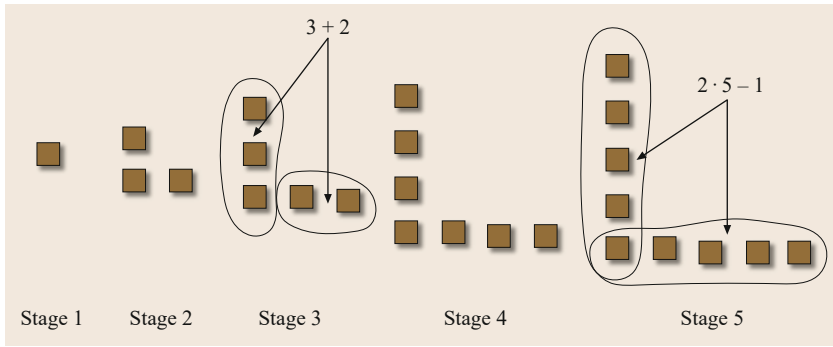


Fig. 25.3 Dung’s growing L shaped pattern (after [25.4])

For Dung, seeing pattern stages in terms of groups enabled him to justify his explicit rules, which became his abductive resource for constructing and justifying an explicit rule for the square array pattern shown in Fig. 25.2. Dung initially saw each pattern stage into parts of separate rows of squares and separate smaller squares per row (Fig. 25.4). Using stage 4, he parsed the whole figure into four disjoint rows and counted the number of sticks per row. In counting the number of sticks per row, he saw four disjoint squares for a total of  $4 \times 4 = 16$  sticks and then subtracted the three overlapping vertical sticks. He then counted the total number of horizontal and vertical sticks counting repetitions and obtained  $(4 \times 4) \times 4 = 52$ . In his written work, he immediately resorted to the use of a variable  $n$  to convey that he was thinking in general terms, which explains the expression  $(4n - (n - 1)) \times n$ . Since he also saw that the four disjoint rows had overlapping sides (i. e., the interior horizontal sticks), he then took away three ( $= 4 - 1$ ) groups of such four horizontal sticks from 52. That concrete step allowed him to complete his explicit rule

for the pattern, that is,  $s = (4n - (n - 1))n - (n - 1)n$ , which he then simplified to  $s = 2n^2 + 2n$ . Dung’s multiplicative thinking ability became his abductive – that is, double descriptive – abstracting resource that enabled him to infer deeper similarity among, and thus generalize to, different kinds of patterns.

In this chapter, we explore the relationship between abductive action and the emergence of necessary mathematical knowledge. The prevailing epistemological perspective on mathematical knowledge values the central role of induction and deduction in the development of necessary mathematical knowledge with a rather taken-for-granted view of abduction that in the past has been characterized as the creative, wild, and messy space of theory generation or construction. However, recent empirical studies on abduction and mathematical knowledge construction have begun to explore ways in which abduction could be implemented in more systematic terms beyond a way of reasoning by detectives from observations to explanations [25.6, p. 24] and merely “studying facts and devising a theory to ex-

A. Find a direct formula for the total number of sticks at any stage in the pattern. Justify your formula.

$$s = 4n^2 - (4n - (n - 1))n - (n - 1)n$$

$$s = 2n^2 + 2n$$

$$s = n(4n - (n - 1)) - n^2 + n$$

$$s = n(3n + 1) - n^2 + n$$

Fig. 25.4 Dung’s construction and justification of his formula for the Fig. 25.2 pattern (after [25.4])

plain them” because “its only justification is that if we are ever to understand things at all, it just be in that way” [25.7, p. 40]. For instance, *Mason et al.* [25.8] associate abductive processing with the construction of structural generalizations, while *Pedemonte* [25.9] situates abduction within a cognitive unity thesis that sees it as being prior and necessary to induction and ultimately deduction. Recent investigations in science and science education that pursue an abductive framework also underscore the central role of abduction in inference systems that model everyday phenomena. For instance, *Addis and Gooding* propose the iterative cycle of “abduction (generation) → deduction (prediction) → induction (validation) → abduction” in modeling the “scientific process of interpreting new or surprising findings by generating a hypothesis whose consequences are then evaluated empirically” [25.10, p. 38]. Another instance involves *Magnani’s* [25.11] formulation of actual computational models in which case abduction is seen as central to the development of creative reasoning in scientific discoveries and can thus be used to generate rational models.

In Sect. 25.2, we provide a characterization of the four types of inferences that students develop in mathematical activity. In Sect. 25.3 we note two key findings

from current research on abduction in mathematics and science education, which should provide the necessary context for understanding the ideas we pursue in the succeeding section. In Sect. 25.4 we explore ways in which students can effectively enact meaningful and purposeful abductive thinking processes and other [25.1, p. 224]

“kinds of *preparing activities* in mathematical learning contexts that will enable learners to become self-consciously engaged in, can get them *ready* to notice, immediately and spontaneously, the kinds of events relevant to their acquiring such relational or orientation understandings – where, by *being ready to do something* means what we often talk of as being in possessions of a *habit*, an *instinct*, an *inclination*, etc.”

Central to such processes and activities involves orchestrating effective tasks and other learning contexts that will engage all students in abductive thinking, which will go a long way in supporting growth in necessary mathematical knowledge and excellence in reasoning that is strategic and has “logical virtue (i. e., avoiding logical fallacies and learning what is and what is not admissible and valid)” [25.12, p. 269].

## 25.2 Inference Types

Table 25.2 lists the characteristics of four types of inferences that students develop in mathematical activity. *Abduction* involves generating a hypothesis or narrowing down a range of hypotheses that is then verified via *induction*. Abduction is the source of original ideas and is initially influenced by prior knowledge and experiences, unlike induction that basically tests an abductive claim on specific instances. The hope, of course, is that possible errors get corrected through the inductive route, which results in the construction of a generalization that draws on the available instances. Like induction, which performs the role of verifying an abductive claim, *deduction* produces results from general rules or laws and thus does not produce any original ideas. Unlike abduction, which is sensitive to empirical data, deduction relies on unambiguous premises in order to ascertain the necessity of a single valid conclusion. An *unambiguous* well-defined set or model assumes the existence of “a finite set of rules and without reference to context” that clearly defines membership or relationships among the elements in the set [25.10, p. 38].

Another useful way to think about abduction and deduction involves truth tables. Deductions depend on truth tables for validity, which also means to say that

the objects and rules of deduction all have to be well established and well defined. Abductions do not depend on truth tables and their validity is established via induction [25.10, p. 37–38]. *Deductive closure* conveys deductively derived arguments and instances and is a necessary condition for algebraic thinking in both symbolic and nonsymbolic contexts [25.13].

Consider, for example, the following four statements below that have been extracted from eighth-grade student *Cherrie’s* generalization of the Fig. 25.3 pattern:

Law (L): I think the rule is  $x = 2(n + 1)n$ .

Case (C): In stage 2, there’s two groups of three twice. There’s two four groups of three in stage 3. There’s two five groups of four in stage 4.

Result (R): Stages 1 through 4 follow the rule  $x$  equals two times  $(n + 1)$  times  $n$ .

All future outcomes (O): Stage 5 has  $26(5) = 60$  sticks, stage 10 has  $211(10) = 220$  sticks, and stage 2035 has 8 286 520 sticks in all.

Deduction assumes a general law and an observed case (or cases) and infers a necessary valid result, which also means that it does not have to depend on real or empirical knowledge for verification [25.14]. Cases

**Table 25.2** Types of inferences in mathematical activity and their characteristics

Inferential type	Inferential form	Intent	Inferential attitude	Sources	Desired construction	Nature of context, verification, and justification
Abduction	From result and law to case	Depth (intentional)	Entertains a plausible inference toward a rule Generates and selects an explanatory theory – that <i>something maybe</i> (conjectural)	Unpredictable (surprising facts; flashes; intelligent guesses; spontaneous conjectures)	Un/ Structured	Context-bound; Structured via induction
Induction	From result and more cases to law	Breadth (extensional)	Tests an abduced inference; measures the value and degree of concordance of an explanatory theory to cases – that <i>something actually is operative</i> (approximate)	Predictable (examples)	Structural based on abduction	Context-derived; empirical (e.g., enumeration, analogy, and experiments)
Deduction	From rule and case to result	Logical proof	Predicts in a methodical way a valid result – that <i>something must be</i> (certain)	Predictable (premises)	Structural (canonical form)	Decontextualized; Steps in a proof
Deductive closure	From an established deduction to future outcomes	Breadth (apply)	Assumes that all future outcomes will behave in the same manner as a result of a valid deductive hypothesis	Predictable (premises are valid deductions)	Structural based on an established deduction	Decontextualized; Mathematical induction (e.g., demonstration of a valid deductive claim)

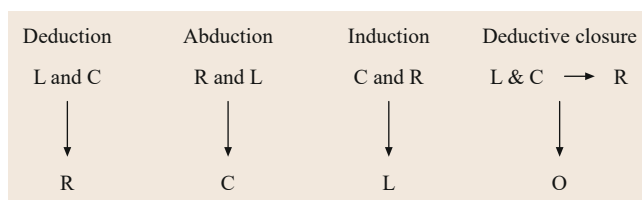
are occurrences or instantiations of the stipulated law. When the first three statements above are switched in two different ways, we obtain the canonical structures for abduction and induction, which are ampliative because the conclusions “amplify or go beyond the information incorporated in the premises” [25.11, p. 511] and invalid (i. e., not necessary) from a deductive point of view. In a deductive closure, an established deduction becomes the cause or hypothesis that is then applied to future outcomes, which are effects. Figure 25.5 visually captures the fundamental differences among the four inferential types.

From a logicopsychological perspective, students need to learn to anticipate inferences that are sensible and valid in any mathematical activity. *Peirce* [25.15, p. 449], of course, reminds us that context matters despite our naturally drawn disposition toward “perpetually making deductions”. As an aside, kindergarten students (ages five to six years) in the absence of formal

learning experiences appear to consider deductive inferences as being more certain than inductive ones and other guesses [25.16].

Students also need to understand the limitations of each inferential process. For *Polya* [25.17], deduction exemplifies *demonstrative reasoning*, which is the basis of the “security of our mathematical knowledge” [25.17, p. v] since it is “safe, beyond controversy, and final”. Abduction and induction exemplify *plausible reasoning*, which “supports our conjectures” and could be “hazardous, controversial, and provisional” [25.17]. Despite such constraints, however, Peirce and Polya seem to share the view that abduction, induction, and deduction are epistemologically necessary. According to *Polya* [25.17], while “anything new that we learn about the world involves plausible reasoning”, demonstrative reasoning uses “rigid standards that are codified and clarified by logic” [25.17, p. v]. *Polya*’s perspectives are narrowly confined to how we come to understand and explain the nature of mathematical objects, unlike Peirce who formulates his view by drawing on his understanding of the nature of scientific practice. “All ideas of science come to it by way of abduction”, *Peirce* writes, which is the fundamental source of the emergence of ideas and “consists in studying facts and devising a theory to explain them” [25.7, p. 90].

In the next three subsections below, we discuss additional characteristics of each inferential type.

**Fig. 25.5** Differences among the four inferential types

### 25.2.1 Abduction

Abduction, the source of original ideas, discoveries, and explanatory theories, emerges and evolves in a continuum of thought processes, uncontrolled and instinctual (e.g., as initial impressions based on perceptions and informed guesses) in the early phase and structured and inferential (e.g., quasiductive) in a much later phase [25.2]. Through abduction, “descriptions of data, patterns, or phenomena” are inferred leading to plausible explanations, hypotheses, or theories “that deserve to be seriously entertained and further investigated” ([25.18, p. 1021], [25.11, p. 511]). Perceptual-like clues provide one possible source of abductive ideas [25.19]. The steps below outline a percept-based “formula that is similar to abduction” [25.19, p. 305].

“A well-recognized kind of object,  $M$ , has for its ordinary predicates  $P[1]$ ,  $P[2]$ ,  $P[3]$ , etc., indistinctly recognized. The suggesting object,  $S$ , has these same predicates,  $P[1]$ ,  $P[2]$ ,  $P[3]$ , etc. Hence,  $S$  is of the kind  $M$ .”

Iconic-based inferences also provide another possible source of abduction [25.19]. Icons, unlike percepts, are pure possible forms of the objects they represent or resemble. Iconic-based abductions employ the following abductive process [25.19, p. 306]:

$$\left. \begin{array}{l} P1 \\ H1 \end{array} \right\} \rightarrow \text{An iconic relationship between } P1 \text{ and } P2$$

$P1$  and  $P2$  are similar (iconically)

$\therefore$  Maybe  $H1$  (or something that is similar to  $H1$ ).

Abduction also involves “the problem of logical goodness, i.e., how ideas fulfill their logical purpose in the world” [25.2, pp. 159, 162]. *El Khachab* [25.2] uses the example of global warming to show how different stakeholders tend to model different kinds of goodness based on their purpose. Following Peirce, he notes that “the purpose of abduction is to provide hypotheses which, when subjected to experimental verification, will provide true explanations” [25.2, p. 162]. True explanations refer to “sustainable belief-habits, that is, as recurring settlements of belief about the world which rely on experientially or experimentally verifiable statements” [25.2, p. 163].

We note the following four important points below about abduction.

*First*, *Tschaep* [25.20] underscores the significance of guessing in abduction, that is [25.20, p. 117],

“guessing is the initial deliberate originary activity of creating, selecting, or dismissing potential solu-

tions to a problem as a response to the surprising experience of that problem.”

Having a guess enables learners to transition from the first to the second premise in Peirce’s general syllogism for abduction (i.e., the surprising fact,  $C$ , is observed; but if  $A$  were true,  $C$  would be a matter of course; hence, there is reason to suspect that  $A$  is true). Following *Kruijff* [25.21], *Tschaep* notes that guessing and perceptual judgment (i.e., observing a surprising fact  $C$ ) are “the two essential aspects that characterize the generation of ideas” [25.20, p. 117], where the event of surprise emerges from every individual knower’s experiences, which is perceptual in nature. Guessing, then [25.20, p. 117],

“follows perceptual judgment, signifying a transition between uncontrolled thought and controlled reasoning. [...] We guess in an attempt to address the surprising phenomenon that has led to doubt; it is our inchoate attempt to provide an explanation.”

*Second*, *Thagard* [25.22] makes sense in saying that an abductive process involves developing and entertaining inferences toward a law that will be tested via induction, which will then produce inferences about a case. For *Eco* [25.23], however [25.23, p. 203],

“the real problem is not whether to find first the Case or the Law, but rather how to figure out both the Law and the Case *at the same time*, since they are inversely related, tied together by a sort of chiasmus.”

*Third*, while the original meaning of abduction based on Peirce’s work refers to inferences that yield plausible or explanatory hypotheses, *Josephson* and *Josephson’s* [25.24] additional condition of *inferences that yield the best explanation* revises the structure of the original meaning of abduction in the following manner:

Case:  $D$  is a collection of data (facts, observations, givens).

Law:  $H$  explains  $D$  (would, if true, explain  $D$ ).

Strong Claim: No other hypothesis can explain  $D$  as well as  $H$  does.

Result:  $H$  is probably true.

*Paavola* [25.25] notes that while the original and revised versions of abductions share the concern toward generating explanations, they are different in several ways. The original version addresses issues related to the *processes of discovery* and the construction of plausible hypotheses, while the revised version models a nondeductive form of reasoning (except induction) that eventually establishes the true explanation. Across

the differences, it is instructive to keep in mind both *Adler's* “simple, conservative, unifying, and yields the most understanding” conditions for constructing strong abductions [25.26, p. 19] and *El Khachab's* logical goodness conditions that characterize good abductions. That is, they [25.2, p. 164]

“(1) need to be clear, i. e., they need to have distinguishable practical effects; (2) they need to explain available facts; and (3) they need to be liable to future experimental verification.”

*Fourth*, it is important to emphasize that abductions provide explanations or justifications that do not prove. Instead, they provide explanations or justifications that primarily assign causal responsibility in *Josephson's* [25.27, p. 7] sense below.

“Explanations give causes. Explaining something, whether that something is particular or general, gives something else upon which the first thing depends for its existence, or for being the way that it is. [...] It is common in science for an empirical generalization, an observed generality, to be explained by reference to underlying structure and mechanisms.”

### 25.2.2 Induction

Unlike abduction, induction tests a preliminary or an ongoing abduction in order to support a most reasonable law and thus develop a generalization that would both link and unite both the known and projected cases together in a meaningful way. By testing an abductive claim over several cases, induction determines whether the claim is right or wrong. So defined, a correct induction does not produce a new concept that explains (i. e., an *explanatory theory*), which is the primary purpose of abductive processing. Instead, it seeks to show that once the premises hold (i. e., the case/s and the result/s), then the relevant conclusions (i. e., the law) must be true by enumeration (number of observed cases), analogy (i. e., structural or relational similarity of features among cases), or scientific analysis (through actual or mental experiments) [25.28] and thus reflect causal relationships that are expressed in the form of (categorical inductive or universally quantified) generalizations [25.11]. In the case of enumeration, in particular, the goal is not to establish an exhaustive count leading to a precise numerical value, but it is about “producing a certain psychological impression [...] brought about through the laws of association, and creating an expectation of a continuous repetition of the experience” [25.28, p. 184]. In all three contexts of inductive justification, inductive inferences do

not necessarily yield true generalizations. However, “in the long run they approximate to the truth” [25.29, p. 207].

Four important points are worth noting about the relationship between abduction and induction, as follows:

*First*, *El Khachab* points out how both abduction and induction appear to be “unclear” about their “practical effects which are essentially similar” [25.2, p. 166]. However, they are different in terms of “degree”, that is [25.2, p. 166],

“an induction is an inference to a rule; an abduction is an inference to a rule *about* an occurrence, or in Peirce’s own words, *an induction from qualities* [...] Induction is a method of experimental verification leading to the establishment of truth in its long-term application.”

*Second*, abduction is not a requirement for induction. That is, there can be an abduction without induction (i. e., abductive generalizations). Some geometry theorems, for example, do not need inductive verification. In some cases, abduction is framed as conjectures that are used to further explain the development of schemes ([25.30] in the case of fractions). However, it is useful to note the insights of *Pedemonte* [25.9] and *Prusak et al.* [25.31] about the necessity of a structural continuity between an abduction argument process and its corresponding justification in the form of a logical proof. That is, a productive abductive process in whatever modal form (visual, verbal) should simultaneously convey the steps in a deductive proof.

Even in the most naïve and complex cases of inductions (e. g., number patterns with no meaningful context other than the appearance of behaving like objects in some sequence), learners initially tend to produce an abductive claim as a practical embodied coping strategy, that is, as a way of imposing some order or structure that may or may not prove to make sense in the long haul. Euler’s numerical-driven generalization of the infinite series  $\sum_{n=1}^{\infty} \frac{1}{n^2}$  is a good example. He initially established an analogical relationship between two different types of equations (i. e., a polynomial  $P$  of degree  $n$  having  $n$  distinct nonzero roots and a trigonometric equation that can be transformed algebraically into something like  $P$  but with an infinite number of terms). Euler’s abductive claim had him hypothesizing an anticipated solution drawn from similarities between the forms of the two equations. Upon inductively verifying that the initial four terms of the two equations were indeed the same, Euler concluded that [25.17, pp. 17–22]

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6} .$$

Third, another consequence of the preceding discussion involves the so-called *inductive leap*, which involves establishing a generalization from concrete instances to a conclusion that seems to contain more than the instances themselves. On the basis of the characterizations we have assigned to abduction and induction, such a leap is no longer an issue since the leap itself is settled by abduction. Hence, criticisms that in effect cite “hazardous inductive leap” as an argument in relation to erroneous patterning questions such as the one shown in Fig. 25.6 is more appropriately and fundamentally a problem of abduction.

Fourth, neither abduction nor induction can settle the issue of *reasonable of context*. For example, the patterning situation in Fig. 25.7 can have a stipulated abduction and an inductively verified set of outcomes based on an interpreted explicit formula. However, as Parker and Baldrige [25.32] have noted, “there is no reason why the rainfall will continue to be given by that expression, or *any* expression”, which implies that the “question cannot be answered” [25.32, p. 90].

### 25.2.3 Deduction and Deductive Closure

While abduction and induction provide support in constructing or producing a theory, both deduction and deductive closure aim to exhibit necessity. Pace Smith [25.33]: “(R)epeated co-instantiation via induction is not the same as inferential necessity” [25.33, p. 5]. A valid deduction demonstrates a logical implication, that is, it shows how a law and a case as premises or hypotheses together imply a necessary result, conclusion, or consequence. It is a “self-contained process” because the validation process relies on “the existence of well-defined sets” and preserves an already established law, thus, “freeing us from the vagaries and changeability of an external world” [25.10, p. 37].

*A certain pattern begins with 1, 2, 4. If the pattern continues, what is the next number?*

- A. 1
- B. 2
- C. 7
- D. 8

**Fig. 25.6** An example of an erroneous generalization problem

It started to rain. Every hour Sarah checked her rain gauge. She recorded the total rainfall in a table. How much rain would have fallen after  $h$  hours?

Hours	Rainfall
1	0.5 in
2	1 in
3	1.5 in

**Fig. 25.7** An Example of a patterning task with an erroneous context (after [25.32])

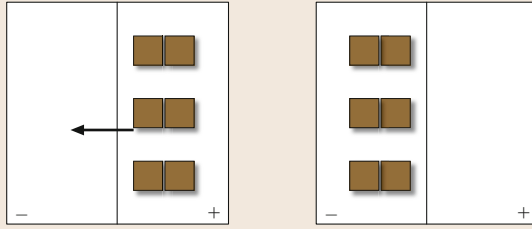
Deductive closure emerges in students’ mathematical thinking and reasoning in at least two ways depending on grade-level expectations, as follows. Among elementary and middle school students, once they (implicitly) form a deduction, they tend to provide an empirical (numerical or visual) structural argument then a formal deductive proof as a form of explanation or justification. For example, Cherrie’s algebraic generalization relative to the pattern in Fig. 25.3 could be expressed in deductive form. When she began to correctly apply her result to any stage in her pattern beyond the known ones, her reasoning entered the deductive closure phase.

Among high school students and older adults, once they formulate a deduction, they tend to provide any of the following types of justification that overlap in some situations: an empirical structural argument; a logical deductive proof; or a mathematical induction proof. Figure 25.8 illustrates how a group of 34 US Algebra 1 middle school students (mean age of 13 years) empirically justified the fact that  $-a \times -b = -(a \times -b)$  by demonstrating a numerical argument following a statement-to-reason template [25.34, pp. 126–130]. Note that when the numbers in the empirical argument shown in Fig. 25.8 are replaced with variables, the argument transforms into a logical deductive proof in which case the steps follow a logical “recycling process” (Duvall, quoted in Pedemonte [25.9, p. 24]), that is, the conclusion of a foregoing step becomes the premise of a succeeding step from beginning to end. Deductive closure for these students occurred when they began to obtain products of integers (and, much later, rational numbers) involving negative factors without providing a justification.

Figure 25.9 shows a mathematical inductive proof of a classic theorem involving the sum of the interior angles in an  $n$ -sided convex polygon that has been drawn from Pedemonte’s [25.9] work with 102 Grade 13 students (ages 16–17 years) in France and in Italy. The “multimodal argumentative process of proof” [25.31, 35] evolved as a result of a structural continuity between a combined abductive-inductive action that was performed on a dynamic geometry tool, which focused on a perceived relationship between the process of constructing nonoverlapping triangles in a polygon and the effects on the resulting interior angle sums, and the accompanying steps that reflected the structure of a mathematical induction proof.



Based on the figure below, Let us illustrate why  $-3 \times 2 = -(3 \times 2)$  using properties of integers.  $-3 \times 2 = -(3 \times 2)$  means pull 3 groups of 2 cubes on the positive region to the negative region



$$\begin{aligned}
 -3 \times 2 &= (-3 \times 2) + 0 \\
 &= (-3 \times 2) + [(3 \times 2) + -(3 \times 2)] \\
 &= [(-3 \times 2) + (3 \times 2)] + -(3 \times 2) \\
 &= [(-3 + 3) \times 2] + -(3 \times 2) \\
 &= 0 + -(3 \times 2) \\
 &= -(3 \times 2)
 \end{aligned}$$

Additive identity property  
 Additive inverse property  
 Associative property  
 Distributive property  
 Additive inverse property  
 Additive identity property

**Fig. 25.8** An empirical structural argument for  $-a \times -b = -(a \times -b)$  (after [25.34])

66. M: If n is equal to 3, f(n) is equal to  $180 \times 1$ ...  
 If n is equal to 4, f(n) is equal to 360, which is equal to  $180 \times 2$   
 67. L: N equal to 5, f(n) is equal to 540, which is equal to  $180 \times 3$ ...  
 68. M: So f(n) is equal to  $180 \times (n-2)$  ...  
 69. L: OK, now we have to understand why ...

70. M: OK... wait!  
 71. L: F(4) is equal to  $180 + f(3)$  because there is one triangle more... so  $180 + 180$ ...  
 72. M: OK, then f(5) is... is  $f(4) + 180$ ... that means that f(n) is equal to  $f(n-1) + 180$   
 73. L: You always add 180 to the previous one  
 74. M: OK we can write  $f(n+1)$  as  $f(n) + 180$ ...

Base  $F(3) = 180^\circ$   
 $F(n+1) = 180^\circ(n-1)$   
 $F(n+1) = F(n) + 180^\circ$   
 It is necessary to add  $180^\circ$  to  $F(n)$  because if we add a side to the polygon, we add a triangle too.  
 The sum of the triangles angles is  $180^\circ$ .  
 So:  
 $F(n+1) = 180^\circ(n-2) + 180^\circ$   
 $F(n+1) = 180^\circ(n-2+1)$   
 $F(n+1) = 180^\circ(n-1)$

**Fig. 25.9** A mathematical inductive proof for the sum of the interior angles in an  $n$ -sided convex polygon (after [25.8, p. 37–38])

The work shown in Fig. 25.10 was also drawn from the same sample of students that participated in Pedemonte's [25.9] study. Unlike Fig. 25.9, the analysis that the students exhibited in Fig. 25.10 shows a structural discontinuity between a combined abductive-inductive action, which primarily focused on the results or out-

comes in a table of values, and steps that might have produced either a valid empirical justification or a logical mathematical induction proof. Deductive closure for these students occurred when they began to obtain the interior angle sum measures of any convex polygon beyond the typical ones.

<p><i>Alice constructs the following table:</i></p> <table border="1" style="width: 100%; border-collapse: collapse; margin-bottom: 10px;"> <thead> <tr> <th style="padding: 5px;">Sides</th> <th style="padding: 5px;">Sum (Angles)</th> </tr> </thead> <tbody> <tr> <td style="padding: 5px; text-align: center;">3</td> <td style="padding: 5px;">180°</td> </tr> <tr> <td style="padding: 5px; text-align: center;">4</td> <td style="padding: 5px;">360°    180° × 2</td> </tr> <tr> <td style="padding: 5px; text-align: center;">5</td> <td style="padding: 5px;">540°    180° × 3</td> </tr> <tr> <td style="padding: 5px; text-align: center;">6</td> <td style="padding: 5px;">720°    180° × 4</td> </tr> </tbody> </table> <p>29. A: So the rule is probably <math>180 \times (n-2)</math> for an <math>n</math>-sided polygon</p> <p>30. L: Yes... <math>n</math> is the number of sides</p>		Sides	Sum (Angles)	3	180°	4	360°    180° × 2	5	540°    180° × 3	6	720°    180° × 4
Sides	Sum (Angles)										
3	180°										
4	360°    180° × 2										
5	540°    180° × 3										
6	720°    180° × 4										
<p>Base for <math>n = 3</math>  <math>180^\circ(3-2) = 180^\circ</math></p> <p>Step            Hp: <math>180^\circ(n-2)</math>            Ts: <math>180^\circ(n-1)</math></p> <p><math>S(n) = 180^\circ (n-2) = 180n - 360</math>  <math>S(n+1) = 180^\circ (n+1) - 360 = 180n + 180 - 360 = n + 1 - 2 = n - 1</math> Th            We have proved the thesis by a mathematical induction</p>											

**Fig. 25.10** Example of an erroneous mathematical inductive argument for the sum of the interior angles in an  $n$ -sided convex polygon (after [25.8, p. 36])

## 25.3 Abduction in Math and Science Education

A nonexhaustive survey of recent published studies dealing with abduction in mathematical and scientific thinking and learning yields two interesting findings, as follows.

### 25.3.1 Different Kinds of Abduction

Drawing on *Eco's* [25.23] work, *Pedemonte* and *Reid* [25.36] provided instances in which traditional 15–17-year-old Grades 12 and 13 students in France and Italy modeled overcoded, undercoded, and creative abductions in the context of proving statements in mathematics. For *Pedemonte* and *Reid*, abduction comes before deduction. Some students in their study generated overcoded abductions, which involve using a single rule to generate a case, while others produced undercoded abductions, which involve choosing from among several different rules to establish a case. Overcoded and undercoded abductions for *Magnani* [25.11] exemplify instances of selective abductions because the basic task involves selecting one rule that would make sense, which, hopefully, would also yield the best explanation. Medical diagnosis, for instance, employs selective abductions [25.11]. In cases when no such rules exist, students who develop new rules of their own yield what *Eco* [25.23] refers to as creative abductions, which also account for “the growth of scientific knowledge” [25.11, p. 511]. *Pedemonte* and *Reid* have noted that students are usually able to construct a deductive proof in cases involving overcoded abductions due to the limited number of possible sets of rules to choose from. Furthermore, they tend to experience considerable difficulties in cases that involve undercoded and creative abductions since they have to deal with

“irrelevant information in the argumentation process, thus confusing, and creating disorder” in their processing [25.36, p. 302]. An additional dilemma that students have with creative abductions is the need to justify them prior to using them as rules in a proof process. “Consequently”, *Pedemonte* and *Reid* write [25.36, p. 302],

“it seems that there is not a simple link between the use of abduction in argumentation and constructing a deductive proof. Both the claim that abduction is an obstacle to proof and the claim that abduction is a support, if considered in a general sense, are oversimplifications. Some kinds of abductions, in some context may make the elements required for the deductions used in a proof more accessible. Some are probably less dangerous to use and can make the construction of a proof easier to get to because they could make easier to find and to select the theorem and the theory necessary to produce a proof. However, other kinds of abductions present genuine obstacles to constructing the proof. This suggests that teaching approaches that involve students conjecturing in a problem solving process prior to proving have potential, but great care must be taken that the abductions expected of the students do not become obstacles to their later proving.”

Aside from selective and creative abductions, *Magnani* [25.11] pointed out the significance of theoretical and manipulative abductions in other aspects of everyday and scientific work that involve creative processing. Theoretical abductions involve the use of logical, verbal or symbolic, and model-based (e.g., diagrams and pic-

tures) processing in reasoning. While valuable, they are unable to account for other possible types of explanations (e.g., statistical reasoning, which is probabilistic; sufficient explanations; high-level kinds and types of creative and model-based abductions; etc.). Manipulative abductions emerge in cases that involve “thinking and discovering through doing”, where actions are pivotal in enabling learners to model and develop insights simultaneously leading to the construction of creative or selective abductions. They operate beyond the usual purpose of experiments and create “extra-theoretical behaviors” that [25.11, p. 517]

“create communicable accounts of new experiences in order to integrate them into previously existing systems of experimental and linguistic (theoretical) practices. The existence of this kind of extra-theoretical cognitive behavior is also testified by the many everyday situations in which humans are perfectly able to perform very efficacious (and habitual) tasks without the immediate possibility of realizing their conceptual explanation.”

Typical accounts of conceptual change processes in science tend to highlight theoretical abductions, however [25.11, p. 519],

“a large part of these processes are instead to due practical and *external* manipulations of some kind, prerequisite to the subsequent work of theoretical arrangement and knowledge creation.”

Manipulative abductions may also emerge in learning situations that provide “conceptual and theoretical details to already automatized manipulative executions” in which case either teacher or learner [25.11, p. 519]

“does not discover anything new from the point of view of the objective knowledge about the involved skill, however, we can say that his conceptual awareness is new from the local perspective of his individuality.”

For example, *Rivera* [25.37] provides a narrative account of US third-grade Mark’s evolving understanding of the long division algorithm involving multidigit whole numbers by a single-digit whole number. Mark’s initial visual representation processing of (sharing-partitive) division (Fig. 25.11) employed the use of place value-driven squares, sticks, and circles. In the case of the division task  $126 \div 6$ , when he could not divide a single (hundreds) box into six (equal) groups, he recorded it as a 0. He then ungrouped the box into ten sticks, regrouped the sticks together, divided the

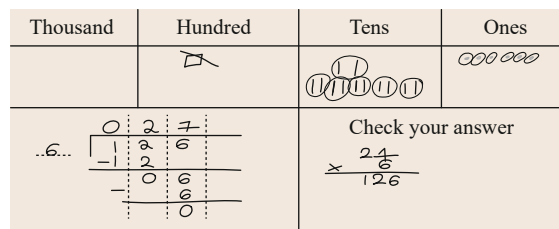


Fig. 25.11 Mark’s initial visual processing of  $126 \div 6$

sticks into six groups, recorded accordingly, and so on until he completed the division process for all subcollections. His numerical recording in Fig. 25.11 also captured every step in his sequence of visual actions. Results of consistent visual processing enabled him to shift his attention away from the visual form and toward the rule for division, which was accompanied by two remarkable changes in his numerical processing. In Fig. 25.12, he performed division on each digit in the dividend from left to right with the superscripts indicating partial remainders that had to be ungrouped and regrouped. In Fig. 25.13, he made another subtle creative revision that remained consistent with his earlier work and experiences. When he was asked to explain his division method, Mark claimed that “it’s like how we do adding and subtracting with regrouping, we’re just doing it with division”. Mark’s manipulative abductive processing for division involving whole numbers necessitated a dynamic experience in which “a first *rough* and concrete experience” [25.11, p. 519] of the process enabled him to eventually develop a version of the long division process that “unfolded in real time” via thinking through doing.

### 25.3.2 Abduction in Mathematical Relationships

A study by *Arzarello* and *Sabena* [25.38] illustrates the important role of abduction in constructing mathematical relationships involving different signs. Signs pertain to the triad of signifier, signified, and an individual learner’s mental construct that enables the linking between signifier and signified possible. Arzarello and Sabena underscore their students’ use of semiotic and theoretic control when they argued and proved statements in mathematics. Semiotic control involves choosing and implementing particular semiotic resources (e.g., graphs, tables, equations, etc.) when they manipulate and interpret signs (i.e., type-1 semiotic action), while theoretic control involves choosing and implementing appropriate theories (e.g., Euclidean theorems) or parts of those theories and related conceptions when they “elaborate an argument or a proof” (i.e., type-3 semiotic action; [25.38, p. 191]). Between type-1 and

7. Eight-hundred thirty-seven divided by three $\begin{array}{r} 8237 \\ \div 3 \\ \hline 279 \end{array}$ $\begin{array}{r} 279 \\ \times 3 \\ \hline 837 \end{array}$	8. Eight-hundred fifty-two divided by three $\begin{array}{r} 8252 \\ \div 3 \\ \hline 284 \end{array}$ $\begin{array}{r} 284 \\ \times 3 \\ \hline 852 \end{array}$
--	---

**Fig. 25.12** Mark's initial numerical division processing

11. Eighty-four divided by 7 $\begin{array}{r} 84 \\ \div 7 \\ \hline 12 \end{array}$ $\begin{array}{r} 12 \\ \times 7 \\ \hline 84 \end{array}$	8. Nine-hundred eighty-four divided by 8 $\begin{array}{r} 984 \\ \div 8 \\ \hline 113 \end{array}$
---	--

**Fig. 25.13** Mark's manipulative abductive processing of the numerical methods shown in Figs. 25.11 and 25.12

type-3 semiotic action is a type-2 semiotic action that involves using abduction to identify relationships between signs and assessing the arguments. Based on their qualitative work with Grade 9 students, such [25.38, p. 202]

“relationships between signs are examined and checked with redundant local arguments, and (economic, explanatory, and testable) hypotheses are detected and made explicit by means of abductions.”

Furthermore, they note how [25.38, p. 204]:

“abduction has an important role at this point. There is an evolution from a phase where the attention is mainly on the given signs, towards a phase where the logical-theoretical organization of the argument becomes the center of the activities and evolves from abductive to deductive and more formal structures. [...] Such an evolution implies a passage from actions of type 1 to actions of type 2 and then 3, and a shift of control by the student, i. e., passing from actions guided by semiotic control to actions guided by theoretical control. [...] Passing from type 1- to type 3-semiotic actions means an evolution from the data to the truth because of theoretical reasons. It is exactly this distinction that makes the difference between [...] a *substantial argument* and an *analytical argument*, which is a mathematical proof.”

Arzarello and Sabena's study foregrounds the role of abduction in inferential processing and documents how a shift from abduction to deduction is likely to occur when students' mathematical thinking shifts in

focus from the semiotic to theoretical, respectively. Studies by *Pedemonte* and colleagues [25.9, 36, 39] and *Boero* and colleagues [25.40, 41] also note the same findings in both algebra and geometry contexts. Across such studies we note how abduction is conceptualized in terms of its complex relationships with induction and deduction. Other studies do not deliberately focus on such shifts and relationships, making it difficult for students to see the value of engaging in abductive processing in the first place. For example, *Watson* and *Shipman* [25.42] documented the classroom event that happened in a Year 9 class of 13–14 year-old students in the UK that investigated the following task: Find a way to multiply pairs of numbers of the form  $a + \sqrt{b}$  that results in integer products. While the emphasis of their study focused on learning through exemplification by using special examples to help students develop meaningful plausible structures, it seems that the abductive process for them became a matter of conjecturing relationships based on their experiences with their constructed examples. But certainly there is more to abductive processing than merely generating conjectures, as follows.

Several studies have suggested inferential model systems that show relationships between and among abduction, induction, and deduction. *Addis* and *Gooding* [25.10], for example, illustrate how the iterative cycle of

abduction (generation) → deduction (prediction) →  
induction (validation) → abduction

might work in the formation of consensus from beliefs. *Radford's* [25.43] architecture of algebraic pattern generalizations emphasizes a tight link between abduction and deduction, that is, hypothetico-deduction, in the fol-

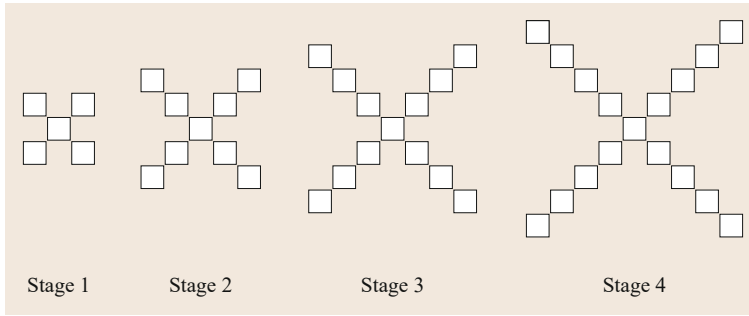


Fig. 25.14 Cross-squares pattern

lowing manner:

abduction (from particulars  $p_1, p_2, \dots, p_k$   
to noticing a commonality  $C$ )  
→ transforming the abduction  
(from noticing  $C$  to making  $C$  a hypothesis)  
→ deduction (from hypothesis  $C$   
to producing the expression of  $p_n$ ).

The studies conducted by Rivera [25.44] with groups of elementary (i. e., first through third grade) and middle school (i. e., sixth through eighth grade) students in the US on similar pattern generalization tasks capture two different inferential structures. Prior to a teaching intervention that involved using multiplicative thinking to establish pattern rules, both elementary and middle school student groups employed the same inferential structure of

abduction → induction → deduction  
→ deductive closure

that enabled them to generalize (correctly and incorrectly). The abductive phase in such a structure tended to be instinctual and iconic- or perceptual-driven. After the teaching intervention, however, Rivera observed that while the elementary student groups continued to model the same inferential structure in pattern gener-

alizing, the middle school student groups skipped the induction phase and instead exhibited the following structure:

abduction and deduction → deductive closure .

Abduction in this phase was combined with deduction and thus became structured and inferential as a consequence of their ability to express generalizations in multiplicative form. For example, when the pattern generalization task shown in Fig. 25.14 was presented to both elementary and middle school student groups after the teaching experiment, sixth-grade student Tamara initially abducted the recursive relation  $+4$ , which enabled her to deduce the explicit rule  $s = n \times 4 + 1$ . She then used her combined abductive-deductive inference to perform deductive closure, in which case she induced the given stages and predicted the correct number of squares for any stage in her pattern. Tamara's empirical justification of her explicit rule for the total number of squares  $s$  involves seeing a fixed square and four copies of the same *leg* that grew according to the stage number  $n$ . In the case of third-grade Anna, her multiplicative-driven abductive processing enabled her to both construct and justify the same explicit rule that Tamara established for the pattern. However, she needed to express her answers inductively, as follows

$$4 \times 1 + 1, 4 \times 2 + 1, 4 \times 3 + 1, \dots, 4 \times 100 + 1, \dots$$

## 25.4 Enacting Abductive Action in Mathematical Contexts

We close this chapter by providing four suggestions for assisting students to enact meaningful, structured, and productive abduction action. Together the suggestions target central features in abductive cognition, that is, thinking, reasoning, processing, and disposition. Empirical research in mathematics education along these features is needed to fully assess the extent and impact of their power in shaping mathematical knowledge construction.

### 25.4.1 Cultivate Abductively-Infused Guesses with Deduction

Students will benefit from knowing how to generate *new* guesses and conjectures that can explain a problem and occur “within the wider scope of the process of inquiry” [25.20, p. 116]. That is, while abductions certainly emerge from perceptual judgments, in actual practice the more useful ones are usually constrained

and logical as a consequence of knowing the problem context and being “compounds of deductions from general rules” (i. e., hypothetico-deductivist) that individual knowers are already familiar with (Peirce, quoted in [25.20, p. 119]). *Tschaepé* writes, “(w)e guess in an attempt to address the surprising phenomenon that has led to doubt; it is our inchoate attempt to provide an explanation” [25.20, p. 118]. Viewed in this sense [25.20, p. 122],

“[a]bduction is a logical operation, and guess is logical insofar as it is a type of reasoning by which an explanation of a surprising phenomenon is first created, selected, or dismissed [...] Guessing is the creative component of abductive inference in which a new idea is first suggested through reasoning.”

### 25.4.2 Support Logically-Good Abductive Reasoning

Students will benefit from knowing how to develop abductions that are logically good, that is, they are: clear (i. e., can be confirmed or disconfirmed); can explain the facts; are capable of being tested and verified; and can lead to true explanations that establish “sustainable belief-habits” [25.2, p. 163]. Such explanations may be new and may emerge from guesses and instincts, but, *Khachab* writes [25.2, pp. 171–172],

“logical goodness is the reason for *abduction*, under its diverse meanings. No matter *how* abduction *actually* generates *new* ideas – whether it is abductive inference, strategic inference, instinctive insight, etc. – its purpose is, ultimately, to provide true explanatory hypotheses for inquiry. And, in this regard, *new* hypotheses should always be evaluated in reference to their goodness.”

### 25.4.3 Foster the Development of Strategic Rules in Abductive Processing

*Paavola* [25.12] distinguishes between definitory and strategic rules. While definitory rules focus on logic and logical relationships, strategic rules pertain to “goal-directed activity, where the ability to anticipate things, and to assess or choose between different possibilities, are important” [25.12, p. 270]. Thus, abductive strategies produce justifications for given explanatory hypotheses, including justifications for “why there cannot be any further explanation” [25.12, p. 271]. Hence, all generated abductive inferences conveyed in the form of discoveries provide an analysis or explanation of

the underlying conceptual issues and are not merely reflective of mechanical recipes or algorithms for generating ideas and discoveries. Furthermore, the analysis or explanation should present “a viable way of solving a particular problem and that it works more generally (and not only in relationship to one, particular anomalous phenomenon)” [25.12, p. 273] and fit the “constraints and clues that are involved in the problem situation in question” [25.12, p. 274].

### 25.4.4 Encourage an Abductive Knowledge-Seeking Disposition

*Sintonen*’s [25.45] interrogative model of inquiry that employs an explicit logic of questions demonstrates the significance of using certain strategic principles and why-questions as starting points in abductive processing. Questions as well as answers drive discoveries and the scientific process. Questions, especially, “pick out something salient that requires special attention, and that it also gives heuristic power and guidance in the search for answers” [25.45, p. 250]. Furthermore, [25.45, p. 263],

“principal questions are often explanation-seeking in nature and arise when an agent tries to fit new phenomena to his or her already existing knowledge. Advancement of inquiry can be captured by examining a chain of questions generated. By finding answers to subordinate questions, an agent approaches step by step toward answering the big initial question, and thus changes his or her epistemic situation.”

Students will benefit from situations and circumstances that engage them in a knowledge-seeking game in which they “subject a source of information [...] to a series of strategically organized questions. This Sherlock Holmes method therefore is at the heart of abductive reasoning” [25.45, p. 254]. Furthermore, the interrogative model allows conclusions (i. e., answers) to emerge. “For abductive tasks”, *Sintonen* writes [25.45, p. 256],

“the goal must be understanding and not just knowledge. A rational inquirer who wants to know why and not only that something is the case must, after hearing the answer, be in the position to say *Now I know (or rather understand) why the (singular or general) fact obtains*. Obviously this condition is fulfilled only if she or he knows enough of the background to be able to insert the offered piece of information into a coherent explanatory account.”

**Acknowledgments.** The research that is reported in this chapter has been funded by the National Science Foundation under Grant Number DRL 0448649. All the

views and opinions expressed in this report are solely the author's responsibility and do not necessarily reflect the views of the foundation.

## References

- 25.1 J. Shotter: Bateson, double description, Todes, and embodiment: Preparing activities and their relation to abduction, *J. Theory Soc. Behav.* **39**(2), 219–245 (2009)
- 25.2 C. El Khachab: The logical goodness of abduction in C. S. Peirce's thought, *Trans. Charles S. Peirce Soc.* **49**(2), 157–177 (2013)
- 25.3 F. Rivera: Changing the face of arithmetic: Teaching children algebra, *Teach. Child. Math.* **12**(6), 306–311 (2006)
- 25.4 F. Rivera: Visual templates in pattern generalization activity, *Educ. Stud. Math.* **73**, 297–328 (2010)
- 25.5 G. Bateson: *Mind in Nature: A Necessary Unity* (Fontana/Collins, London 1979)
- 25.6 D. Holton, K. Stacey, G. FitzSimons: Reasoning: A dog's tale, *Aust. Math. Teach.* **68**(3), 22–26 (2012)
- 25.7 C. Peirce: *Collected Papers of Charles Saunders Peirce*, Vol. 5 (Harvard Univ. Press, Cambridge 1934)
- 25.8 J. Mason, M. Stephens, A. Watson: Appreciating mathematical structures for all, *Math. Educ. Res. J.* **21**(2), 10–32 (2009)
- 25.9 B. Pedemonte: How can the relationship between argumentation and proof be analyzed?, *Educ. Stud. Math.* **66**, 23–41 (2008)
- 25.10 T. Addis, D. Gooding: Simulation methods for an abductive system in science, *Found. Sci.* **13**, 37–52 (2008)
- 25.11 L. Magnani: Conjectures and manipulations: Computational modeling and the extra-theoretical dimension of scientific discovery, *Minds Mach.* **14**, 507–537 (2004)
- 25.12 S. Paavola: Abduction as a logic and methodology of discovery: The importance of strategies, *Found. Sci.* **9**, 267–283 (2004)
- 25.13 A. Heeffer: Learning concepts through the history of mathematics: The case of symbolic algebra. In: *Philosophical Dimensions in Mathematics Education*, ed. by K. Francois, J.P. Van Bendegem (Springer, Dordrecht 2010) pp. 83–103
- 25.14 U. Goswami: Inductive and deductive reasoning. In: *The Wiley-Blackwell Handbook of Childhood Cognitive Development*, ed. by U. Goswami (Wiley-Blackwell, Malden 2011) pp. 399–419
- 25.15 C. Peirce: *Collected Papers of Charles Saunders Peirce*, Vol. 1/2 (Belnap Press of Harvard Univ. Press, Cambridge 1960)
- 25.16 B. Pillow, R. Pearson, M. Hecht, A. Bremer: Children's and adults' judgments of the certainty of deductive inference, inductive inferences, and guesses, *J. Genet. Epistemol.* **171**(3), 203–217 (2010)
- 25.17 G. Polya: *Induction and Analogy in Mathematics*, Mathematics and Plausible Reasoning, Vol. 1 (Princeton Univ. Press, Princeton 1973)
- 25.18 B. Haig: Precis of "an abductive theory of scientific method, *J. Clin. Psychol.* **64**(9), 1019–1022 (2008)
- 25.19 S. Paavola: Diagrams, iconicity, and abductive discovery, *Semiotica* **186**(1/4), 297–314 (2011)
- 25.20 M. Tschaeppe: Guessing and abduction, *Trans. Charles S. Peirce Soc.* **50**(1), 115–138 (2014)
- 25.21 G.-J. Kruijff: Peirce's late theory of abduction: A comprehensive account, *Semiotica* **153**(1/4), 431–454 (2005)
- 25.22 P. Thagard: Semiosis and hypothetic inference in C. S. Peirce, *Versus Quaderni Di Studi Semiotici* **19/20**, 163–172 (1978)
- 25.23 U. Eco: Horns, hooves, insteps: Some hypotheses on three types of abduction. In: *The Sign of Three: Dupin, Holmes, Peirce*, ed. by U. Eco, T. Sebeok (Indiana Univ. Press, Bloomington 1983) pp. 198–220
- 25.24 J. Josephson, S. Josephson: *Abductive Inference: Computation, Philosophy, Technology* (Cambridge University Press, New York 1994)
- 25.25 S. Paavola: Hansonian and Harmanian abduction as models of discovery, *Int. Stud. Philos. Sci.* **20**(1), 93–108 (2006)
- 25.26 J. Adler: Introduction: Philosophical foundations. In: *Reasoning: Studies of Human Inference and Its Foundations*, ed. by J. Adler, L. Rips (Cambridge Univ. Press, Cambridge 2008) pp. 1–34
- 25.27 J. Josephson: Smart inductive generalizations are abductions. In: *Abduction and Induction: Essays on Their Relation and Integration*, ed. by P. Flach, A. Kakas (Kluwer, Dordrecht 2000) pp. 31–44
- 25.28 J. Hibben: *Logic: Deductive and Inductive* (Charles Scribner's Sons, New York 1905)
- 25.29 C. Peirce: Grounds of validity of the laws of logic: Further consequences of four incapacities, *J. Specul. Philos.* **2**, 193–208 (1869)
- 25.30 A. Norton: Josh's operational conjectures: Abductions of a splitting operation and the construction of new fractional schemes, *J. Res. Math. Educ.* **39**(4), 401–430 (2008)
- 25.31 N. Prusak, R. Hershkowitz, B. Schwarz: From visual reasoning to logical necessity through argumentative design, *Educ. Stud. Math.* **74**, 185–205 (2012)
- 25.32 T. Parker, S. Baldrige: *Elementary Mathematics for Teachers* (Sefton-Ash Publishing, Okemos 2004)
- 25.33 L. Smith: *Reasoning by Mathematical Induction in Children's Arithmetic* (Elsevier Science Ltd., Oxford 2002)
- 25.34 F. Rivera: *Toward a Visually-Oriented School Mathematics Curriculum: Research, Theory, Practice, and Issues* (Springer, New York 2011)
- 25.35 F. Arzarello: The proof in the 20th century. In: *Theorems in Schools: From History, Epistemology, and Cognition in Classroom Practices*, ed. by P. Boero (Sense Publishers, Rotterdam 2006) pp. 43–64

- 25.36 B. Pedemonte, D. Reid: The role of abduction in proving processes, *Educ. Stud. Math.* **76**, 281–303 (2011)
- 25.37 F. Rivera: From math drawings to algorithms: Emergence of whole number operations in children, *ZDM* **46**(1), 59–77 (2014)
- 25.38 F. Arzarello, C. Sabena: Semiotic and theoretic control in argumentation and proof activities, *Educ. Stud. Math.* **77**, 189–206 (2011)
- 25.39 M. Martinez, B. Pedemonte: Relationship between inductive arithmetic argumentation and deductive algebraic proof, *Educ. Stud. Math.* **86**, 125–149 (2014)
- 25.40 P. Boero, R. Garuti, M. Mariotti: Some dynamic mental processes underlying producing and proving conjectures, *Proc. 20th Conf. Int. Group Psychol. Math. Educ.*, Vol. 2, ed. by L. Puig, A. Gutierrez (IGPME, Valencia 1996) pp. 121–128
- 25.41 P. Boero, N. Douek, F. Morselli, B. Pedemonte: Argumentation and proof: A contribution to theoretical perspectives and their classroom implementation, *Proc. 34th Conf. Int. Group Psychol. Math. Educ.*, Vol. 1, ed. by M. Pinto, T. Kawasaki (IGPME, Belo Horizonte 2010) pp. 179–204
- 25.42 A. Watson, S. Shipman: Using learner generated examples to introduce new concepts, *Educ. Stud. Math.* **69**, 97–109 (2008)
- 25.43 L. Radford: Iconicity and contraction: A semiotic investigation of forms of algebraic generalizations of patterns in different contexts, *ZDM* **40**, 83–96 (2008)
- 25.44 F. Rivera: *Teaching and Learning Patterns in School Mathematics: Psychological and Pedagogical Considerations* (Springer, New York 2013)
- 25.45 M. Sintonen: Reasoning to hypotheses: Where do questions come?, *Found. Sci.* **9**, 249–266 (2004)



---

# Model-Based Reasoning

## Part F

### Part F Model-Based Reasoning in Cognitive Science

Ed. by Athanassios Raftopoulos

- 26 Vision, Thinking, and Model-Based Inferences**  
Athanassios Raftopoulos, Nicosia, Cyprus
- 27 Diagrammatic Reasoning**  
William Bechtel, La Jolla, USA
- 28 Embodied Mental Imagery in Cognitive Robots**  
Alessandro Di Nuovo, Enna, Italy  
Davide Marocco, Plymouth, UK  
Santo Di Nuovo, Catania, Italy  
Angelo Cangelosi, Plymouth, UK
- 29 Dynamical Models of Cognition**  
Mary Ann Metzger, Baltimore, USA
- 30 Complex versus Complicated Models of Cognition**  
Ruud J.R. Den Hartigh, Groningen, The Netherlands  
Ralf F.A. Cox, Groningen, The Netherlands  
Paul L.C. Van Geert, Groningen, The Netherlands
- 31 From Neural Circuitry to Mechanistic Model-Based Reasoning**  
Jonathan Waskan, Urbana, USA

Model-based reasoning refers to the kinds of inferences performed on the basis of a knowledge-context that guides them. This context constitutes a model of a domain of reality, that is, an approximative and simplifying to various degrees representation of the factors that underlie, and the interrelations that govern, the behavior of the entities in this domain. Model-based reasoning is ubiquitous in the human (and not only the human) brain. Various studies have shown that most likely we do not draw inferences by applying some abstract, formal rules; instead inference rules are applied within concrete-knowledge contexts that determine which rules should be used and when.

Model-based reasoning is not limited to the cognitive functions of the brain but it is likely that it extends to perceptual functions that retrieve information from the environment. [Chapter 26](#) defends the view that the processes of visual perception constitute a case of model-based reasoning. It discusses, first, the problem of whether vision involves model-based inferences and, if so, what kind. Secondly, it discusses the problem of the nature of the context that guides visual inferences. It finally addresses the broader problem of the relation between visual processing and thinking; various modes of inferences, the most predominant conceptions about visual perception, the stages of visual processing, the problem of the cognitive penetrability of perception, and the logical status of the processes involved in all stages of visual processing are discussed and assessed.

Reasoning is usually considered to consist of actions that occur exclusively within the brain of the agent that reasons. Reasoning is a mental activity in which various inference rules are applied to mentally represented sentences. This is not always true, however. On many occasions agents use external representations to enhance their inferential capabilities by overcoming limitations in working memory capacities, by simplifying the problem space, etc. In the second chapter, [Bechtel](#) argues that humans often reason by constructing, manipulating, and responding to external representations, whether the reasoning be deductive, abductive, or inductive. These representations are not only linguistic expressions (symbols on a piece of paper, for instance) but also include diagrams. Although diagrams are used in everyday reasoning, they are particularly important in science; diagrams, for example, figure in the processes through which scientists analyze data and construct their explanations. In [Chap. 27](#) [Bechtel](#) discusses what is known about how people, including scientists, reason with diagrams.

Another consequence of the assumption that reasoning takes place within the human mind is that reasoning

depends only on the mental properties of the reasoning agent and is independent of the agent's body. More generally, cognition is limited within the mind of the cognitive agent. This assumption has been recently challenged on many grounds and the argument has been made that reasoning is embodied in that it constitutively and not merely causally involves the body of the reasoning agent. In this vein, [Chap. 28](#) focuses on the role of the concept of mental imagery as a fundamental cognitive capability that enhances the performance of cognitive robots. The authors discuss the embodied imagery mechanisms applied to build artificial cognitive models of motor imagery and mental simulation to control complex behaviors of humanoid platforms that represent the artificial body.

If reasoning is model based, the reasoning agent draws from a variety of sources in order to choose the more salient and useful rules in a particular problem context, to choose the information that will be brought to bear on the problem at hand, to determine how to update her knowledge basis in view of the outputs of the rule, etc. The complexity of the task paves the way for dynamic approaches to cognition, as they are better suited to handle complexities of this magnitude and explain animals' intelligent behavior more adequately. Dynamical models of cognition put emphasis on time and complexity, both of which relate context to behavior. In [Chap. 29](#), [Metzger](#) argues that temporal processes allow memory, feedback, the effects of non-linear recursion, and the generation of expectation to be brought to bear on cognitive activities, whereas complexity allows stable patterns of coordination to emerge from the interaction of sub-processes. [Metzger](#) reviews several models of cognition, and their dynamical features. She focuses on the manner in which each model deals with time and complexity, thought, and action.

Dynamical models could also be used to model the ways that humans continuously adapt their behavior to changes in their environment, and the way their cognitive abilities continuously develop over time. In [Chap. 30](#), [P. van Geert](#), [R. den Hartigh](#), and [R. Cox](#) argue that an important question for psychologists in this direction has been the discovery of the (cognitive) mechanism that underlies the control of human behavior in real time, as well as the process of cognitive development in the long term. Their chapter discusses two kinds of general approaches, namely, the reductionist approach and the complex dynamic systems (CDS) approach. The reductionist approach, on the one hand, assumes that separate components, such as brain areas or cognitive processing mechanisms, are the main determinants of behavior and development, by processing (and responding to) specific environmental inputs.

---

The CDS approach, on the other hand, assumes that cognition, and, hence, the control of behavior and development, are distributed over the brain, the body, and the environment, all three continuously interacting over time.

Thus, dynamic system theory proposes ways in which embodied cognition (that is, the view that cognitive processes constitutively involve the body) and extended cognition (that is, the view that cognitive processes constitutively involves the environment and, in this sense, are extended to the world breaking the boundaries of the mind/brain) could be brought together with the received view that cognition is restricted to what happens within the boundaries of the brain, and provide a more adequate account of animal cognition. To substantiate this claim, the authors compare the two approaches with respect to their assumptions, research strategies, and analyses. Furthermore, they discuss the extent to which current research data in the cognitive domain can be explained by the two

different approaches. They conclude that the CDS approach provides the most plausible approach to cognition.

In *Chap. 31*, *Waskan* argues that model-based reasoning in science is often carried out in an attempt to understand the kinds of mechanical interactions that might give rise to particular occurrences. Scientists do that by constructing and using mental models that are like scale models in crucial respects. Behavioral evidence points to the existence of these mental models, but the neural plausibility of this hypothesis is still questioned. *Waskan* provides an overview of the psychological literature on mental models of mechanisms, focusing on the problem of how representations that share the distinctive features of scale models might be realized by neural machinations. He argues that lessons brought together from the computational simulation of mechanisms and from neurological research on mental maps in rats, could be applied to explain how neurophysiological processes might realize mental models.

## 26. Vision, Thinking, and Model-Based Inferences

Athanasios Raftopoulos

Model-based reasoning refers to the sorts of inferences performed on the basis of a knowledge context that guides them. This context constitutes a model of a domain of reality, that is, an approximative and simplifying to various degrees representation of the factors that underlie, and the interrelations that govern, the behavior of this domain.

This chapter addresses both the problem of whether vision involves model-based inferences and, if yes, of what kind; and the problem of the nature of the context that acts as the model guiding visual inferences. It also addresses the broader problem of the relation between visual processing and thinking. To this end, the various modes of inferences, the most predominant conceptions about visual perception, the stages of visual processing, the problem of the cognitive penetrability of perception, and the logical status of the processes involved in all stages of visual processing will be discussed and assessed.

The goal of this chapter is, on the one hand, to provide the reader with an overview of the main broad problems that are currently debated in philosophy, cognitive science, and visual science, and, on the other hand, to equip them with the knowledge necessary to allow them to follow and assess current discussions on the nature of visual processes, and their relation to thinking and cognition in general.

26.1	<b>Inference and Its Modes</b> .....	576
26.2	<b>Theories of Vision</b> .....	577
26.2.1	Constructivism .....	577
26.2.2	Theory of Direct Vision or Ecological Theory of Visual Perception .....	580
26.2.3	Predictive Visual Brain: Vision and Action .....	581
26.3	<b>Stages of Visual Processing</b> .....	585
26.3.1	Early Vision .....	585
26.3.2	Late Vision .....	586
26.4	<b>Cognitive Penetrability of Perception and the Relation Between Early Vision and Thinking</b> .....	588
26.4.1	The Operational Constraints in Visual Processing .....	589
26.4.2	Perceptual Learning .....	590
26.5	<b>Late Vision, Inferences, and Thinking</b> .....	591
26.5.1	Late Vision, Hypothesis Testing, and Inference .....	593
26.5.2	Late Vision and Discursive Understanding .....	594
26.6	<b>Concluding Discussion</b> .....	596
26.A	<b>Appendix: Forms of Inferences</b> .....	597
26.A.1	Deduction .....	597
26.A.2	Induction .....	597
26.A.3	Abduction or Inference to the Best Explanation .....	597
26.A.4	Differences Between the Modes of Inference .....	598
26.B	<b>Appendix: Constructivism</b> .....	598
26.C	<b>Appendix: Bayes' Theorem and Some of Its Epistemological Aspects</b> .....	600
26.D	<b>Appendix: Modal and Amodal Completion or Perception</b> .....	600
26.E	<b>Appendix: Operational Constraints in Visual Processing</b> .....	601
	<b>References</b> .....	602

*Helmholtz* [26.1] famously maintained that perception is a form of inference; the brain uses probabilistic knowledge-driven inferences to induce the causes of the sensory input from this input, that is, to extract from the bodily effects of the light emanating from the objects in a visual scene as it impinges on our transducers the various aspects of the world that cause the input. The brain integrates computationally the retinal properties of the image of an object with other relevant sources of information to determine the object's intrinsic properties. *Rock* [26.2] claimed that the perceptual system combines inferential information to form the percept. From visual angle and distance information, for example, the perceptual system infers and perceives size. This inference may be automatic and outside the authority of the viewer who does not have control over it, but is an inference nevertheless.

Similarly, *Spelke* [26.3] suggests "perceiving objects may be more akin to thinking about the physical world than to sensing the immediate environment". The reason is that the perceptual system, to solve the underdetermination problem of both the distal object from the retinal image and of the percept from the retinal image, employs a set of object principles (the Spelke principles) that reflect the geometry and the physics of our environment. Since the principles can be thought of as some form of knowledge about the world, perception engages in inferential processes from some pieces of worldly knowledge and visual information to the percept, that is, the object of our ordinary visual encounters with the world.

Recently *Clark* [26.4] argued that:

"To perceive the world just is to use what you know to explain away the sensory signal across multiple spatial and temporal scales. The process of perception is thus inseparable from rational (broadly Bayesian) processes of belief fixation [...] As thought, sensing, and movement here unfold, we discover no stable or well-specified interface or interfaces between cognition and perception. Believing and perceiving, although conceptually distinct, emerge as deeply mechanically intertwined."

The aim of this conglomeration of faculties that constitute perception is, therefore, to enable perceivers to respond, modify their responses, and eventually adapt their responses as they interact with the environment so as to tune themselves to the environment in such a way that this interaction be successful; success in such an endeavor relies on inferring correctly (or nearly so) the nature of the source of the incoming signal from the signal itself.

In all these views, the visual system constructs the percept in the way thinking constructs new thoughts on

the basis of thoughts that are already entertained. In this sense, vision is a cognitive, that is, thought involving, process.

If perception is to be thought of as some sort of thinking, its processes must necessarily first include transformations of states that are expressed in symbolic or propositional form, and, second, these transformations must be inferences from some states that function as premises to a state that is the conclusion of the inference. That is to say, visual processes must be inferences or arguments, exactly like the processes of rational belief formation. These two conditions follow directly from the claim that perception is some sort of thinking, since the characteristic trait of thinking is drawing inferences (whether it be deductive, abductive, or inductive) operating on symbolic forms by means of inference rules that are represented in the system, although thinking is not reduced to drawing inferences this way. In view of these considerations, the principles guiding the transformations of perceptual states, that is, the principles (such as Spelke's principles) acting as the inference rules in perceptual inferences, must be expressed in the system and, specifically, must be represented in a symbolic form. Whenever the system needs some of the principles to draw an inference, it simply activates and uses them. In addition, the premises and the conclusion of a visual argument be represented in the viewer in a propositional-like, symbolic form.

If these conditions are met, perception involves discursive inferences, that is, drawing propositions or conclusions from other propositions acting as premises by applying (explicitly or implicitly) inferential rules that are also represented in the system. Clark's view quoted above seems to echo this thesis in so far as Clark conceives the processes of visual perception as a rational process of belief fixation. It follows that the inferences used in perception are no different from the inferences used in thought. That is, they are discursive inferences.

A short digression is needed here, however, lest we attribute to Clark intentions that he may not have. The previous analysis assumes the standard view of the brain as a physical machine that processes symbols in purely formal or syntactic way on the basis of the physical properties of the symbols; the brain performs digital computations. These symbols have meaning, of course, and so do the transformations of these symbols, but the processes in the brain are independent of any meaning. To put it differently, the brain is a syntactic machine that processes symbols that have meaning. The standard view can be modified by adding the thesis that digital computations are not merely formal syntactic manipulations but also involve semantics, that is, the contents of the states that participate in computations

are causally relevant in the production of the computations' outputs [26.5].

Although this is the standard, algorithmic, view of cognition, it is by no means unequivocally endorsed. There is another, competing view of cognition, according to which the brain is not a syntactic machine that processes symbols through algorithms. The brain represents information in a nonsymbolic, analogue-like form, as activation patterns across a number of units. Furthermore, the processes in the brain do not assume the form of algorithmic but of algebraic transformations; this is the connectionist view of cognition, of which Clark is a stern proponent. This is not the place to expand and explain connectionism, but I wish to stress that in this view of cognition, the brain does not use at all discursive inferences, although some of its behavior certainly simulates the usage of discursive inferences. If this is so, Clark's thesis that perception is inseparable from the rational processes of belief fixation does not commit him to the view that perception employs discursive inferences for the simple reason that thinking itself does not implicate such inferences.

Furthermore, given the propositional or symbolic form of the format in which the states of the visual system must be represented if vision is akin to thinking, the contents of these states, that is the information carried by the states, consists of concepts that roughly correspond to the symbols implicated; it is conceptual content. If vision is some sort of thinking, therefore, its contents must be conceptual contents. This means two things. Either the visual circuits store conceptual information that they use to process the incoming information, or they receive from the inception of their function such information from the cognitive areas of the brain while they are processing the information impinging on the retina. Spelke's principles that guide visual processing and render the percept possible are examples of conceptual content.

It should be noted that discursive inferences are distinguished from *inferences* as understood by vision scientists according to whom any transformation of signals carrying information according to some rule is an inference [26.6]:

"Every system that makes an estimate about unobserved variables based on observed variables performs inference [...] We refer to such inference problems that involve choosing between distinct and mutually exclusive causal structures as causal inference."

One could claim, therefore, that although inferences, in this liberal sense, occur in the brain during visual perception, they are not like the inferences used

in thought. One might even go further than that and claim that these inferences, or rather state transformations, do not involve representational states at all [26.7]. Although the percept is certainly a representational state, the processes that lead to its formation are not representations. It follows that visual perception is not a cognitive process, if *cognitive* is taken to entail the use of mental representations; "a system is cognitive because it issues mental representations" [26.7].

In this chapter, I examine vision and its processes and discuss the relation of vision with thinking. I do not have the space here to discuss the problem of whether visual processes involve representations. I proceed by assuming that they do although, first, as I will argue, the state transformations do not presuppose the application of inference rules that are represented in the system, and, second, not all visual states are representational.

In Sect. 26.1, in view of the close relationship between thinking and inference, I chart and briefly discuss inference and its modes, namely, deduction, induction, and abduction or inference to the best explanation.

In Sect. 26.2, I sketch an overview of the main conceptions concerning vision, to wit constructivism, direct or ecological theory of vision, and the more recent proposals that view vision as inseparable from action.

In Sect. 26.3, I present the two stages of which visual perception consists, namely early vision and late vision.

In Sect. 26.4, I discuss the problem of the cognitive penetrability (CP) of perception, because if vision is akin to thinking, visual processes necessarily involve concepts and are thus cognitively penetrated. If it turns out that some stage of vision is cognitively impenetrable (CI) and conceptually encapsulated, the status of the logical characterization of the visual processes of that stage remains open since, being nonconceptual in nature, they cannot be discursive inferences. I am going to argue that a stage of vision, early vision, is CI and has nonconceptual content. This content is probably iconic, analogue-like and not symbolic. By not being symbolic, the contents of the states of early vision cannot be transformed to some other contents by means of discursive inferences in so far as the latter operate on symbolic forms. The second main visual stage, namely late vision, is CP and implicates concepts. I also address in this section two problems with my claim that early vision is conceptually encapsulated. The first is raised by the existence of some general regularities that seem to guide the functioning of the perceptual system, of which the Spelke principles are a subset, and which operate at all levels of visual processing. The problem is, first, whether the existence of such principles entails that at least some part of the information

processed in early vision is inherently conceptual, and, second, whether the existence of such principles entails that vision in general is theory-laden. The second concerns the effects of perceptual learning, since one might argue that through perceptual learning some concepts are embedded in the perceptual circuits of early vision. If either of these two is correct, the states of early vision have conceptual contents and thus the processes of early vision may involve discursive inferences rendering early vision akin to thought and belief formation. I argue, however, that neither the principle nor the effects of perceptual learning entail that early vision has conceptual content.

## 26.1 Inference and Its Modes

Let us grant that vision is like thinking and, therefore, involves discursive inferences. The question that arises concerns the nature of the inferences involved; are they deductive, inductive, or abductive? (Appendix 26.A for a definition of deductive, inductive, and abductive inference).

I think it is safe to assume that the whole visual process fits better the description of an abductive inference. The main reason for this thesis is that vision constructs a representation, (i. e., the percept) that best fits the visual scene. Specifically, given that the retinal image is sparse and thus underdetermines both the distal object and the percept, the visual system has to fill in the missing information to arrive at the best explanation, that is, the percept that best fits the retinal information. In essence, given the sparsity of the incoming information in the retinal image, the brain attempts to construct a representation that consists of the properties that an object should have in order to produce the specific retinal image. That is, the brain works back from the information that the retinal image contains to the object that could produce such a retinal image. Many objects could produce this image and the brain attempts to figure out which one of them best fits the retinal image. This is the trait par excellence of an abductive inference. Recent work (see [26.4] for an overview) suggests that this abductive inference or inference to the best explanation is a Bayesian inferences in which the brain constructs the percept that best explains the visual input by selecting the hypothesis that has the highest probability given the visual input.

It follows that the inference is ampliative, that is, the conclusion has a wider content than that of the premises and thus is not implicitly included in the premises; as such, the inference is not deductive. This is easy to grasp if we consider that the only information im-

In Sect. 26.5, I examine the logical status of the processes of early and late vision and argue that the processes of early vision are abductive nondiscursive inferences that do not involve any concepts, while the processes of late vision despite the fact that they are abductive inferences guided by concepts, are not discursive inferences either. I argue that the abductive inferences involved in visual perception are not sentential inferences but, instead, they rely on pattern-matching mechanisms that explore both iconic, analogue-like information and symbolic information. In this sense, visual abduction could be construed as consisting of a series of model-based inferences.

pinging on the retina consists of differences of light intensities and electromagnetic wavelengths. The percept that which the visual processes output (and since we have assumed that vision is a complex inference, the premises of the inference consist in the impinging information and the percept is the conclusion of this inference), however, is the object of our ordinary experience with its shape, size, color, motion, texture, etc. All these properties far exceed the impinging information concerning light intensities and wavelengths.

Moreover, and related to the first consideration, even if the premises of a visual inference that outputs the percept are correct, that is, even if the principles that guide perception reflect correctly the physical and geometrical regularities, and the impinging information being what it is, the percept may still not be a correct representation of the object in the environment that emanated the light rays and caused the perception. In other words, the conclusion may be wrong even though the premises are correct. This is why vision should be better understood as an abductive process or as an inference to the best explanation. Traditionally, abduction is thought as synonymous to the inference to the best explanation (for a recent reaffirmation see [26.8]). Recently, however, this thesis has come under attack mainly on the ground that abduction is for the generation of theories, whereas the inference to the best explanation is for their evaluation [26.9, 10]. Although I agree with Lipton, I will not dwell on this issue here any further. I will continue to use *abduction* as synonymous to *inference to the best explanation* because nothing important in the discussion in this chapter hinges on the outcome of this debate.

One may wonder why this ampliative, non-truth-preserving inference should be construed as an abductive inference and not as an inductive inference. One

might argue that all inductions are abductions or inferences to the best explanation [26.11]. Most authors, however, think that abduction is a subspecies of induction since it bears the basic marks of induction as it is ampliative and does not preserve truth. However, it is more specific than induction since it aims exclusively to pinpoint the cause or causes for some phenomena, that is, it aims to yield an explanation of a set of phenomena. Not all inductions are focused towards this aim. Several times a good induction leads to a generalization that subsumes a set of phenomena under the heading of a generalization, which, however, does not explain the phenomena. Consider the following induction.

Bird  $\alpha$  is a crow and is black ( $Ca&Ba$ )  
 Bird  $\beta$  is a crow and is black ( $Cb&Bb$ )  
 ...  
 Bird  $\kappa$  is a crow and is black ( $C\kappa&B\kappa$ )  
 Therefore (inductively)  
 All crows are probably black ( $(x)(Cx \rightarrow Bx)$ )

Under certain conditions this is a good induction in which from the colors of specific specimens of crows

one infers the color of all crows. This is hardly a good explanation though. A good explanation seeks to explain, that is, make us or the scientific community understand why crows  $\alpha$  and  $\beta$  are black. The generalization *All crows are probably black* fails to accomplish this since all that it does is gather together all instances of black crows in a generalization. Moreover, a good explanation of a set of phenomena is expected to have a wider range than these specific phenomena in the sense that it can be used as a springboard to explain a wider class of phenomena. In our case, a good explanation of why crows  $\alpha$  and  $\beta$  are black should certainly involve genetics. Such an account not only would provide understanding of the correlation of crows with the color black, but it could also be used to explain the colors of other species. Now, it is widely agreed that the discovery of the relevant laws of genetics would fall within the purview of abduction. To put this point differently, all abductions are inductive inferences but not all inductions are abductions.

When I examine in Sects. 26.3 and 26.5 the visual processes in some detail, I shall adduce more evidence supporting the claim that visual processing is an abductive inference.

## 26.2 Theories of Vision

I have claimed that vision is a complex process that starts when light impinges on the retina and culminates with the formation of the percept, that is, the object of our ordinary experience and its properties. If vision as a whole is a complex process, it consists of a series of processes, or, in other words, in a series of state transformations in which one state containing some information is transformed via the visual mechanisms to a state containing some other sort of information. According to this view, vision is a process in which the visual system constructs the percept from the incoming visual information. All these processes take place within the visual system and although information from the other modalities and the actions of the viewer may either facilitate or inhibit the visual processing, vision in principle is autonomous from the other modalities and action.

This thesis can be assaulted from at least two fronts. The first is to deny that vision is a complex process involving information processing. It may be the direct retrieval of visual information from the environment without any need for mediating processes. The proponents of this view are divided into two camps. The first maintain that the retrieval of information from the environment is mediated by representations, while the

second deny the necessity of invoking representations to explain how visual perception works. The second is to claim that although vision necessarily involves inferences, vision cannot be separated from action in that actions figure inherently and constitutively in vision. In this section, I present the three different conceptions of vision.

### 26.2.1 Constructivism

Visual perception begins with information impinging on the retina, this is the stimulation of the sensory organs, and culminates with the construction of the percept, which is a visual representation of the worldly objects (they are called *distal objects*) that emanate the light that stimulates the sensory organs. This is made possible through a series of transformations whereby the information impinging on the retina is progressively transformed into a final visual representation, the percept. The construction of the final visual representation is preceded by the construction of a host of intermediate visual representations of increasing complexity.

The transformation from one visual representation to the other, which are both mental representations being located in the brain, is effectuated through the



processes of vision that consist of the application of transformational rules that take as input representation  $r_1$  at time  $t_1$  and output representation  $r_{t+1}$  at time  $t_2$ . These rules could be construed as abductive inferences since the brain is called upon to fill in the gaps in the information contained in the retinal image in order to construct a representation of the distal object that is the most likely candidate for being the object that could have produced the retinal image. It could be argued, hence, that the brain guesses which object is the best fit to explain the retinal image.

Since visual perception consists of a series of constructions of visual representations, vision is a constructive process. Let us call this construal of visual perception *constructivism*. According to one of the most influential visual scientists that espouse constructivism, Marr [26.12], there are three levels of representation. The initial level of representation involves Marr's *primal sketch*, which consists of the *raw primal sketch* and the *full primal sketch*. The *raw primal sketch* provides information about the edges and blobs present in a scene, their location and their orientation; this information is gathered by locating and coding individual intensity changes. Grouping procedures applied to the edge fragments formed in the *raw primal sketch* yield the *full primal sketch*, in which larger structures with boundaries and regions are recovered. Through the *primal sketch* contours and textures in an image are captured. The primal sketch can be thought of as a description of the image of a scene but not as a description of the real scene. This latter involves the relative distances of the objects and their motions. This information is provided by the viewer-centered representation, which is Marr's  $2^{1/2}$  *sketch*. At this level information about the distance and layout of each surface is computed using various depth cues and by means of analysis of motion and of shading. This information describes only the parts of the object that are visible to the viewer and thus is relative to the viewer.

The computations leading to the formation of the  $2^{1/2}$  *sketch* are determined by three factors:

1. The input to the visual system, that is, the optical array
2. The physiological mechanisms involved in vision, and the computations they allow, and
3. Certain principles that restrict and guide the computation.

These principles are constraints that the system must satisfy in processing the input. These constraints are needed because perception is underdetermined by any particular retinal image; the same retinal image could lead to distinct perceptions. Thus, unless the

observer makes some assumptions about the physical world that give rise to the particular retinal image, perception is not feasible.

It is important at this juncture to stress that according to Marr, all the processes that lead to the formation of the  $2^{1/2}$ D sketch are data-riven; they are driven solely by the input.

One of the aims of vision is the recognition of objects. This requires the matching of the shape of a structure with a particular object, a matching that requires an object-centered representation. This is Marr's *three-dimensional (3-D) model*. The recovery of the objects present in a scene cannot be purely data-driven, since what is regarded as an object depends on the subsequent usage of the information, and thus is task dependent and cognitively penetrable. Most computational theories of vision [26.12, 13] hold that object recognition is based on part decomposition, which is the first stage in forming a structural description of an object. It is doubtful, however, whether this decomposition can be determined by general principles reflecting the structure of the world alone, since the process appears to depend upon knowledge of specific objects [26.14]. Object recognition, which is a top-down process and requires knowledge about specific objects, is accomplished by the high-end vision. The construction of the percept, which is the end product of visual perception, therefore requires the synergy of both top-down and bottom-up transfer of information between the visual circuits and the cognitive centers of the brain. Object recognition requires matching the internal representation of an object stored in memory against the representation of an object generated from the image. In Marr's model of object recognition the 3-D model provides the representation extracted from the image that will be matched against the stored structural descriptions of objects (perceptual classification). (It should be emphasized that these *object recognition* units are not necessarily semantic, since we may recognize an object that we had seen before, even though we have no idea of its name, of what it does and how it functions, that is, even if we have no semantic and lexical information about it. Ref. [26.15] introduces a distinction between the *perceptual classification* and *semantic classification* and *naming*. These processes are independent one of the other. ). See Appendix 26.B for an overview of constructivism.

Marr's and Biederman's hypothesis that object recognition occurs through part decomposition is based on the conception of three-dimensional objects as arrangements of some set of primitive 3-D shapes. According to Marr, these primitive 3-D shapes are generalized cylinders (Fig. 26.1) that are defined in terms of major axes and radii of objects.

According to Biederman, the primitive 3-D shapes are the so-called geons (Fig. 26.2). All objects can be decomposed into a set of 36 specific geons related in various ways. The properties that identify geons and allow them to function as volumetric perceptual primitives are viewpoint invariant, that is, they do not change as the angle of view changes. As such, they are called nonaccidental features since they are features not only of the image but also of the worldly objects (that is, they are properties that exist in the environment outside the viewer) that do not depend on what the viewpoint may be accidentally. Examples of nonaccidental properties are parallel lines and collinearity. If an object has parallel lines many rotations of this object yield an image in which these lines are still nearly parallel; that is to say, parallelism is a property that is rotation- or perspective-invariant.

Let me close the account of constructivism by reminding the reader that the theories of visual perception presented in this part of the chapter are some among the many different theoretical accounts of visual processing. The differences between the various theories notwithstanding, all constructivist theories share a common core, namely that visual perception involves state transformations in the course of which visual representations of increasing complexity are being gradually constructed by the visual system. The visual processes start from the meager information contained in the retinal image and which consists of local distributions of light intensities and wavelengths. These transformations can also be construed as computations in which the brain computes an output state given an input state. Many of these transformations (but not all of them) act on and therefore essentially involve mental representations that are within the brain of the viewer, and can be independent of any other activities on the part of the viewer. The transformations are made possible through the application of transformational rules, such as, for example, the rule that abrupt changes in light intensity signify the presence of edges that is used by the perceptual system to construct the raw primal sketch. Such a rule takes as input states that carry information about various light intensities distributed in space and delivers states that carry information about edges. It follows that the transformations taking place in visual processing are information-processing operations. (I said that not all of the transformations operate on representations because many of these transformations operate on states that are not representational. It would require another chapter to discuss the conditions under which a state is representational or not and, of course, much depends on how one defines the term *representation*. I confine myself to pointing out

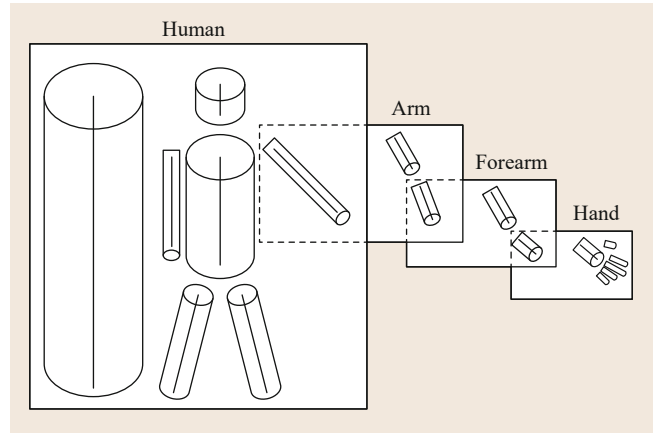


Fig. 26.1 Marr's generalized cylinders

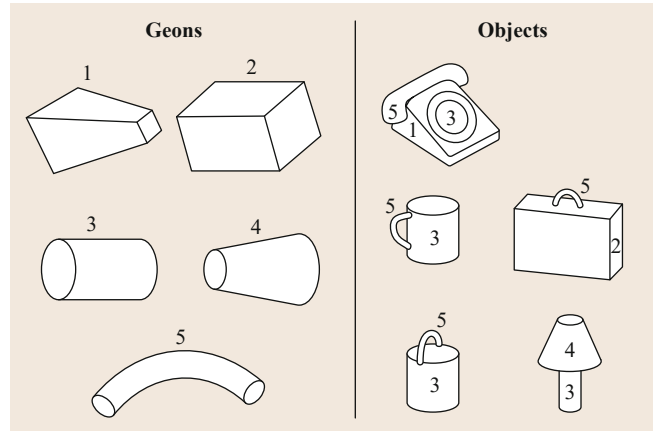


Fig. 26.2 Biederman's geons

that many of the earlier visual states are probably not representational because they do not meet the criteria that an adequate definition of representation posits, and to referring the reader to the discussion in Chap. 4. As we shall see in Sect. 26.4, one could claim that there is a sharp distinction between internal probabilistic dependencies between states that can be explained by internal causal connections between the circuits of the brain and those that cannot; only those that cannot be explained internally carry information about the external world and thus involve representational states.)

The fact that the visual brain transforms states to other states through the usage of some rules means that the function of the brain can be understood as a series of inferences from some state/premises to some other states/conclusion. In view of our discussion in the beginning of this section, as well as in the previous one, the inferences most likely are abductive in nature.

### 26.2.2 Theory of Direct Vision or Ecological Theory of Visual Perception

Gibson [26.16] started from a very different assumption than that of constructivism. In contradistinction to the latter, Gibson argued that perception begins not with the informationally sparse retinal image but with the informationally rich *optic array*. The spatial pattern of light intensities and the mixture of wavelengths that impinge on the receptors of the retina form the optic array. This light, however, carries a lot of information about the solid objects in the environment (the distal objects) because the intensities of light and its wavelengths vary from one solid visual angle to another (as the rays of light emanating from solid objects travel in space and between the surfaces of the objects that fill the space, given that at any point in space light converges from all directions, at each point in space there is a nested set of solid visual angles). As a result, the optical array is determined by, and therefore carries information about, the nature and location of the three-dimensional worldly surfaces from which it is being reflected.

Unlike the retinal image, the optic array is an external source of information, or, better, an external information-bearing structure since it exists outside the viewer, is independent of the constitution of the brain of the viewer, and carries information about the environment. Gibson's central claim is that the information contained in the optic array suffices to allow perceivers to specify the visual scene that causes the optic array, that is, to specify the solid surfaces that surround them, and to use the information included in the optic array to interact with their environment.

When perceivers move in their environment, moreover, the entire optic array is transformed to reflect the new environment since it depends exclusively on it. As perceivers move around, they sample different optic arrays and therefore receive a variety of information about the changing environment, since the transformations of the optic array as perceivers move contain information about the layout of the objects in the environment as well. As in realistic situations perceivers are not static, motion enriches the visual information that the perceivers receive from the environment enabling them to recover the visual scene much easier than if they were static. Furthermore, this motion by effecting transformations of the optic array allows the perceivers to identify those aspects of objects that remain invariant under movement (the nonaccidental properties that we have discussed). It goes without saying that this information is made available only to perceivers that move in their environment and effect a change in the optic array that they receive from the environment; a static perceiver would never be able to detect the properties

of objects that remain invariant under motion. Note that information about the invariant properties is available in the environment, but viewers can retrieve or detect it only as they move. This entails that perception becomes entangled with action, since moving around is a form of action.

The richer the information that the light impinging on the retina carries, the less information processing the visual brain is required to do in order to form the percept. Taking this view to its extreme end, one might claim that if the optic array suffices all by itself to enable viewers to recover the visual scene, there is no need to posit any internal information processing on information-bearing states. Visual perception involves no information processing and no inferences of any sort; it just recovers the visual scene directly from the information contained in the optic array (which explains coining this theory a theory of direct vision). This interpretation of the theory received a devastating criticism in Fodor and Pylyshyn's [26.17] paper entitled *How Direct is Visual Perception*. I think it safe to assume that the radical interpretation that excludes information processing from visual perception has not recovered from this critique since most of the counterarguments raised in that paper have not been adequately answered. Whether, however, Gibson subscribed to this radical view is debatable. Be that as it may, the radical interpretation is not the only possible interpretation of direct vision.

The fact that the input to the visual system may contain more information than that envisaged by constructivism does not entail that visual perception does not involve any internal information processing. It only entails that the internal information processing needed for the formation of the percept is less than in constructivist theories since a part of it is being replaced by the manipulation through motion and transformation of the optic array, which as you recall is an external information-bearing structure. Nor does the richness of the information in the input entail that no representations are needed; it entails that visual perception allows positing less representations than those required in constructivist theories. As Rowlands [26.18] remarks:

“Here is nothing in Gibson's theory itself – as opposed, perhaps, to his statements about his theory – that entails or even suggests that all of the role traditionally assigned to manipulation and transformation of internal information-bearing structures can be taken over by the manipulation and transformation of external information-bearing structures.”

In this moderate interpretation of Gibson's theory of direct vision, the need for some information processing over internal representational states still persists, except

that, in view of the fact that the information contained in the visual input is richer than previously thought, this need is attenuated. Therefore, visual perception involves some sort of inferences.

Gibson's theory was coined the *theory of direct perception* because it relinquished the need for internal information processing; instead, the viewers retrieve all the information they need to detect the environment directly from the environment without any internal processing of any sort mediating the process of information retrieval. If, however, some information processing over internal representations is needed as well, as a moderate form of Gibson's theory asserts, can the qualification *direct* be salvaged?

There is a sense in which it might. Suppose that *direct* is construed so as to emphasize not the lack of information processing operating on internal representations, but the fact that the information processing is entirely data-driven, that is, guided by environmental input and some principles that reflect regularities in the environment, and the whole process is not influenced by other internal nonvisual states of the viewer, such as the viewer's cognitive or emotional states. If this supposition is borne out, then visual perception is direct in the sense that the whole process is data-driven and, as such, the information processing used operates over information retrieved exclusively from the environment. Note that this presupposes that the principles guiding visual processing do not constitute some form of intervention on the part of the viewer whose contribution exceeds what is given in the environment.

This assumption is borne out if visual perception or at least some stage of it, is purely data-driven, that is, cognitively and emotionally impenetrable. If cognitive states penetrate and thus influence perceptual processing, the viewer's cognitive states actively contribute to the formation of the percept and the visual processing does not retrieve information directly from the environment but only through some cognitive intervention; visual perception, in this case, is not direct. Norman [26.19] has argued along this line that the processing along the dorsal visual pathway that guides our on-line interactions with the environment, owing to the fact that when it operates immediately on the visual input it is entirely data-driven, is a visual function that conforms very closely to Gibson's *direct* theory. The ventral visual pathway, in contradistinction that is responsible for object recognition and categorization is clearly affected by cognition and, in this sense, is not a *direct* visual function. Since both visual pathways are found in the brain, the constructivist and the ecological theories of perception can be reconciled.

Even though it seems abundantly clear that visual perception requires a significant amount of information

processing, and in this sense one of Gibson's main insights is considered to be wrong, several of Gibson's insights have been incorporated in the constructivist information-processing research program. For example, most, if not all, information-processing theories hold that most of the ambiguities that occur during the information processing of the retinal input cannot be resolved by that input alone and need top-down assistance only when information comes from a static monocular image. When additional information can be derived from stereopsis and motion of real scenes, then the information-processing program can resolve the ambiguity without the need of a top-down flow of information. If one takes into account the real input to human vision, which is binocular and dynamic, there are few ambiguities that cannot be resolved through a full consideration of the products of the early visual processing *modules* [26.20]. This shows that the dynamic and interactive character of vision solves several problems encountered within the information-processing research program.

Our discussion about direct vision revealed an aspect of visual processing that traditional constructivist theories did not initially consider, namely the interaction of perception and action. The next kind of theory of perceptual processing that we will examine views visual perception as inextricably linked with action and uses the most recent neuropsychological evidence, vision science research, and computer modeling to both substantiate this claim, and draw the details of how the active visual brain works in order to provide a fully fledged unifying model of perception and action. Although this model aims to cover all modalities, for the purpose of this chapter I will restrict the presentation and discussion to visual perception.

### 26.2.3 Predictive Visual Brain: Vision and Action

The basic tenet of the theory of ecological or direct vision is that all the information viewers need to recover the visual scene that causes the retinal image is already included in the incoming information in the optic array. Little or no information processing is required for the construction of the percept. The constructivist theories of visual perception, in contradistinction, underline the necessity of information processing and state transformations in the brain. The flow of information in the brain is bidirectional; both top-down and bottom-up signals are transmitted and the ensuing percept is the result of the synergy between top-down and bottom-up processing. This class of theories assumes that the representation constructed at some level is transmitted bottom-up to the neuronal assembly at the next

immediate level where it is further processed. Moreover, recurrent signals return top-down to earlier levels mainly to test hypotheses concerning aspects of the visual scene (recall that visual perception aims to recover the visual scene that causes the retinal image and does that by constructing increasingly more complex representations of the probable aspects of the visual scene at various spatial and temporal scales) until the percept is constructed.

Recent empirical findings and modeling shed light on the way the brain actually effectuates these processes. These details, as we shall see, entail certain deviations from the traditional constructivism image, which concern (a) the sort of information transmitted bottom-up; only prediction errors are transmitted to the next level, (b) the nature of the representations constructed; they are distributions of probabilities rather than having a unique value (note that this new approach emphasizes the indispensable role of representations in visual processing), and (c) the interaction between perception and cognition. This last trait is very important has important repercussions for our discussion on the relation between visual processing and thinking.

According to this view of visual perception, brains are predictive machines [26.4]:

“They are bundles of cells that support perception and action by constantly attempting to match incoming sensory inputs with top-down expectations or predictions. This is achieved using a hierarchical generative model that aims to minimize prediction error within a bidirectional cascade of cortical processing.”

A hierarchical generative model as applied to visual processing is a model of perceptual processes according to which the brain uses top-down flow of information (enabled by top-down neural connections) in an attempt to generate a visual (meaning, in the brain) representation of the visual scene (in the environment) that causes the light pattern impinging on the transducers and the low-level visual responses to this light pattern. The brain attempts to recover gradually the causal matrix (the various aspects of a visual scene) that causes and thus is responsible, for the retinal image seen as a data-structure (i. e., the sensory data). The brain does that by capturing the statistical structure of the sensory data, that is, by discovering the deep regularities underlying the retinal structure, on the very plausible assumption that the deep structure underneath the sensory data reflects, so to speak, the causal structure of the visual scene.

Hierarchical generative models attempt to achieve this by constructing, at each level, hypotheses about the probable cause of the information represented in

the immediately previous level, and testing these hypotheses by matching their predictions with the actual sensory data at the preceding processing level. Suppose, for example that a neuronal assembly at level  $l$  receives from level  $l-1$  information concerning differences in light intensities. The higher level attempts to recover the probable edges that cause the variation in light intensity and forms a hypothesis involving such edges. Now, and this is the crucial part, if this hypothesis were correct, that is, if the edges as represented in the hypothesis were present in the environment, then a certain pattern of variation of light intensities at the appropriate local scale would have been present in the sensory data. This prediction is transmitted top-down to level  $l-1$  and matched against the actual pattern of variations in light intensities. If there is a match (with an acceptable degree of error deviation due to the inherent noise of the signal, of course) no further action is needed since the perceptual system *assumes* that it has constructed the correct, at this spatial scale, representation of the relevant environmental input. If the match reveals a discrepancy, that is, if an error in the prediction is detected, this prediction error is transmitted bottom-up to level  $l$  so that a new hypothesis be formulated and tested until, eventually, no unacceptable prediction error persists. If one thinks of the discovered error as a surprise for the system, the system strives to correct its hypotheses so that by making correct predictions, the testing of the hypotheses yields no surprises; this is a typical error-driven learning process where a system learns, i. e., constructs a correct representation, by gradually reducing error. The hierarchical generative models hence generate, in essence, low-level states (the predictions they make about the activities at the lower levels) from high-level causes (the hypotheses that would, if correct, explain the activity at the lower levels).

Bidirectional hierarchical structure allows the system to [26.4]:

“infer its own priors (the prior beliefs essential to the predicting routines) as it goes along. It does this by using its best current model at one level as the source of the priors for the level below, engaging in a process of *iterative estimation* that allows priors and models to coevolve across multiple linked layers of processing so as to account for the sensory data.”

To form hypotheses concerning the probable cause of the sensory data at a certain level, at a specific spatial and temporal scale, the neuronal assembly at the next level, say level  $l$ , uses information not only about the sensory data at the previous level (or, to be precise, information regarding its prediction error) that is transmitted bottom-up, but also higher-level information that

is transmitted to *l* either laterally, that is, from neuronal assemblies at the same level (neurons in V1 processing wavelengths inform other neurons in V1 processing shape information, for example), or top-down from levels higher in the hierarchy (neurons in V4, for instance, are informed about the color of incoming information from neurons in the inferotemporal cortex in the brain (IT) as a result of precueing – that is, when a viewer has been informed about the color of an object that *will* appear on a screen). This higher-level information may and usually does concern general aspects of the world (such as *solid objects do not penetrate each other, or solid objects do not occupy exactly the same space at the same time, etc.*), and may also reflect knowledge about specific objects learned through experience. All this lateral and top-down flow of information provides the context in which each neuronal assembly constructs the most probable hypothesis that would explain the sensory data at the lower level. Thus, context-sensitivity is a fundamental and pervasive trait of the processing of hierarchical predictive coding; the contextualized information significantly affects, and on occasions (as in hallucinations) may override, the information carried by the input.

The hierarchical predictive processing model can be naturally extended to include action and thus closely ties perception with action [26.4]. This is the action-oriented predictive processing. Action-oriented predictive processing extends the standard hierarchical predictive model by suggesting that motor intentions actively elicit, via the motor actions they induce, the ongoing streams of sensory data at various levels that our brains predict. In other words, once a prediction is made about the state in the world that causes the transduced information, the action-oriented predictive processes engage in a search in the environment of the appropriate worldly state. Suppose, for example, that owing to bad illumination conditions, a perceiver is unsure about the identity of an object in view. Its brain makes a prediction about the putative object that causes the sensory data the perceiver receives, and the perceiver moves around the object in order to acquire a better view that will confirm the prediction. By moving around, the perceiver's expectations about the proprioceptive consequences of moving and acting directly cause the moving and acting since where and when the perceiver moves is guided by the aim that the perceiver's action brings the object into a better view.

It is worth pausing at this point to discuss briefly the problem of nature of the relation between visual perception and action and, specifically, motion. Is this relation constitutional, which means that if someone cannot or does not move they cannot visually perceive anything? This claim was initially made by *Noe* although,

in view of vehement criticism, *Noe* has attempted to modify it without compromising the main tenets of his views [26.21]:

“When you experience an object as cubical merely on the basis of its aspect, you do so because you bring to bear, in this experience, your sensorimotor knowledge of the relation between changes in cube aspects and movement. *To experience a figure as a cube, on the basis of how it looks, is to understand how it looks changes as you move* (emphasis added).”

The sensorimotor knowledge consists of the expectations of how our perception of an object changes as we move around it, or as this object moves with respect to us. These expectations constitute a form of practical knowledge, a *knowing how* as opposed to a *knowing that*. Thus, to be able to experience visually an object, one needs to have the ability to move around the object and explore it. Visually experiencing the object literally consists of grasping the relevant sensorimotor contingencies, that is, the sensorimotor knowledge associated with this specific object. There are two ways to read this claim. According to the first reading, which *Noe* seems to espouse judging from the previously cited passage, to be able to visually perceive requires the actual exercise of the ability to probe the world. According to the second reading, visually perceiving an object only requires the ability to probe the world but not the actual exercise of this ability. The first reading entails immediately that prior to exercising this ability, one does not visually perceive the object. Since this is absurd, one has to concede that viewers do not need to exercise actually the ability to probe the environment, it suffices that they take recourse to their experience with similar objects and retrieve the requisite sensorimotor contingencies from experience. Even if one takes this line, however, the problem remains that at the time of a first encounter with an object to be able to see its, say, shape, one should be able to probe the object either by moving around the object, or by having the object move around them. Thus, when stationary viewers perceive a stationary novel object, lacking any knowledge of sensorimotor contingencies, they do not see its shape or its other properties.

It follows that infants upon opening their eyes for the first time and facing the world, by lacking any sensorimotor knowledge and by not probing the environment, they do not see anything. This claim flies to face of countless empirical evidence, which shows that there is something fundamentally wrong with equating visual perception with understanding sensorimotor contingencies and deploying the relevant practical knowledge. This entails, in turn, that the relation between visual

perception and action, no matter how important it is, is not a constitutional relation; one gets to see the world even if both they and the world are stationary, although it goes without saying that their experience will be restricted compared to other viewers who can probe the environment. They could not visually experience, for example, Marr's 3-D sketch because they lack knowledge of the unseen surfaces of objects.

This unity between perception and action emerges most clearly in the context of *active inference*, where the agent moves its sensors in ways that amount to actively seeking or generating the sensory consequences that their brains expect. "Perception, cognition, and action work closely together to minimize sensory prediction errors by selectively sampling, and actively sculpting, the stimulus array" [26.4]. Their synergy moves a perceiver in ways that fulfill a set of expectations that constantly change in space and time. Accordingly, perceptual inference is necessary to induce prior expectations about how the sensorium unfolds and action is engaged to resample the world to fulfill these expectations.

Since the construction of the representations of the putative causes of the sensory inputs is made possible through a synergy of bottom-up processing transmitting the prediction errors and top-down processing transmitting for testing the hypotheses concerning the probable causes of the input and in so far as the processes constructing these hypotheses are informed by high-level knowledge of the sort discussed above, visual perception unifies cognition and thinking with sensation; these two become intertwined. This means that perception inextricably involves thinking. Notice that this account of visual perception necessarily involves representations; it requires that each level retain a representation of the data represented at this level so that the top-down transmitted predictions of the hypotheses formed at subsequent higher levels be matched against the information represented at the lower level in order for the hypothesis to be tested. It also requires the representation of the putative causes of the sensory data at the preceding level; these are called the representation-units, which operate along the error units (the units that compute the error signal, that is, the discrepancy between prediction and actual data) in a hierarchical generative system.

Furthermore, testing hypotheses and altering them as a result of any prediction errors until the prediction error is minimized and thus until the most probable cause of the sensory data has been discovered, is an inference. Being a probabilistic inference that aims to

discover the most probable hypothesis that explains away a set of data, it is most likely a Bayesian inference. It is very plausible, therefore, that the computational framework of hierarchical predictive processing realizes a Bayesian inferential strategy (see Appendix 26.C for an analysis of Bayes' theorem). Indeed, recent work on Bayesian causal networks [26.22] presents the brain as a Bayesian net operating at various space and time scales.

What Bayes' theorem, on which this strategy is based, ensures is that a hypothesis is eventually selected that makes the best prediction about the sensory data minimizing thereby the prediction error and thus best explains them away; that is a hypothesis that by having the highest posterior probability provides the best fit for the sensory data. The construction of this hypothesis crucially and necessarily involves the context, as it is clearly expressed in Bayes' equation in the form of the prior probability for the hypothesis  $P(A)$ , whose value depends on the context. That is to say, it is the context that provides the initial plausibility of a hypothesis before the hypothesis is tested.

This enables Clark [26.4] to claim that in the framework of predictive brains that use hierarchical generative processing perception becomes theory-laden in the specific sense that what viewers perceive depends crucially on the set of priors (that is, the hypotheses that guide the predictions about the matrix of the sensory data at the lower processing levels, which the hypothesis projects) that the brain brings to bear in its attempt to predict the current sensory data. This remark brings us back to the main theme of this chapter, namely, the relation between perceiving and thinking. If thinking necessarily implicates discursive inferences and deploying concepts, as it usually does, Clark's claims entail that perception employs from its onset concepts and draws discursive inferences. To assess this dual claim, we must examine the processes of vision to determine first whether concepts are used and if the answer is affirmative the extent to which they are being used, and second, whether the inferences that are undoubtedly used in perception must necessarily be discursive. I hasten to note that, with respect to this last problem, nowhere in his account does Clark suggest that the inferences must be discursive. In fact, the sources he refers to, especially those concerning connectionist neural networks, suggest that the inferences on which perception relies may take another form and need not necessarily involve propositionally structured premises and conclusions.

## 26.3 Stages of Visual Processing

I said above that we must examine the processes of vision with a view to determine whether and, depending on the answer to this question, the extent to which, cognition penetrates visual perception in the sense that perceptual processing uses conceptual information that is either transmitted top-down to perceptual circuits, or is inherently embedded in visual circuits. In the literature, visual processing is divided into two main stages, to wit, early vision and late vision.

### 26.3.1 Early Vision

Early vision is a term used to denote the part of perceptual processing that is preattentive, where attention means top-down, cognitively driven attention. *Lamme* [26.23, 24] argues for two kinds of processing that take place in the brain, the feedforward sweep (FFS) and recurrent processes (RP). In the FFS, the signal is transmitted only from the lower (hierarchical) or peripheral (structural) levels of the brain to the higher or more central ones. There is no feedback; no signal can be transmitted top-down as in RP. Feedforward connections in conjunction with lateral modulation and recurrent feedback that occurs and is restricted within the early perceptual areas (local recurrent processing – LRP) extract high-level information that is sufficient to lead to some initial categorization of the visual scene and selective behavioral responses.

When a visual scene is being presented, the feedforward sweep reaches *V1* in about 40 ms. Multiple stimuli are all represented at this stage. The unconscious FFS extracts high-level information that could lead to categorization, and results in some initial feature detection. LRP produces further binding and segregation. The representations formed at this stage are restricted to information about spatiotemporal and surface properties (color, texture, orientation, motion, and perhaps to the affordances of objects), in addition to the representations of objects as bounded, solid entities that persist in space and time. (*Affordances* is the term *Gibson* [26.16] used to refer to the functional properties of objects (an object affords eating to an organism, grasping to an organism, etc.). *Clark* [26.25] defines *affordance* as “the possibilities for use, intervention and action which the physical world offers a given agent and are determined by the *fit* between the agent’s physical structure, capacities and skills and the action-related properties of the environment itself”. Affordances are directly perceivable by an organism in the sense that an object does not have to be classified as a member of a certain category in order for the organism to draw the conclusion, or use the relevant knowledge, that this object can be

used in a certain way by the organism; the organism just perceives the affordance, that is, the opportunity of action on this specific object. Affordances have two important properties. First, they are determined by the functional form of an object, that is, a combination of the object’s visible properties should suffice to determine whether this object has an affordance relative to some viewer. Affordances are based on certain invariant characteristics of the environment. Second, the affordance is always relative to the viewing organism; this is a consequence of the fact that affordances provide organisms with the opportunity to interact with objects in their environment. This interaction depends on the objects’ properties but it also depends on the needs and the constitution of the organism. A fly, for instance, affords eating to a frog but not to a human.)

At this level there are nonattentional selective mechanisms that prevent many stimuli from reaching awareness, even when attended to. Such stimuli are the high temporal and spatial frequencies, physical wavelength (instead of color), crowded or masked stimuli and so forth. FFS results in some initial feature detection. Then this information is fed forward to the extrastriate areas. When it reaches area *V4* recurrent processing occurs. Horizontal and recurrent processing allows interaction between the distributed information along the visual stream. At this stage, features start to bind and an initial coherent perceptual interpretation of the scene is provided. Initially, RP is limited to within visual areas; it is local. At this level one can be phenomenally aware of the content of perceptual states. At these intermediate levels there is already some competition between multiple stimuli, especially between close-by stimuli. The receptive fields that get larger and larger going upstream in the visual cortical cannot process all stimuli in full and crowding phenomena occur. Attentional selection intervenes to resolve this competition. Signals from higher cognitive centers and output areas intervene to modulate processing; this is global RP and signifies the inception of late vision.

*Lamme* [26.23, 24] discusses the nature of information that has achieved local recurrent embedding. He suggests that local RP may be the neural correlate of binding or perceptual organization. However, it is not clear whether at this preattentional stage the binding problem has been solved. The binding of some features, such as its color and shape, may require attention, while other feature combinations are detected preattentively. So, before attention has been allocated, the percept consists of only tentatively but uniquely bound features that form the proto-objects [26.26]. *Lamme* [26.24] argues that Marr’s  $2\frac{1}{2}$ D surface representation of objects and



their surface properties are extracted during the local RP stage. Other research [26.27] suggests that spatial relations are extracted at this recurrent stage. In addition motion and size are represented in cortical areas in which local RP take place.

It should be added that Marr thought of the 2½D sketch as the final product of a cognitively unaffected stage of visual processing, since, as we have seen, the formation of the 3-D sketch relies on semantic, conceptual knowledge. If, as is usually thought, cognitive effects on perception are mediated by cognitively-driven top-down attention, Lamme's proposal that early vision is not affected by this sort of attention echoes Marr's view that early vision is not affected by cognition and is thus CI, a view shared by *Pylyshyn* [26.28].

Current research (see [26.4] for a discussion) sheds light on the nature of inferences involved in the hypothesis testing implicated in early vision. Specifically, the top-down and lateral effects within early vision aim to test hypotheses concerning the putative distal causes of the sensory data encoded in the lower neuronal assemblies in the visual processing hierarchy. This testing assumes the form of matching predictions made on the basis of this hypothesis about the sensory information that the lower levels should encode assuming that the hypothesis is correct, with the current, actual sensory information encoded at the lower levels. Eventually, the hypothesis that best matches the sensory data is selected and the whole process of hypothesis selection can be construed as an abductive inference or inference to the best explanation, which could very well be carried through by Bayesian nets. One should note that this account of early vision shows that the standard constructivist theories of visual processing can be reconciled and greatly benefit from the recent conceptions of the brain as a generative, predictive machine.

There seems to be, however, a crucial discrepancy between the account of early vision presented here and Clark's account of generative hierarchical predictive models. It concerns the role of context, or previously acquired knowledge, in the formation of the working hypotheses and its direct consequence that because of this trait, visual perception and discursive thinking are inseparable. If early vision is restricted to processes occurring within the visual cortex and excludes any cognitive influences, then first, previous knowledge seems to play no role in the formation of the working hypotheses, and second, early vision does not involve any thinking since the latter requires the participation of the cognitive centers of the brain. Moreover, the representations in early vision are analogue-like, iconic and not symbolic and this entails that early vision cannot be some sort of discursive thinking since the latter operates on symbolic forms.

With respect to the first point, there is actually no real discrepancy. Recall that lateral and local recurrent processes play a fundamental role in the formation of the hypotheses that are constructed in early vision. Moreover, as we shall see in the next section, all visual processes including those of early vision, are restricted by certain principles, or better constraints, that reflect general regularities about the world and its geometry. Now, one could say that these constraints constitute a body of knowledge that informs early vision processing and affects early vision from the within and not in a top-down manner, since as we saw there are no cognitive top-down effects in early vision. This as we shall see, however, is misleading because these constraints do not constitute some form of knowledge that by affecting early vision renders it theory-laden, as Clark claims. Finally, early vision is also affected by associations of object properties that reflect statistical regularities in the environment and are stored in the early visual circuits through perceptual learning. I argue in the next section that these associations do not constitute a body of knowledge that affects early vision rendering it theory-laden. The lateral and local recurrent processes, the constraints, and the associations built in the early visual circuits constitute a rich context that contributes significantly to the formation of the working hypotheses that early vision neuronal assemblies construct to explain the sensory data at the lower processing levels. This context, however, does not involve any body of knowledge that renders perception theory-laden, as theories are traditionally understood.

As far as the second point is concerned, there is indeed a discrepancy because the account of early vision and Clark's views. Early vision, by being CI and conceptually encapsulated does not involve thinking and is radically different from thinking. In fact, as I argue in Sect. 26.5, not even late vision that involves concepts and is affected by the viewers' knowledge about the world is like thinking.

### 26.3.2 Late Vision

The conceptually modulated stage of visual processing is called late vision. Starting at 150–200ms, signals from higher executive centers including mnemonic circuits intervene and modulate perceptual processing in the visual cortex and this signals the onset of global recurrent processing (GRP). In 50ms low spatial frequency (LSF) information reaches the IT and in 100ms high spatial frequency (HSF) information reaches the same area. (LSF signals precede HSF signals. LSF information is transmitted through fast magnocellular pathways, while HSF information is transmitted through slower parvocellular pathways.)

Within 130 ms, parietal areas in the dorsal system but also areas in the ventral pathway (IT cortex) semantically process the LSF information and determine the gist of the scene based on stored knowledge that generates predictions about the most likely interpretation of the input. This information reenters the extrastriate visual areas and modulates (at about 150 ms) perceptual processing facilitating the analysis of HSF, for example by specifying certain cues in the image that might facilitate target identification [26.29–31]. Determining the gist may speed up the FFS of HSF by allowing faster processing of the pertinent cues, using top-down connections to preset neurons coding these cues at various levels of the visual pathway [26.32].

At about 150 ms, specific hypotheses regarding the identity of the object(s) in the scene are formed using HSF information in the visual brain and information from visual working memory (WM). The hypothesis is tested against the detailed iconic information stored in early visual circuits including V1. This testing requires that top-down signals reenter the early visual areas of the brain, and mainly V1. Indeed, evidence shows that V1 is reentered by signals from higher cognitive centers mediated by the effects of object- or feature-centered attention at 235 ms post-stimulus [26.33, 34]. This leads to the recognition of the object(s) in the visual scene. This occurs, as signaled by the P3 event-related-potentials (ERP) waveform, at about 300 ms in the IT cortex, whose neurons contribute to the integration of LSF and HSF information. (The P3 waveform is elicited about 250–600 ms and is generated in many areas in the brain and is associated with cognitive processing and the subjects' reports. P3 may signify the consolidation of the representation of the object(s) in working memory.)

A detailed analysis of the form that the hypothesis testing might take is provided by *Kosslyn* [26.35]. Note that one need not subscribe to some of the assumptions presupposed by Kosslyn's account, but these disagreements do not undermine the framework. Suppose that one sees an object. A retinotopic image is formed in the visual buffer, which is a set of visual areas in the occipital lobe that is organized retinotopically. An attentional window selects the input from a contiguous set of points for detailed processing. This is allowed by the spatial organization of the visual buffer. The information included in the attention window is sent to the dorsal and ventral system where different features of the image are processed. The ventral system retrieves the features of the object, whereas the dorsal system retrieves information about the location, orientation, and size of the object. Eventually, the shape, the color, and the texture of the object are registered in anterior portions of the ventral pathway. This information is transmitted to the

pattern activation subsystems in the IT cortex where the image is matched against representations stored there, and the compressed image representation of the object is thereby activated. This representation (which is a hypothesis regarding the identity, that is, class membership of an object) provides imagery feedback to the visual buffer where it is matched against the input image to test the hypothesis against the fine pictorial details registered in the retinotopical areas of the visual buffer. If the match is satisfactory, the category pattern activation subsystem sends the relevant pattern code to associative or WM, where the object is tentatively identified with the help of information arriving at the WM through the dorsal system (information about size, location, and orientation). Occasionally the match in the pattern activation subsystems is enough to select the appropriate representation in WM. On other occasions, the input to the ventral system does not match well a visual memory in the pattern activation subsystems. Then, a hypothesis is formed in WM. This hypothesis is tested with the help of other subsystems (including cognitive ones) that access representations of such objects and highlight their more distinctive feature. The information gathered shifts attention to a location in the image where an informative characteristic or an object's distinctive feature can be found, and the pattern code for it is sent to the patternactivation subsystem and to the visual buffer where a second cycle of matching commences.

Thus, the processes of late vision rely on recurrent interactions with areas outside the visual stream. This set of interactions is called *global recurrent processing* (GRP). In GRP, standing knowledge, i. e., information stored in the synaptic weights is activated and modulates visual processing that up to that point was conceptually encapsulated. During GRP the conceptualization of perception starts and the states formed have partly conceptual and eventually propositional contents. Thus, late vision involves a synergy of perceptual bottom-up processing and top-down processing, where knowledge from past experiences guides the formation of hypotheses about the identity of objects. This is the stage where the 3-D sketch (that is, the representation of an object as a volumetric structure independently of the viewer's perspective) is formed. This recovery cannot be purely data-driven since what is regarded as an object depends on the subsequent usage of the information and thus depends on the knowledge about objects. Seeing 3-D sketches is an instance of amodal completion, i. e., the representation of object parts that are not visible from the viewer's standpoint. (Amodal completion is the perception of the whole of an object or surface when only parts of it affect the sensory receptors. An object will be perceived as a complete volumetric structure even if

only part of it, namely, its facing surface, projects to the retina and thus is viewed by the viewer; it is perceived as possessing internal volume and hidden rear surfaces despite the fact that only some of its surfaces are exposed to view. Whether this perception involves visual awareness, in which case the brain completes the missing features through mental imagery, or visual understanding only, which means that the hidden features are not present in the phenomenology of the visual scene but are thought of, is a matter of debate. In amodal completion, one does not have a perceptual impression of the object's hidden features since the perceptual system does not fill in the missing features as happens in modal perception; the hidden features are not perceptually occurrent (see Appendix 26.D for a discussion of modal and amodal perception or completion).

One readily notices that Kosslyn's account of hypothesis testing naturally fits the schema of hierarchical generative predictive models as discussed in Clark [26.4]. The main themes of this schema are present in Kosslyn's account. These are: the generation of hypotheses at a higher level of visual processing, the crucial role of context or previously acquired knowledge in the formation of these hypotheses, and the testing of these hypotheses through their predictions against the rich iconic information stored in the lower

levels in the visual hierarchy. The whole process fits the scheme of an abductive inference or inference to the best explanation that could be carried out by means of Bayesian networks.

There is a marked difference between the abductive inferences involved in early vision and those involved in late vision; the latter but not the former are informed by knowledge properly speaking, that is, by information that is articulated in thought and thus contains concepts. This might tempt one to think that late vision may be akin to thought and thus that there is a stage of visual processing that has the most crucial traits of thinking, i. e., it involves discursive inferences justifying thus in part Clark's, Spelke's and others' belief to that effect. Against this, I am going to argue in Sect. 26.5, that late vision despite its being informed by conceptually articulated knowledge, differs in significant ways from thinking, the most important difference being that late vision does not engage in discursive inferences.

I have claimed that late vision constructs gradually a representation that best matches the visual scene through a set of processes that test a series of hypotheses by matching these hypotheses against stored iconic information. In other words, the output of late vision, a recognitional belief, is the result of an abductive inference.

## 26.4 Cognitive Penetrability of Perception and the Relation Between Early Vision and Thinking

In assessing claims relating perception to thinking and cognition, it is of paramount importance to examine the role that concepts play in modulating perceptual processing. This is so because if the processes of visual perception are the same as those that lead to belief formation, which means that perception and thinking are of the same nature and cannot be separated, then since belief formation is a process that requires the deployment of concepts, so should perception; perception should be conceptual through and through.

I have argued elsewhere [26.36] that early vision, the first stage of visual processing, is CI and conceptually encapsulated in the sense that its processes are not affected directly, that is, in an on-line manner from cognitive states. There are, as a matter of course, many indirect cognitive effects on early vision, such as precueing effects and the effects of spatial attention in its capacity as a determinant of the focus of gaze, but these effects do not constitute cases of genuine CP [26.36] because, first, concepts do not enter the content of the states of early vision although they causally affect it, and second, because of the preceding fact, these sorts

of cognitive effects can be mitigated and thus do not threaten the epistemological role of early vision as a neutral arbiter of perceptual judgments. If this view is correct, early vision being CI does not employ any concepts and thus it cannot be like thinking, which necessarily involves concepts.

One might object that this claim overlooks the possibility that concepts are embedded in the circuits subserving early vision, rendering it conceptual from the within as it were and not because of any top-down cognitive influences. Being conceptually affected and by using inferences, there is no obstacle in thinking of early vision as akin to thinking. This objection is reinforced by two empirical facts. First, as we have seen, all stages of visual processing are restricted by a set of principles or constraints that aim to solve the various problems of underdetermination. These principles contain concepts and exemplify some form of knowledge that renders early vision theory-laden; it follows that early vision can be like thinking owing to its inherent structure. Second, as a result of perceptual learning, many an environmental regularity are

learned and stored in the early visual circuits to facilitate the processing of familiar input. Since these associations could arguably be construed as involving concepts, a claim could be made that early vision is affected by concepts.

In what follows, I examine these two objections and argue that both sorts of phenomena do not signify the CP and theory-ladenness of perception. This is so because, first, they do not entail that there are any concepts embedded in early vision, and second, because it is doubtful whether they contain any representations. This is also important for the wider claim that visual perception is like thinking, since thinking necessarily involves inferences driven by representations of both premises and the rules of inferences. If it turns out, as I argue here, that the transformation rules that visual perception employs to process its states are not represented anywhere in the system, this would severely undermine the claim that perceptual inferences are the same as the inferences used in belief formation.

### 26.4.1 The Operational Constraints in Visual Processing

There is extensive evidence that there is an important *body of information* that affects perception not in a top-down manner but from within and this might be construed as evidence for the CP of visual perception from its inception. The perceptual system does not function independently of any kind of internal restrictions. Visual processing at every level is constrained by certain principles or rather operational constraints that modulate information processing. Such constraints are needed because distal objects are underdetermined by the retinal image, and because the percept itself is underdetermined by the retinal image. Unless the processing of information in the perceptual system is constrained by some *assumptions* about the physical world, perception is not feasible. Most computational accounts hold that these constraints substantiate some reliable generalities of the natural physical world as it relates to the physical constitution and the needs of the perceiving agents. There is evidence that the physiological visual mechanisms reflect these constraints. Their physical making is such that they implement these constraints, which are thus hardwired in perceptual systems (see Appendix 26.E for a list of some of these constraints).

These are Raftopoulos' [26.36] *operational constraints* and Burge's [26.37] *formation principles*. The operational constraints reflect higher-order physical regularities that govern the behavior of worldly objects and the geometry of the environment and which have been incorporated in the perceptual system through causal interaction with the environment over the evolu-

tion of the species. They allow us to lock onto medium size lumps of matter, by providing the discriminatory capacities necessary for the individuation and tracking of objects in a bottom-up way; they allow perception to generate perceptual states that present worldly objects as cohesive, bounded, solid, and spatiotemporally continuous entities.

The constraints are not available to introspection, function outside the realm of consciousness, and cannot be attributed as acts to the perceiver. One does not believe implicitly or explicitly that an object moves in continuous paths, that it persists in time, or that it is rigid, though one uses this information to parse and index the object. These constraints are not perceptually salient but one must be *sensitive* to them if one is to be described as perceiving their world. The constraints constitute the *modus operandi* of the perceptual system and not a set of rules used by the perceptual system as premises in perceptual inferences even though the *modus operandi* of the visual system consists of operations determined by laws describable in terms of computation principles. They are reflected in the functioning of the perceptual system and can be used only by it. They are available only for visual processing, whereas *theoretical* constraints are available for a wide range of cognitive applications. These constraints cannot be overridden since they are not under the perceiver's control; one cannot decide to substitute them with another body of constraints even if one knows that they lead to errors.

Being hardwired, the constraints are not even contentful states of the perceptual system. A state is formed through the spreading of activation and its modification as it passes through the synapses. The hardwired constraints specify the processing, i. e., the transformation from one state to another, but they are not the result of this processing. They are computational principles that describe transitions between states in the perceptual system. Although the states that are produced by means of these mathematical transformations have contents, there is no reason to suppose that the principles that specify the mathematical transformation operations are states of the system or contents of states in the system. If they are not states of the visual system, the principles that express them linguistically cannot be contents of any kind. Even though the perceptual system uses the operational constraints to represent some entity in the world and thus operates in accord with the principles reflected in the constraints (since the constraints are hardwired in the perceptual system, physiological conditions instantiate the constraints), the perceiver does not represent these principles or the constraints in any form. By the same token, these principles cannot be thought of as implicating concepts, since concepts are

representational. For this reason, perceptual operations should not be construed as inference rules, although they are describable as such, and they do not constitute either a body of knowledge or some theory about the world.

Recent work on Bayesian causal networks [26.4] draws a picture of the brain as a Bayesian net operating at various space and time scales, and suggests that there is a sharp distinction between internal probabilistic dependencies that can be explained by internal causal connections and those that cannot. Only those that cannot be explained internally carry information about the external world. Applying this to the case of the neuronal mechanisms that implement the operational constraints at work in visual processing, one could say that these mechanisms perform transformations that depend entirely on the internal probabilistic dependencies in the system as they are determined by the hardwired circuitry that realizes the internal causal connections and thus there is nothing representational about them.

These considerations allow us to address *Cavanagh's* [26.38] claim that the processes that lead to the formation of a conscious percept constitute *visual cognition* in virtue of their use of inferences. The construction of a percept is “the task of visual cognition and, in almost all cases, each construct is a choice among an infinity of possibilities, chosen based on likelihood, bias, or a whim, but chosen by rejecting other valid competitors” [26.38]. This process is an inference in that “it is not a guess. It is a rule-based extension from partial data to the most appropriate solution”; in the terminology of this chapter, the selection process is an abduction.

According to *Cavanagh* [26.38], for inference to take place the visual system should not rely to purely bottom-up analyses of the image that use only retinal information, such as sequences of filters that underlies facial recognition, or the cooperative networks that converge on the best descriptions of surfaces and contours. Instead, the visual system should use some object knowledge, which is nonretinal, context-dependent information. By *object knowledge* *Cavanagh* means any sort of nonretinal information that may be needed for the filling in that leads to the construction of the percept. This knowledge consists of rules that guide or constrain visual processing in order to solve the underdetermination problem that I mentioned above; they provide the rule-based extension from partial data that constitutes an inference. These rules do not influence visual processing in a top-down way, since they reside within the visual system; they are “from the side” [26.39].

The discussion concerning the nature of the operational constraints suggests that, their crucial role

in perceptual processing notwithstanding, these constraints do not justify *Cavanagh's* characterization of visual perception as *visual cognition*, if cognition is thought of as involving discursive inferences.

## 26.4.2 Perceptual Learning

Evidence from studies showing early object classification effects suggests that to the extent that object classification presupposes object knowledge, this knowledge affects early vision in a top-down manner rendering it theory-laden. Moreover, even if one could show that these effects do not entail the CP of early vision, one could argue that since perceptual learning affects the way one sees the world, some experiences are learned and form memories that are stored in visual memory and affect perceptual processing from its inception. Our experiences shape the way we see the world.

Indeed, visual memories affect perception. Familiarity with objects or scenes that is built through repeated exposure to objects or scenes (sometimes one presentation is enough), or even repetition memory [26.40] facilitate search, affect figure from ground segmentation, speed up object identification and image classification, etc. [26.41–43].

Familiarity can affect visual processing in different ways. It may facilitate object identification and categorization, which are processes that take time since their final stage occurs between 300–600 ms after stimulus onset as is evidenced by the P3 responses in the brain, but their earlier stage starts about 150 ms after stimulus onset [26.44–46]. One notices that familiarity intervenes during the latest stage of visual processing (300–360 ms). These effects involve the higher cognitive levels of the brain at which semantic information and processing, both being required for object identification and categorization, occur [26.30]. In this sense, these sort of familiarity effects do not threaten the CI of early vision, which has ended about 120 ms after stimulus onset.

Familiarity, including repetition memory, also affects object classification (whether an image portrays an animal or a face), a process that occurs in short latencies (95–100 ms and 85–95 ms respectively) [26.47–49]. These early effects may pose a threat to the CI of early vision since they cannot be considered post-sensory. The threat would materialize should the classification processes either require semantic information to intervene or require the representations of objects in working memory to be activated, since that would, too, mean conceptual involvement.

Researchers however unanimously agree that the early classification effects in the brain result from the FFS and do not involve top-down semantic information,

nor do they require the activation of object memories. The brain areas involved are low-level visual areas (including the FEF – front eye fields) from V1 to no higher than V4 [26.48] or perhaps a bit more upstream to posterior IT [26.42] and lateral occipital complex (LOC) [26.49].

The early effects of familiarity may be explained by invoking contextual associations (target-context spatial relationships) that are stored in early sensory areas to form unconscious perceptual memories [26.50] which, when activated from incoming signals that bear the same or similar target-context spatial relationships, modify the FFS of neural activity resulting in the facilitating effects mentioned above. Thus, what is involved in the phenomenon are certain associations built in the early visual system that once activated speed up the feedforward processing. This is a case of rigging-up the early visual processing; it is not a case of top-down cognitive effects on early visual processing.

The early effects may also be explained by appealing to configurations of properties of objects or scenes. Currently, neurophysiological research [26.40, 49], psychological research [26.42], and computation modeling [26.51] suggest that what is stored in early visual areas are implicit associations representing fragments of objects and shapes, or *edge complexes*, as opposed to whole objects and shapes. One of the reasons that have led researchers to argue that it is object and shape fragments that are used in rapid classifications instead of whole objects and shapes is the following: If these associations reflecting some sort of object recognition can affect figure-ground segmentation as we have reasons to believe [26.42] in view of the fact that figure-ground segmentation occurs very early (80–100 ms) [26.52] these associations must be stored in early visual areas (up to V4, LO and posterior IT) and cannot be the representations stored in, say, anterior IT. The earlier visual areas store object and shape fragments and not holistic figures and shapes [26.40, 51].

The associations that are built in, through learning, in early visual circuits reflect in essence the statistical distribution of properties in environmental scenes [26.32, 53]. The statistical differences in physi-

cal properties of different subsets of images are detected very early by the visual system before any top-down semantic involvement as is evidenced by the elicitation of an early deflection in the differential between animal-target and nontarget ERP's at about 98 ms (in the occipital lobe) and 120 ms (in the frontal lobe). The low-cues could be retrieved very early in the visual system from a scene by analyzing the energy distribution across a set of orientation and spatial frequency-tuned channels [26.54]. This suggests that the rapid image classification may rely on low-level, or intermediate-level cues [26.51] that act diagnostically, that is, they allow the visual system to predict the gist of the scene and classify images very fast. These cues may be provided by coarse visual information, say by low-level spatial frequency information and thus the visual system does not have to rely on high-level fully integrated object representations in order to be able to classify rapidly visual scenes.

It follows that the classification of an object that occurs very early during the fast FFS at about 85–100 ms is due to associations regarding shape and object fragments stored in early visual areas and does not reflect any top-down cognitive effects on, that is, the CP of, early vision. Thus, early object classification is not a sign of the theory-ladenness of early vision, since the knowledge about the world does not affect it in a top-down manner.

To recapitulate the results of our discussion in this section, I have argued that neither the operational constraints operating in visual perception, nor perceptual learning entail that concepts affect early vision. Moreover, they do not entail that visual processing in general is theory-laden because of the role of these constraints, since they are not representational elements and any theory constitutively implicates representational elements. On the other hand, both the operational constraints and the effects of perceptual learning provide the context in which early vision constructs its hypotheses, and part of the context in which late vision operate, the other part being the viewer's knowledge of the world, which, as I have said, affects late vision but not early vision.

## 26.5 Late Vision, Inferences, and Thinking

*Jackendoff* [26.55] distinguishes visual awareness from visual understanding. There is a qualitative difference between the experience of a 3-D sketch and the experience of a  $2\frac{1}{2}$ -D sketch. Although one is in some sense aware of the 3-D sketch or of category-based representations, however, this is not visual awareness but some

other kind of awareness. Visual awareness is awareness of Marr's  $2\frac{1}{2}$ -D sketch, which is the viewer-centered representation of the visible surfaces of objects, while the awareness of the 3-D sketch is visual understanding. Thus, the 3-D sketch, which includes the unseen surfaces that are not represented in the  $2\frac{1}{2}$ -D sketch,

is a result of an inference. These views belong to the belief-based account of amodal completion: the 3-D sketch is the result of beliefs abductively inferred from the object's visible features and other background information from past experiences (see Appendix 26.D for an explanation of amodal and modal completion or perception).

The problem is whether the object identification that occurs in late vision (which, as we have seen most likely constitutes in essence an abductive inference) and depends on concepts should be thought of as a purely visual process or as a case of discursive understanding involving discursive inferences. If late vision involves conceptual contents and if the role of concepts and stored knowledge consists of providing some initial interpretation of the visual scene and in forming hypotheses about the identity of objects that are tested against perceptual information, one is tempted to say that this stage relies on inferences and thus differs in essence from the purely perceptual processes of early vision. Perhaps it would be better to construe late vision as a discursive stage involving thoughts, in the way of epistemic seeing, where *seeing* is used in a metaphorical nonperceptual sense, as where one says of his friend whom she visited *I see he has left*, based on perceptual evidence [26.56]. It is, also possible that Dretske [26.57, 58] thinks that seeing in the doxastic sense is not a visual but rather a discursive stage.

One might object, first, that abandoning this usage of *to see* violates ordinary usage. A fundamental ingredient of visual experience consists of meaningful 3-D solid objects. Adopting this proposal would mean that one should resist talking of seeing tigers and start talking about seeing viewer-centered visible surfaces. "By this criterion, much of the information we normally take to be visually conscious would not be, including the 3-D shape of objects as well as their categorical identity" [26.59].

More to the point, I think that one should not assume either that late vision involves abductive inferences construed as inferential discursive-state transformations that constitutively involve thoughts in the capacity of premises in inferences whose conclusion is a recognitional belief, or that late vision consists of discursively entertaining thoughts; if thinking is construed as constitutively implicating discursive argumentation, visual perception is different from thinking in some radical ways. The reason is twofold. First, seeing an object is not the result of a discursive inference, that is, a movement in thought from some premises to a conclusion, even though it involves concepts and intrastate transformations. Second, late vision is a stage in which conceptual modulation and perceptual processes form an inextricable link that differentiates late vision from

discursive stages and renders it a different sort of a set of processes than understanding, even though late vision involves implicit beliefs regarding objects that guide the formation of hypotheses concerning object identity, and an explicit belief of the form *that O is F* eventually arises in the final stages of late vision. Late vision has an irreducible visual ingredient that makes it different from discursive understanding.

Let me clarify two terminological issues. First, judgments are occurrent states, whereas beliefs are dispositional states. To judge that *O is F* is to predicate *F*-ness to *O* while endorsing the predication [26.60]. To believe that *O is F* is to be disposed to judge under the right circumstances that *O is F*. This is one sense in which beliefs are dispositional items. There is also a distinction between standing knowledge (information stored in long term memory, LTM) and information that is activated in working memory (WM). The belief that *O is F* may be a standing information in LTM, a memory about *O* even though presently one does not have an occurrent thought about *O*. Beliefs need not be consciously or unconsciously apprehended, that is, activated in the mind, in order to be possessed by a subject, which means that beliefs are dispositional rather than occurrent items; this is a second sense in which beliefs are dispositional. When this information is activated, the thought that *O is F* emerges in WM; all thoughts are occurrent states.

It follows that a belief qua dispositional state may be either a piece of standing knowledge, in which case it is dispositional in the sense that when activated it becomes a thought, or a thought that awaits endorsement to become a judgment, in which case the belief is dispositional in the sense that it has the capacity to become a judgment. In the first case, beliefs differ from thoughts. In the second case, a belief is a thought held in WM, albeit one that has not been yet endorsed. In what follows, I assume that beliefs are either pieces of standing information or thoughts that have not been endorsed and thus are not judgments. Finally, by *implicit belief* I mean the belief held by a person who is not aware that she is having that belief.

As I said in the introduction, this chapter examines whether the abductive processes that take place in late vision should be construed as discursive inferences. Specifically, my claim is that the processes in late vision are not inferential processes where *inference* is understood as discursive, that is, as a process that involves drawing propositions or conclusions from other propositions, that are represented in the system, acting as premises by applying (explicitly or implicitly) inferential rules that are also represented. As we saw, these inferences are distinguished from *inferences* as understood by vision scientists according to whom

any transformation of signals carrying information according to some rule is an inference.

### 26.5.1 Late Vision, Hypothesis Testing, and Inference

I think that the states of late vision are not inferences from premises that include the contents of early vision states, even though it is usual to find claims that one infers that a tiger, for example, is present from the perceptual information retrieved from a visual scene. An inference relates some propositions in the form of premises with some other proposition, the conclusion. However, the objects and properties as they are represented in early vision do not constitute contents in the form of propositions, since they are part of the nonpropositional, iconic nonconceptual content of perception. In late vision, the perceptual content is conceptualized but the conceptualization is not a kind of inference but rather the application of stored concepts to some input that enters the cognitive centers of the brain and activates concepts by matching their content. Thus, even though the states in late vision are formed through the synergy of bottom-up visual information and top-down conceptual influences, they are not inferences from perceptual content.

Late vision involves hypotheses regarding the identity of objects and their testing against the sensory information stored in iconic memory. One might think that inferences are involved since testing hypotheses is an inferential process even though it is not an inference from perceptual content to a recognitional thought. It is, rather, an argument of the form of: if  $A$  and  $B$  then (conclusion)  $C$ , where  $A$  and  $B$  are background assumptions and the hypothesis regarding the identity of an object respectively, and  $C$  is the set of visual features that the object is likely to have.  $A$  consists of implicit beliefs about the features of the hypothesized visual object. If the predicted visual features of  $C$  match those that are stored in iconic memory in the visual areas, then the hypothesis about the identity of the object is likely correct. The process ends when the best possible fit is achieved. However, the test basis or evidence against which these hypotheses are tested for a match, that is, the iconic information stored in the sensory visual areas, is not a set of propositions but patterns of neuronal activations whose content is nonpropositional.

There is nothing inference-like in this matching. It is just a comparison between the activations of neuronal assemblies that encode the visual features in the scene and the activations of the neuronal assemblies that are activated top-down from the hypotheses. If the same assemblies are activated then there is a match. If they are not, the hypothesis fails to pass the test.

This can be done through purely associational processes of the sort employed, say, in connectionist networks that process information according to rules and thus can be thought of as instantiating processing rules, without either representing these rules or operating on language-like symbolic representations. Such networks perform vector completion and function by satisfying soft constraints in order to produce the best output given the input into the system and the task at hand. Note that the algebraic and thus continuous nature of state transformations in neural networks, as opposed to the algorithmic discrete-like operations of classical AI (which assumes that the brain is a syntactic machine that processes discrete symbols according to rules that are also represented in the system) suits best the analogue nature of iconic representations.

In perceptual systems construed as neural networks, the fundamental representational unit is not some linguistic or linguistic-like entity but the activation pattern across a proprietary population of neurons. If one wishes to understand the workings of the visual brain, one should eschew sentences and propositions as bearers of representations and meanings and reconceptualize representations as activation patterns. This does not mean, of course, that the brain does not have symbolic representations but only that, first, these are a subset of the representations that the brain uses in its various functions, and, second and most importantly, the symbolic representations are constructed somehow out of the more fundamental context-dependent representations that the brain uses and are, consequently, a later construct, phylogenetically speaking. This has an important corollary for any theory of cognition that employs activation patterns as the fundamental units of representation, namely, that it must be able to explain the existence and usage of symbolic representations. This means also that the processing at work in the brain, that is, the transformation of the representational units to other representational units is not exclusively the transformation of complex or simple symbols by means of a set of syntactic rules as in the algorithms that, according to the classical view, the brain is supposed to run. Instead, it can be the algebraic transformation of activation patterns (in essence the algebraic transformations from one multidimensional matrix or tensor to another). The transformation is effected by the synaptic connections among the neurons as the signal passes from one layer to another. These connections have weights that constitute a filter through which the signal is transformed as it passes through.

The above also explain the holistic nature of the abductive visual processes that classical cognitive theories (the family of theories that assume that the brain is a syntactic machine that processes symbols that are



constant, context independent, and freely repeatable elements) have failed to capture. It is interesting that if I am right, Fodor's attempt to differentiate the perceptual systems from cognitive functions in order to protect the former from the abductive holistic reasoning implicated in the latter fails since late vision is abductive and holistic as well.

Since discursive inferences are carried out through rules that are represented in the system and operate on symbolic structures, the processing in a connectionist network does not involve discursive inferences, although it can be described in terms of inference making. Thus, even though seeing an object in late vision involves the application of concepts that unify the appearances of the object and of its features under some category, it is not an inferential process.

I have said that the noninferential process that results in the formation of a recognitional thought or belief can be recast in the form of an argument from some premise to a conclusion. However, this does not entail that the formation of the perceptual thought is a piece of reasoning, that is, a transition from a set of premises that act as a reason for holding the thought to the thought itself. Admittedly, the perceiver can be asked on what grounds she holds the thought that *O* is *F*, in which case she may reply *because I saw it or I saw that O is F*. However, this does not mean that the reason she cites as a justification of her thought is a premise from which she inferred the thought. The perceiver does not argue from her thought *I saw it to be thus and so* to the thought *It is thus and so*. She just forms the thought on the basis of the evidence included in her relevant perceptual state in the noninferential way I described above. What warrants the recognitional thought *O is F* is not the thought held by the perceiver that she sees *O* to be *F* but the perceptual state that presents to her the world as being such and such. "When one knows something to be so by virtue of seeing to be so, one's warrant for believing it to be so is that one sees it to be so, not one's believing that one sees it to be so" [26.57].

*Spelke* [26.3] who echoes *Rock's* [26.2] views that the perceptual system combines inferential information to form the percept (for example, from visual angle and distance information, one infers and perceives size) – argues "perceiving objects may be more akin to thinking about the physical world than to sensing the immediate environment". The reason is that the perceptual system, to solve the underdetermination problem of both the distal object from the retinal image and of the percept from the retinal image, employs a set of object principles and that reflect the geometry and the physics of our environment. Since the contents of these principles consist of concepts, and thus the principles can be thought of as some form of knowledge about the world,

perception engages in discursive, inferential processes. Against this, I argued above that the processes that constrain the operations of the visual system should not be construed as discursive inferences. They are hardwired in the perceptual circuits and are not represented in it. Thus, perceptual operations should not be construed as inference rules, although they are describable in terms of discursive inferential rules. It follows that the abduction that takes place in late vision is not an Aristotelian inference; it is better described by the ampliative vector completion of connectionism.

### 26.5.2 Late Vision and Discursive Understanding

Even if I am right that seeing in late vision is not the result of a discursive abductive inference but the result of a pattern-matching process that ensures the best fit with the available data, it is still arguable that late vision should be better construed as a stage of discursive understanding rather than as a visual stage. If object recognition involves forming a belief about class membership, even if the belief is not the result of an inference, why not say that recognizing an object is an experience-based belief that is a case of understanding rather than vision?

#### Late Vision Is more than Object Recognition

A first problem with this view is that late vision involves more than a recognitional belief. Suppose that *S* sees an animal and recognizes it as a tiger. In the parallel preattentive early vision, the proto-object that corresponds to the tiger is being represented amongst the other objects in the scene. After the proto-objects have been parsed, the object recognition system forms hypotheses regarding their identity. However, for the subject's confidence to reach the threshold that will allow her to form beliefs about the identity of the objects and report it, these hypotheses must be tested [26.61].

For this to happen, the relevant sensory activations enter the parietal and temporal lobes, and the prefrontal cortex, where the neuronal assemblies encoding the information about the objects in the scene are activated and the relevant hypotheses are formed. To test these hypotheses, the visual system allocates resources to features and regions that would confirm or disconfirm the hypotheses. To accomplish this, activation spreads through top-down signals from the cognitive centers to the visual areas of the brain where the visual sensory memory and the fragile visual memory store the proto-objects extracted from the visual scene. This way, conceptual information about the tiger affects visual processing and after some hypothesis testing the animal is recognized as a tiger through the synergy of

visual circuits and WM. At this point the explicit belief *O is F* is formed. This occurs after 300 ms, when the viewer consolidates the object in WM and identifies it with enough confidence to report it, which means that beliefs are formed at the final phases of late vision.

However, semantic modulation of visual processing and the process of conceptualization that eventually leads to object recognition starts at about 130–200 ms. There is thus a time gap between the onset of conceptualization and the recognition of an object, which is a prerequisite for the formation of an explicit recognitional belief. As *Treisman* and *Kanwisher* [26.62] observe, although the formation of hypotheses regarding the categorization of objects can occur within 130–200 ms after stimulus onset, it takes another 100 ms for subsequent processes to bring this information into awareness so that the perceiver could be aware of the presence of an object. To form the recognitional belief that *O is F*, one must be aware of the presence of an object token and construct first a coherent representation. This requires the enhancement through attentional modulation of the visual responses in early visual circuits that encode rich sensory information in order to integrate them into a coherent representation, which is why beliefs are delayed in time compared with the onset of conceptualization; not all of late vision involves explicit beliefs.

#### Late Vision as a Synergy of Bottom-Up and Top-Down Information Processing

A second reason why the beliefs formed in late vision are partly visual constructs and not pure thoughts is that the late stage of late vision in which explicit beliefs concerning object identity are formed constitutively involves visual circuits (that is, brain areas from LGN to IT in the ventral system). Pure thought, on the other hand, involves an amodal form of representation formed in higher centers of the brain, even though these amodal representations can trigger in a top-down manner the formation of mental images and can be triggered by sensory stimulation. The point is that amodal representations can be activated without a concomitant activation of the visual cortex. The representations in late vision, in contrast, are modal since they constitutively involve visual areas. Thus, what distinguishes late vision beliefs from pure thoughts is mostly the fact that the beliefs in late vision are formed through a synergy of bottom-up and top-down activation and their maintenance requires the active participation of the visual circuits. Pure thoughts can be activated and maintained in the absence of activation in visual circuits.

The constitutive reliance of late vision on the visual circuits suggests that late vision relies on the presence of the object of perception; it cannot cease to function

as a perceptual demonstrative that refers to the object of perception, as this has been individuated through the processes of early vision. As such, late vision is constitutively context dependent since the demonstration of the perceptual particular is always context dependent. Thought, on the other hand, by its use of context independent symbols, is free of the particular perceptual context. Even though recognitional beliefs in late vision and pure perceptual beliefs involve concepts, the concepts function differently in the two contexts [26.37]:

“Perceptual belief makes use of the singular and attributive elements in perception. In perceptual belief, pure attribution is separated from, and supplements, attributive guidance of contextually purported reference to particulars. Correct conceptualization of a perceptual attributive involves taking over the perceptual attributive’s range of applicability and making use of its (perceptual) mode of presentation.”

The attributive and singular elements in perception correspond to the perceived objects and their properties respectively. The attributive elements or properties guide the contextual reference to particulars or objects since the referent in a demonstrative perceptual reference is fixed through the properties of the referent as these properties are presented in perception.

Concepts enter the game in their capacity as pure attributions that make use of the perceptual mode of presentation. Burge’s claim that in perceptual beliefs pure attributions supplement attributions that are used for contextual reference to particulars may be read to mean that perceptual beliefs are hybrid states involving both visual elements (the contextual attributions used for determining reference to objects and their properties) and conceptualizations of these perceptual attributives in the form of pure attributions. In this case, the role of perceptual attributives is ineliminable. In late vision, unlike in pure beliefs, there can be no case of pure attribution, that is, of attribution of features in the absence of perceptually relevant particulars since the attributions are used to single out these particulars.

The inextricable link between thought and perception in late vision explains the essentially contextual [26.63, 64] character of beliefs in late vision. The proposition expressed by the belief cannot be detached from the perceptual context in which it is believed and cannot be reduced to another belief in which some third person or objective content is substituted for the indexicals that figure in the thought (in the way one can substitute via Kaplan’s characters the indexical terms with their referents and get the *objective* truth-evaluable content of the belief); the belief is tied to a idiosyncratic

viewpoint by making use of the viewer's physical presence and occupation of a certain location in space and time; the context in which the indexical thought is believed is essential to the information conveyed.

The discussion on late vision and the inferences it uses to construct the percept suggests that late vision, its conceptual nature notwithstanding, does not involve discursive inferences and in this sense is fundamentally different from thinking, if the latter is thought to implicate constitutively discursive inferences. Late vision employs abductive inferences, in that it constructs the representation that best fits the sensory image, but these inferences are not the result of the application of rules that are represented in the system. Even the operational constraints that restrict visual processing in late vision and could be thought of as transformation rules that the system follows to make inferences, are not, as we have seen, propositional structures or even representations in the brain. The inferences involved are informed and guided by conceptual information in pattern-matching processes but fall short of being discursive inferences.

## 26.6 Concluding Discussion

I have argued in this chapter that visual processing involves abductive inferences that aim to construct a representation, namely, the percept, that best matches the sensory information. To achieve this, the brain probably uses Bayesian strategies since abductive inferences are probabilistic in nature. I also argued that these inferences are not discursive inferences and since the latter are the characteristic trait of thinking, visual processing is not akin to thinking despite its usage of abductive inferences; my claim applies to both early vision and late vision.

The discussion in this chapter, and especially the view on the relation between perceptual inference and perceptual judgments, is in agreement with *Magnani's* [26.65] elaboration on *Peirce's* [26.66, 67] views on visual perception, which Peirce also conceived of perception as an abductive inference. In particular, I have tried to defend the thesis eloquently expressed by Peirce that the transition between abductive inferences and perceptual judgment is a continuous one without any sharp line of demarcation between them despite their many differences that I elaborated on in the previous section. My discussion also reinforces *Magnani's* view that "judgments in perception are fallible but indubitable abductions we are not in any condition to psychologically conceive that they are false, as they are unconscious habits of inference" [26.65]. Most importantly, my account of the abductive inferences in-

The fact that both conceptual and nonconceptual representations are in essence activation patterns allows us to understand how conceptual, symbolic information and nonconceptual iconic information could interact. The main difference between the two forms of representations is that the former are not homogeneous and have a syntactic structure that has a canonical decomposition, whereas the latter are homogeneous and lack a canonical decomposition. To appreciate the difference think of it in following way: the fact that a symbolic representation has a canonical decomposition means that not every subpart of the representation is a representation; only those subparts that satisfy the syntactic rules of the representational systems are symbols or representations. The expression (p&Q), for instance, is a symbol or a representation, but the expression (p(&q)) is not. Any subpart of an image, on the other hand, is an image and thus a representation.

The output of late vision, namely the percept, enters the space of reasons and participates in discursive inferences and thus in thought.

Involved in visual perception fully justifies *Magnani's* claim that visual abduction is not sentential, that is, it does not employ symbolic, or discursive as I have called them, inferences. Instead it relies on pattern matching in which activation patterns that take on continuous values are compared. Thus, the representational medium employed is analogue and not symbolic in nature and the usage of stored knowledge in drawing inferences resembles more the use of models that put the incoming information in a context so that conclusions could be drawn rather than the recruitment of sentences and inference rules. In other words, visual abduction is model-based.

In constructing the percept, the brain uses a set of operational constraints that aim to solve the various underdetermination problems that the visual perception encounters in order to construct the percept. I have argued that these constraints should not be thought of as rules that are represented in the system or that have some representational contents and guide the perceptual inferences rendering them discursive. Instead, they are hardwired in the visual system and are not representations.

I also suggested, although I did not discuss this issue in full, that the recent developments in vision studies tend to bring together the different theories of vision by showing the points of contact between them, rather than to underline their differences.

To recapitulate, the main conclusion of this chapter is, first, that to the extent that thinking is associated with the use of discursive inferences, perception differs radically from thinking. If the meaning of thinking is extended to comprise nondiscursive inferences, the claim may be made that perception is thinking. In this case, however, a distinction should be drawn between discursive thinking that characterizes cognition and nondiscursive thinking that characterizes perceptual processes. Second, if thinking also necessarily involves the deployment of concepts, then there is a stage of visual processing, namely, early vision, which is not akin to thinking since its contents are nonconceptual. The other stage of visual processing, namely late vision, uses conceptual information. Since, as I will argue, the processes of late vision are not discursive inferences,

**Table 26.1** Visual perception and thinking

Perception	Thinking	
	Thinking narrow	Thinking wide
Early vision	No	Yes/no concepts
Late vision	No	Yes/yes concepts

if thinking is conceived as necessarily implicating discursive inferences late vision is not akin to thinking, notwithstanding the conceptual involvement. If the concept of thinking is extended to include other sorts of inferences, such as the model-based abductive inferences discussed in this chapter, late vision could be thought of as a sort of thinking, which, unlike early vision, implicates concepts (see Table 26.1 for a taxonomy).

## 26.A Appendix: Forms of Inferences

These are the three forms of inferences in which all syllogisms can be categorized.

### 26.A.1 Deduction

An inference is deductive if its logical structure is such that the conclusion of the inference is a logical consequence of the premises of the inference. This entails that if the premises of a deductive argument are true then its conclusion is necessarily true as well. In this sense, deductive arguments are truth preserving. This is equivalent to saying that in any interpretation of the inference in which the premises are true, the conclusion is true too. Differently put, if an argument is deductively valid, there is no model under which the premises are true but the conclusion is false. This is why deductive inferences are sometimes characterized as conclusive.

A typical example of a deductive argument is this: All men are mortal; Socrates is a man. Therefore Socrates is mortal.

### 26.A.2 Induction

An argument is inductive if its conclusion does not follow logically from the premises. The premises of an inductive argument may be true and still its conclusion false. The premises of an inductive argument provide epistemic support or epistemic warrant for its conclusion; they constitute evidence for the conclusion. By definition, inductive arguments are not truth preserving.

A typical example of an inductive argument is the following: Bird  $\alpha$  is a crow and is black; bird  $\beta$  is a crow and is black; . . . bird  $\kappa$  is a crow and is black. Therefore: All crows are probably black.

If the examined specimens are found in a variety of places and under different environmental conditions, the premises of the inference provide solid evidence for the conclusion. Yet, the conclusion may still be wrong since the next crow that we will examine may not be black. This example shows that the conclusion does not follow logically from the premises. It is still possible, no matter how good the premises, that is the evidence, are that the conclusion be false, which explains the qualification *probably* in the conclusion of an inductive argument. The world could be such that even crows  $\alpha$  through  $\kappa$  are black, crow  $\kappa + 1$  is white. For this reason inductions are considered to be nonconclusive but tentative [26.68].

### 26.A.3 Abduction or Inference to the Best Explanation

It is an inference in which a series of facts, which are either new, or improbable, or surprising on their own or in conjunction, are used as premises leading to a conclusion that constitutes an explanation of these facts. This explanation makes them more probable and more comprehensible in that it accounts for their appearance. As such, with abductive inferences the mind reaches conclusions that go far beyond what is given. For this reason, abductions are the main theoretical tools for building models and theories that explain reality. Ab-

duction is inductive since it is ampliative, does not preserve truth and is thus probabilistic in that the conclusion is tentative.

#### 26.A.4 Differences Between the Modes of Inference

##### Induction versus Deduction

Induction is an ampliative inference, whereas deduction is not ampliative. This means that the information conveyed by the conclusion of an inductive argument goes beyond the information conveyed by the premises and, in this sense, the conclusion is not implicitly contained in the premises. In deduction, the conclusion is implicitly contained in the premises and the inference just makes it explicit. If all men are mortal and Socrates is a man, for example, the fact that Socrates is mortal is implicitly contained in these two propositions. What the deduction does is to render it explicit in the form of the conclusion. When we deduce that Socrates is mortal, our knowledge does not extend that which we already knew; it only makes it explicit. When, on the other hand, we inductively infer that all crows are probably black from the premise that all the specimens of crows that we have examined thus far are black, we extend the scope of our knowledge because the conclusion concerns all crows and not just the crows thus far examined.

The above discussion entails the main difference between deductive and inductive arguments. Deductive arguments are monotonous, while inductive arguments are not. This means that a valid deductive argument remains valid no matter how many premises we add to the argument. The reason is that the validity of the deductive argument presupposes that the conclusion is a logical conclusion of its premises. This fact does not change by the addition of new premises, no matter what these premises stipulate and thus the deductive argument remains valid. Things are radically different in induction. A new premise may change the conclusion even if the previous premises strongly supported the conclusion. For example, if we discover that crow  $\kappa + 1$  is white, this undermines that previously drawn and well-supported conclusion that all crows are black.

### 26.B Appendix: Constructivism

Some of Marr's particular proposals of his model have been criticized on many grounds (see, for example, [26.59]). In particular, against Marr's model of object recognition, it has been argued by several researchers that object recognition may be more image-based than based on object-centered representations,

##### Induction versus Abduction

Both abduction and induction are tentative forms of inference in that they do not warrant the truth of their conclusion even if the premises are true. They are, also, both ampliative in that the conclusion introduces information that was not contained implicitly in the premises. As we have seen, in abduction one aims to explain or account for a set of data. Induction is a more general form of inference. When, for instance, one successfully tests a hypothesis by making predictions that are borne out, the predicted data provide inductive, but not abductive, support for the hypothesis. In general, the evaluation phase in hypothesis, or theory, construction is considered to be inductive. Conceiving the explanatory hypothesis, on the other hand, is an abductive process that may assume the form of a pure, educated guess that need not have involved any previous testing. In this case, the abductively produced hypothesis is not, a priori, the best explanation for the set of data that need explanation; this is one of the occasions in which abduction can be distinguished from the inference to the best explanation. However, it should be stressed, although I do not have the space to elaborate on this problem, that in realistic scientific practice abduction as theory construction could not be separated from the evaluative inductive phase since they both form an inextricable link. This justifies the claim that abduction is an inference to the best explanation.

A further difference between abduction and induction is that even though both kinds of inference are ampliative, in abduction the conclusion may, and usually does, contain terms that do not figure in the premises. Almost all theoretical entities in science were conceived as a result of abduction. The nucleus of an atom, for example, was posited as a way of explaining the scattering of particles after the bombardment of atoms. Nowhere in the premises of the abductive argument was the notion of an atom present; the evidence consisted in measurements of the deviation of the pathways particles from their predicted values after the bombardment. The conclusion *all crows are probably black*, on the other hand, contains only terms that are available in the premises.

which means that the latter may be less important than Marr thought them to be. Neurophysiological studies [26.69] also suggest that both object-centered and viewer-centered representations play a substantial role in object recognition. Nevertheless, his general ideas about the construction of gradual visual representations

remain useful. According to this form of constructivism, vision consists of four stages, each of which outputs a different kind of visual representation:

1. The *formation of the retinal image*; the immediate stimulus for vision, that is the first stimulus that affects directly the sensory organs (this is called the proximal stimulus) is the pair of two-dimensional (2-D) images projected from the environment to the eyes. This representation is based on a 2-D retinal organization. At this stage, the information impinging on the retina (which as you may recall concerns intensity of illumination and wavelengths, and which is captured by the retinal receptors) is organized so that all of the information about the spatial distribution of light (i. e., the light intensity falling on each retinal receptor) be recast in a reference frame that consists of square image elements (*pixels*), each indicating with a numerical value the light intensity falling on each receptor. Sometimes, the processes of this stage are called *sensation*.
2. The *image-based stage*; it includes operations that receive as input the retinal image (that is, the numerical array of values of light intensities in each pixel) and process it in order to detect local edges and lines, to link these edges and lines in a more global scale, to match up corresponding images in the two eyes, to define 2-D regions in the image, and to detect line terminations and blobs. This stage outputs 2-D surfaces at some particular slant that are located at some distance from the viewer in 3-D space.

In general, the image-based representation has the following properties: First, it receives as input and thus operates first on information about the 2-D structure of the retinal image rather than on information concerning the physical, distal, objects. Second, its geometry is inherently two-dimensional. Third, the image-based representation of the 2-D features is cast in a coordinate reference system that is defined with respect to the retina (as a result, the organization of the information is called *retino-topic*). This means that the axes of the reference system are aligned with the eye rather than the body or the environment. This stage is the first stage of *perception* proper:

3. The *surface-based*; in this stage, vision constructs representations of the intrinsic properties of sur-

faces in the environment that might have produced the features constructed in the image-based model. At this stage, and in contradistinction to the preceding stage, the information about the worldly surfaces is represented in three dimensions. Marr's two-and-a-half-dimensional (2.5-D) sketch is a typical example of a surface-based representation. Note that the surface-based representation of a visual scene does not contain information about all the surfaces that are present in the scene, but only those that are visible for the viewer's current viewpoint.

In general, the surface-based representation has the following properties: First, The elements that the surface-based stage outputs consist of the output of the image-based stage, that is, in 2-D surfaces at some particular slant that are located at some distance from the viewer in 3-D space. Second, these 2-D surfaces are represented within a 3-D spatial framework. Third, the aforementioned reference framework is defined in terms of the direction and distance of the surfaces from the observer's standpoint (it is egocentric):

4. The *object-based*; this is the stage in which the visual system constructs 3-D representations of objects that include at least some of the occluded surfaces of the objects, that is, the surfaces that are invisible from the standpoint of the viewer, such as the back parts of objects. In this sense, this is the stage in which explicit representations of whole objects in the environment are constructed. It goes without saying that in order for the visual system to achieve this aim, it must use information about the whole objects that viewers have stored from their previous visual encounters with objects of the same type. The viewer retrieves from memory this information and fills in with it the surface-based image constructed at the previous stage.

In general, the object-based representation has the following properties: First, this stage outputs volumetric representations of objects that may include information about unseen surfaces. Second, the space in which these objects are represented is three-dimensional. Third, the frame of reference in which the object-based representations are cast is defined in terms of the intrinsic structure of the objects and the visual scene (it is scene-based or allocentric).

## 26.C Appendix: Bayes' Theorem and Some of Its Epistemological Aspects

Bayes' theorem is the following probabilistic formula (in its simple form because there is another formulation when one considers two competing hypotheses), where  $A$  is a hypothesis purporting to explain a set of data  $B$

$$P(A/B) = P(B/A)P(A)/P(B),$$

where  $P(A)$  is the prior probability, that is, the initial degree of belief in  $A$ ;  $P(A/B)$  is the conditional probability of  $A$  given  $B$ , or posterior probability, that is, the degree of belief in  $A$  after taking into consideration  $B$ ;  $P(B)$  is the probability of  $B$ .  $P(B/A)$  is the likelihood of  $B$  given  $A$ , that is, the degree of belief that  $B$  is true given that  $A$  is true. The ratio  $P(B/A)/P(B)$  represents the degree of support that  $B$  provides for  $A$ .

Suppose that  $B$  is the sensory information encoded by a neuronal assembly at level  $l-1$ , and  $A$  is the hypothesis that the neuronal assembly at level  $l$  posits as an explanation of  $B$ . Bayes' theorem tells us that the probability that  $A$  is true, that is, the probability that level  $l$  represent a true pattern in the environment given the sensory data  $B$ , depends first on the prior probability of hypothesis  $A$ , that is the probability of  $A$  before the predictions of  $A$  are tested. This prior probability depends on both the incoming signal to  $l$  but, also and most crucially because many different causes could have caused the incoming signal, on the contextual effects because these are the factors that determine which is the most

likely explanation of the data among the various possible alternative accounts.

The probability of  $A$  also depends on the  $P(B/A)$ , that is, the probability that  $B$  be true given  $A$ . This reflects a significant epistemological insight, namely, that since a correct account of a set of data explains away, these data are a *natural consequence* of the explaining hypothesis, or naturally fit into the conceptual framework created by the hypothesis. The various gravity phenomena, for instance, become very plausible in view of the law of gravity; they are not so much so if the hypothesis purporting to explain these same phenomena involves some accidents of nature, even if they are systematic. To put in a reverse way, if gravity exists, then the probability that unsupported objects will fall down is greater than the probability of these objects falling down if some other hypothesis is postulated to explain the fall of unsupported objects.

The probability of the hypothesis  $A$  depends inversely on the probability of the data  $B$ . Since probabilities take values from (0 to 1), the smaller the probability in the denominator, that is, the more surprising and thus improbable  $B$  is, the greater the probability that  $A$  be true given  $B$ . This part of the equation also reflects an important epistemological insight, namely that the more surprising a set of data is, the more likely is to be true a hypothesis that successfully explains them. Finally, the ratio  $P(B/A)/P(B)$  expresses the support  $B$  provides to  $A$  in the sense that the greater this ratio, the greater the probability that the hypothesis  $A$  is true.

## 26.D Appendix: Modal and Amodal Completion or Perception

There are two sorts of completion. In modal completion the viewer has a distinct visual impression of a hidden contour or other hidden features even though these features are not occurrent sensory features. The perceptual system fills in the missing features, which thus become as phenomenally occurrent as the occurrent sensory features of the object.

In amodal completion, one does not have a perceptual impression of the object's hidden features since the perceptual system does not fill in the missing features as it happens in modal perception, although as we shall see mental imagery can fill in the missing phenomenology; the hidden features are not perceptually occurrent.

There are cases of amodal perception that are purely perceptual, that is, bottom-up. In these cases, although no direct signals from the hidden features impinge on the retina (there is no local information available), the

perceptual system can extract information regarding them from the global information contained in the visual scene without any cognitive involvement, as the resistance of the ensuing percepts to beliefs indicates. However, in such cases, the hidden features are not perceived. One simply has the visual impression of a single concrete object that is partially occluded and not the visual impression of various disparate image regions. Therefore, in these perceptually driven amodal completions there is no mental imagery involved, since no top-down signals from cognitive areas are required for the completion, and since the hidden features are not phenomenologically present.

There are also cases of amodal completion that are cognitively driven, such as the formation of the 3-D sketch of an object, in which the hidden features of the object are represented through the top-down acti-

vation of the visual cortex from the cognitive centers of the brain. In some of these cases, top-down processes activate the early visual areas and fill in the missing features that become phenomenologically present. In

other cases of cognitively driven amodal completion, the viewer simply forms a pure thought concerning the hidden structure in the absence of any activation of the visual areas and thus in the absence of mental imagery.

## 26.E Appendix: Operational Constraints in Visual Processing

Studies by [26.3, 70–72] show that infants, almost from the very beginning, are constrained by a number of domain-specific principles about material objects and some of their properties. As *Karmiloff-Smith* [26.72] remarks, these constraints involve “attention biases toward particular inputs and a certain number of principled predispositions constraining the computation of those inputs”. Such predispositions are the conception of object persistence, and four basic principles (boundness, cohesion, rigidity, and no action at a distance).

The *cohesion principle*: “two surface points lie on the same object only if the points are linked by a path of connected surface points”. This entails that if some relative motion alters the adjacency relations among points at their borders, the surfaces lie on distinct objects, and that “all points on an object move on connected paths over space and time. When surface points appear at different places and times such that no connected path could unite their appearances, the surface points do not lie on the same object”.

According to the *boundness principle* “two surface points lie on distinct objects only if no path of connected surface points links them”. This principle determines the set of those points that define an object boundary and entails that two distinct objects cannot interpenetrate, because two distinct bodies cannot occupy the same place at the same time.

Finally the *rigidity* and *no action at a distance* principles specify that bodies move rigidly (unless the other mechanisms show that a seemingly unique body is, in fact, a set of two distinct bodies) and that they move independently of one another (unless the mechanisms show that two seemingly separate objects are in fact connected).

Further studies shed light on the nature of these principles or constraints and on the neuronal mechanisms that may realize them. There is evidence that the physiological mechanisms underlying vision reflect these constraints; their physical making is such that they implement these constraints, from cells for edge detection to mechanisms implementing the epipolar constraint [26.73, 74]. Thus, one might claim that these principles are hardwired in our perceptual system.

The formation of the *full primal sketch* in *Marr's* [26.12] theory relies upon the principles of *local proximity* (adjacent elements are combined) and of *similarity* (similarly oriented elements are combined). It also relies upon [26.20] the more general principle of *closure* (two edge-segments could be joined even though their contrasts differ because of illumination effects).

Other principles used by early visual processing to solve the problem of the underdetermination of perception by the retinal image are those of *continuity* (the shapes of natural objects tend to vary smoothly and usually do not have abrupt discontinuities), *proximity* (since matter is cohesive, adjacent regions usually belong together and remain so even when the object moves), and *similarity* (since the same kind of surface absorbs and reflects light in the same way the different subregions of an object are likely to look similar).

The formation of the  $2\frac{1}{2}$ D *sketch* is similarly underdetermined, in that there is a great deal of ambiguity in matching features between the two images form in the retinas of the two eyes, since there is usually more than one possible match. Stereopsis requires a unique matching, which means that the matching processing must be constrained. The formation of the  $2\frac{1}{2}$ D *sketch*, therefore, relies upon a different set of operational constraints that guide stereopsis. “A given point on a physical surface has a unique position in space at some time” [26.69] and matter is cohesive and surfaces are generally smooth. These operational constraints give rise to the general constraints of *compatibility* (a pair of image elements are matched together if they are physically similar, since they originate from the same point of the surface of an object), of *uniqueness* (an item from one image matches with only one item from the other image), and of *continuity* (disparities must vary smoothly). Another constraint posited by all models of stereopsis is the *epipolar* constraint (the viewing geometry is known). *Mayhew* and *Frisby's* [26.75] account of stereopsis posits some additional constraints, most notably, the principle of *figural continuity*, according to which figural relationships are used to eliminate most of alternative candidate matches between the two images.



## References

- 26.1 H. von Helmholtz: *Treatise on Psychological Optics* (Dover, New York 1878/ 1925)
- 26.2 I. Rock: *The Logic of Perception* (MIT Press, Cambridge 1983)
- 26.3 E.S. Spelke: Object perception. In: *Readings in Philosophy and Cognitive Science*, ed. by A.I. Goldman (MIT Press, Cambridge 1988)
- 26.4 A. Clark: Whatever next? Predictive brains, situated agents, and the future of cognitive science, *Behav. Brain Sci.* **36**, 181–253 (2013)
- 26.5 M. Rescorla: The causal relevance of content to computation, *Philos. Phenomenol. Res.* **88**(1), 173–208 (2014)
- 26.6 L. Shams, U.R. Beierholm: Causal inference in perception, *Trends Cogn. Sci. (Regul. Ed.)* **14**, 425–432 (2010)
- 26.7 N. Orlandi: *The Innocent Eye: Why Vision Is not a Cognitive Process* (Oxford Univ. Press, Oxford 2014)
- 26.8 P. Lipton: *Inference to the Best Explanation*, 2nd edn. (Routledge, London, New York 2004)
- 26.9 D.G. Campos: On the Distinction between Peirce's Abduction and Lipton's inference to the best explanation, *Synthese* **180**, 419–442 (2011)
- 26.10 G. Minnameier: Peirce-suit of Truth-why inference to the best explanation and abduction ought not to be confused, *Erkenntnis* **60**, 75–105 (2004)
- 26.11 G. Harman: Enumerative induction as inference to the best explanation, *J. Philos.* **68**(18), 529–533 (1965)
- 26.12 D. Marr: *Vision: A Computational Investigation into Human Representation and Processing of Visual Information* (Freeman, San Francisco 1982)
- 26.13 J. Biederman: Recognition by components: A theory of human image understanding, *Psychol. Rev.* **94**, 115–147 (1987)
- 26.14 A. Johnston: Object constancy in face processing: Intermediate representations and object forms, *Ir. J. Psychol.* **13**, 425–438 (1992)
- 26.15 G.W. Humphreys, V. Bruce: *Visual cognition: Computational, Experimental and Neuropsychological Perspectives* (Lawrence Erlbaum, Hove 1989)
- 26.16 J.J. Gibson: *The Ecological Approach to Visual Perception* (Houghton-Mifflin, Boston 1979)
- 26.17 J. Fodor, Z. Pylyshyn: How direct is visual perception? Some reflections on Gibson's 'Ecological Approach, *Cognition* **9**, 139–196 (1981)
- 26.18 M. Rowlands: *The New Science of Mind: From Extended Mind to Embodied Phenomenology* (MIT Press, Cambridge 2010)
- 26.19 J. Norman: Two visual systems and two theories of perception: An attempt to reconcile the constructivist and ecological approaches, *Behav. Brain Sci.* **25**, 73–144 (2002)
- 26.20 V. Bruce, P.R. Green: *Visual Perception: Physiology, Psychology and Ecology*, 2nd edn. (Lawrence Erlbaum, Hillsdale 1993)
- 26.21 A. Noe: *Action in Perception* (MIT Press, Cambridge 2004)
- 26.22 J. Pearl: *Causality: Models, Reasoning and Inference* (Cambridge Univ. Press, Cambridge 2009)
- 26.23 V.A.F. Lamme: Why visual attention and awareness are different, *Trends Cogn. Sci.* **7**, 12–18 (2003)
- 26.24 V.A.F. Lamme: Independent neural definitions of visual awareness and attention. In: *The Cognitive Penetrability of Perception: An Interdisciplinary Approach*, ed. by A. Raftopoulos (Nova-Science Books, Hauppauge 2004)
- 26.25 A. Clark: An embodied cognitive science?, *Trends Cogn. Sci.* **3**(9), 345–351 (1999)
- 26.26 P. Vecera: Toward a biased competition account of object-based segmentation and attention, *Brain Mind* **1**, 353–384 (2000)
- 26.27 E.C. Hildreth, S. Ulmann: The computational study of vision. In: *Foundations of Cognitive Science*, ed. by M.I. Posner (MIT Press, Cambridge 1989)
- 26.28 Z. Pylyshyn: Is vision continuous with cognition? The case for cognitive impenetrability of visual perception, *Behav. Brain Sci.* **22**, 341–423 (1999)
- 26.29 M. Barr: The proactive brain: Memory for predictions, *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **364**, 1235–1243 (2009)
- 26.30 K. Kihara, Y. Takeda: Time course of the integration of spatial frequency-based information in natural scenes, *Vis. Res.* **50**, 2158–2162 (2010)
- 26.31 C. Peyrin, C.M. Michel, S. Schwartz, G. Thut, M. Seghier, T. Landis, C. Marendaz, P. Vuilleumier: The neural processes and timing of top-down processes during coarse-to-fine categorization of visual scenes: A combined fMRI and ERP study, *J. Cogn. Neurosci.* **22**, 2678–2780 (2010)
- 26.32 A. Delorme, G.A. Rousselet, M.J.-M. Macé, M. Fabre-Thorpe: Interaction of top-down and bottom up processing in the fast visual analysis of natural scenes, *Cogn. Brain Res.* **19**, 103–113 (2004)
- 26.33 L. Chelazzi, E. Miller, J. Duncan, R. Desimone: A neural basis for visual search in inferior temporal cortex, *Nature* **363**, 345–347 (1993)
- 26.34 P.R. Roelfsema, V.A.F. Lamme, H. Spekreijse: Object-based attention in the primary visual cortex of the macaque monkey, *Nature* **395**, 376–381 (1998)
- 26.35 S.M. Kosslyn: *Image and Brain* (MIT Press, Cambridge 1994)
- 26.36 A. Raftopoulos: *Cognition and Perception: How Do Psychology and the Neural Sciences Inform Philosophy?* (MIT Press, Cambridge 2009)
- 26.37 T. Burge: *Origins of Objectivity* (Clarendon Press, Oxford 2010)
- 26.38 P. Cavanagh: Visual cognition, *Vis. Res.* **51**, 1538–1551 (2011)
- 26.39 R. Gregory: *Concepts and Mechanisms of Perception* (Charles Scribners and Sons, New York 1974)
- 26.40 K. Grill-Spector, T. Kushnir, T. Hendler, S. Edelman, Y. Itzhak, R. Malach: A sequence of object-processing stages revealed by fMRI in the human occipital lobe, *Human Brain Mapping* **6**, 316–328 (1998)
- 26.41 H. Liu, Y. Agam, J.R. Madsen, G. Krelman: Timing, timing, timing: Fast decoding of object information from intracranial field potentials in human visual cortex, *Neuron* **62**, 281–290 (2009)

- 26.42 M. Peterson: Overlapping partial configurations in object memory. In: *Perception of Faces, Objects, and Scenes: Analytic and Holistic Processes*, ed. by M. Peterson, G. Rhodes (Oxford Univ. Press, New York 2003)
- 26.43 M. Peterson, J. Enns: The edge complex: Implicit memory for figure assignment in shape perception, *Percept. Psychophys.* **67**(4), 727–740 (2005)
- 26.44 M. Fabre-Thorpe, A. Delorme, C. Marlot, S. Thorpe: A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes, *J. Cogn. Neurosci.* **13**(2), 171–180 (2001)
- 26.45 J.S. Johnson, B.A. Olshausen: The earliest EEG signatures of object recognition in a cued-target task are postsensory, *J. Vis.* **5**, 299–312 (2005)
- 26.46 S. Thorpe, D. Fize, C. Marlot: Speed of processing in the human visual system, *Nature* **381**, 520–522 (1996)
- 26.47 S.M. Crouzet, H. Kirchner, S.J. Thorpe: Fast saccades toward faces: Face detection in just 100 ms, *J. Vis.* **10**(4), 1–17 (2010)
- 26.48 H. Kirchner, S.J. Thorpe: Ultra-rapid object detection with saccadic movements: Visual processing speed revisited, *Vis. Res.* **46**, 1762–1776 (2006)
- 26.49 K. Grill-Spector, R. Henson, A. Martin: Repetition and the brain: Neural models of stimulus-specific effects, *Trends Cogn. Sci.* **10**, 14–23 (2006)
- 26.50 M. Chaumon, V. Drouet, C. Tallon-Baudry: Unconscious associative memory affects visual processing before 100 ms, *J. Vis.* **8**(3), 1–10 (2008)
- 26.51 S. Ullman, M. Vidal-Naquet, E. Sali: Visual features of intermediate complexity and their use in classification, *Nat. Neurosci.* **5**(7), 682–687 (2002)
- 26.52 V.A.F. Lamme, H. Super, R. Landman, P.R. Roelfsema, H. Spekreijse: The role of primary visual cortex (V1) in visual awareness, *Vis. Res.* **40**(10–12), 1507–1521 (2000)
- 26.53 R. VanRullen, S.J. Thorpe: The time course of visual processing: From early perception to decision making, *J. Cogn. Neurosci.* **13**, 454–461 (2001)
- 26.54 A. Torralba, A. Oliva: Statistics of natural image categories, *Network* **14**, 391–412 (2013)
- 26.55 R. Jackendoff: *Consciousness and the Computational Mind* (MIT Press, Cambridge 1989)
- 26.56 F. Jackson: *Perception: A Representative Theory* (Cambridge Univ. Press, Cambridge 1977)
- 26.57 F. Dretske: Conscious experience, *Mind* **102**, 263–283 (1993)
- 26.58 F. Dretske: *Naturalizing the Mind* (MIT Press, Cambridge 1995)
- 26.59 S.E. Palmer: *Vision Science: Photons to Phenomenology* (MIT Press, Cambridge 1999)
- 26.60 J. McDowell: *Mind and World* (Harvard Univ. Press, Cambridge 2004)
- 26.61 A. Treisman: How the deployment of attention determines what we see, *Vis. Cogn.* **14**, 411–443 (2006)
- 26.62 A. Treisman, N.G. Kanwisher: Perceiving visually presented objects: Recognition, awareness, and modularity, *Curr. Opin. Neurobiol.* **8**, 218–226 (1998)
- 26.63 J. Perry: *Knowledge, Possibility, and Consciousness*, 2nd edn. (MIT Press, Cambridge 2001)
- 26.64 R.C. Stalnaker: *Our Knowledge of the Internal World* (Clarendon Press, Oxford 2008)
- 26.65 L. Magnani: *Abductive Cognition. The Epistemological and Eco-Cognitive Dimensions of Hypothetical Reasoning* (Springer, Berlin 2009)
- 26.66 C.S. Peirce: Perceptual judgments (1902. In: *Philosophical Writings of Peirce*, ed. by J. Buchler (Dover, New York 1955)
- 26.67 C.S. Peirce, N. Houser: *The Essential Peirce: Selected Philosophical Writings*, Vol. 2 (Indiana Univ. Press, Bloomington 1998)
- 26.68 S. Toulmin: *The Uses of Argument* (Cambridge Univ. Press, Cambridge 1958)
- 26.69 D.I. Perrett, M.W. Oram, J.K. Hietanen, P.J. Benson: Issues of representation in object vision. In: *The Neuropsychology of Higher Vision: Collated Tutorial Essays*, ed. by M.J. Farah, G. Ratcliff (Lawrence Erlbaum, Hillsdale 1994)
- 26.70 E.S. Spelke, R. Kestenbaum, D.J. Simons, D. Wein: Spatio-temporal continuity, smoothness of motion and object identity in infancy, *Br. J. Dev. Psychol.* **13**, 113–142 (1995)
- 26.71 E.S. Spelke: Principles of object perception, *Cogn. Sci.* **14**, 29–56 (1990)
- 26.72 A. Karmiloff-Smith: *Beyond Modularity: A Developmental Perspective on Cognitive Science* (MIT Press, Cambridge 1992)
- 26.73 G.F. Poggio, W.H. Talbot: Mechanisms of static and dynamic stereopsis in foveal cortex of the rhesus monkey, *J. Physiol.* **315**, 469–492 (1981)
- 26.74 D. Ferster: A comparison of binocular depth mechanisms in areas 17 and 18 of the cat visual cortex, *J. Physiol.* **311**, 623–655 (1981)
- 26.75 J.F.W. Mayhew, J.P. Frisby: Psychophysical and computational studies towards a theory of human stereopsis, *Artif. Intell.* **17**, 349–385 (1981)

# Diagrammatic Reasoning

William Bechtel

Diagrams figure prominently in human reasoning, especially in science. Cognitive science research has provided important insights into the inferences afforded by diagrams and revealed differences in the reasoning made possible by physically instantiated diagrams and merely imagined ones. In scientific practice, diagrams figure prominently both in the way scientists reason about data and in how they conceptualize explanatory mechanisms.

To identify patterns in data, scientists often graph it. While some graph formats, such as line graphs, are used widely, scientists often develop specialized formats designed to reveal specific types of patterns and not infrequently employ multiple formats to present the same data, a practice illustrated with graph formats developed in circadian biology. Cognitive scientists have revealed the spatial reasoning and iterative search processes scientists deploy in understanding graphs.

In developing explanations, scientists commonly diagram mechanisms they take to be

27.1	<b>Cognitive Affordances of Diagrams and Visual Images</b> .....	606
27.2	<b>Reasoning with Data Graphs</b> .....	608
27.2.1	Data Graphs in Circadian Biology .....	608
27.2.2	Cognitive Science Research Relevant to Reasoning with Graphs .....	611
27.3	<b>Reasoning with Mechanism Diagrams</b> .....	613
27.3.1	Mechanism Diagrams in Circadian Biology .....	613
27.3.2	Cognitive Science Research Relevant to Reasoning with Mechanism Diagrams .....	615
27.4	<b>Conclusions and Future Tasks</b> .....	616
	<b>References</b> .....	617

responsible for a phenomenon, a practice again illustrated with diagrams of circadian mechanisms. Cognitive science research has revealed how reasoners mentally animate such diagrams to understand how a mechanism generates a phenomenon.

Human reasoning is often presented as a mental activity in which we apply inference rules to mentally represented sentences. In the nineteenth century, Boole presented the rules for natural deduction in logic as formalizing the rules of thought. Even as cognitive scientists moved beyond rules of logical inference as characterizing the operations of the mind, they tended to retain the idea that cognitive operations apply to representations that are encoded in the mind (e.g., in neural activity). But in fact humans often reason by constructing, manipulating, and responding to external representations, and this applies as well to deductive as to abductive and inductive reasoning. Moreover, these representations are not limited to those of language but include diagrams. While reliance on diagrams extends far beyond science, it is particularly important in science. Scientific papers and talks are replete with diagrams and these are often the primary focus as scien-

tists read papers and engage in further reasoning about them. *Zacks et al.* [27.1] determined that the number of graphs in scientific journals doubled between 1984 and 1994. One would expect that trend has continued. Although many journals now limit the number of figures that can appear in the published paper, they have increasing allowed authors to post supplemental material, which often includes many additional diagrams.

Scientists clearly use diagrams to communicate their results with others. But there is also evidence that they make extensive use of these diagrams in their own thinking – in developing an understanding of the phenomenon to be explained and in advancing an explanation of it. Diagrams also figure prominently in the processes through which scientists analyze data. Since far less attention has been paid, both in philosophy of science and in the cognitive sciences, to how diagrams figure in reasoning activities, my objec-

tive in this chapter is to characterize what is known about how people, including scientists, reason with diagrams.

An important feature of diagrams is that they are processed by the visual system, which in primates is a very highly developed system for extracting and relating information received by the eyes (approximately one-third of the cerebral cortex is employed in visual processing). I begin in Sect. 27.1 by focusing on the distinctive potential of diagrams to support reasoning by enabling people to employ visual processing to detect specific patterns and organize together relevant pieces of information and examine the question of whether images constructed in one's imagination work equally well. In this chapter, I employ the terms *diagram* in its inclusive sense in which it involves marks arranged in

a two or more dimensional layout where the marks are intended to stand for entities or activities or information extracted from them and the geometrical relations between the marks are intended to convey relations between the things represented. In Sects. 27.2 and 27.3, I will discuss separately two types of diagrams that I designate data graphs and mechanism diagrams. In each case I introduce the discussion with examples from one field of biological research, that on circadian rhythms – the endogenously generate oscillations with a period of approximately 24 h that are entrainable to the light-dark cycle of our planet and that regulate a wide range of physiological activities. I then draw upon cognitive science research relevant to understanding how people reason with each type of diagram and relate this to the diagrams used in the science.

## 27.1 Cognitive Affordances of Diagrams and Visual Images

Two different traditions have dominated cognitive science research on vision. One, associated with Marr [27.2], has emphasized how, from the activation of individual neurons in the retina, people can build up a representation of what is seen. The other, advanced by Gibson [27.3], drew attention to the rich information, often highly structured, available to the visual system. The latter is especially relevant to addressing diagrams, since they involve structured perceptual objects in the environment. A key theoretical claim Gibson advanced was that different objects of perception afford different activities for different organisms – the back of a chair affords landing for an insect but draping a garment for humans. One can extend the account of affordance to external representations, and so focus, as Zhang [27.4] does, on how different representations activate different cognitive operations [27.4, pp. 185–186]:

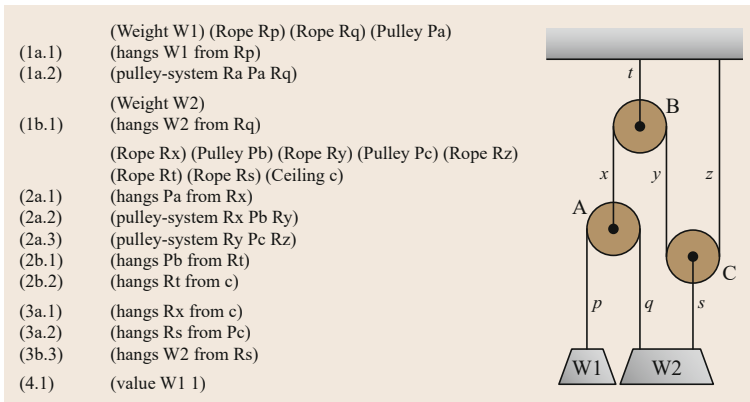
“Different representations activate different operations, not vice versa. It follows that operations are representation-specific. External representations activate perceptual operations, such as searching for objects that have a common shape and inspecting whether three objects lie on a straight line. In addition, external representations may have invariant information that can be directly perceived [...] such as whether several objects are spatially symmetrical to each other and whether one group has the same number of objects as another group. Internal representations activate cognitive operations, such as adding several numbers to get the sum.”

To investigate how diagrams afford different reasoning than other representations, Zhang compared a game formally equivalent to tic-tac-toe in which players pick

numbers from the pool 1 through 9 with the objective of being the first to pick three numbers totaling 15. Representing the numbers on a tic-tac-toe board (Fig. 27.1) shows that the two games are formally equivalent – all sequences of three numbers totaling 15 can be mapped onto a winning solution to tic-tac-toe and vice versa. Despite being formally equivalent, the tic-tac-toe board representation engages different cognitive operations than the number game represented as picking numbers from a pool. On the tic-tac-toe board, players can identify winning combinations by detecting lines but in the number variant they must perform arithmetic over many sets of numbers. In Zhang's experiments, humans played against a computer, which always made the first play and was programmed never to lose. If participants chose the best moves, however, they could always gain a tie. Participants required much longer to figure out a strategy to tie the computer when playing the number version than traditional tic-tac-toe, indicating that they deployed different operations in the two games. (See [27.5], for experiments showing similar results with variants of the Tower of Hanoi problem that placed

4	3	8
9	5	1
2	7	6

**Fig. 27.1** The game of picking three numbers that add to 15 is mapped onto a tic-tac-toe board, establishing their formal equivalence



**Fig. 27.2** Larkin and Simon's pulley problem presented in sentential form on the left and in a diagram on the right (after [27.6])

different demands on internal processes.) Zhang further claims that by limiting winning strategies to lines, traditional tic-tac-toe reduces the cognitive demands, freeing up cognitive resources for other activities.

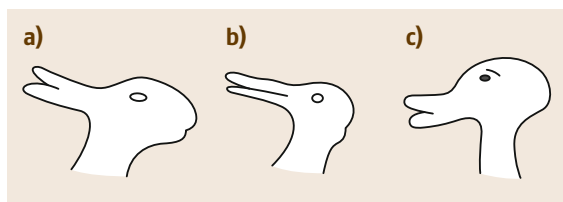
In a provocative pioneering study addressing the question *why a diagram is (sometimes) worth 10 000 words?* Larkin and Simon [27.6] also focused on how diagrammatic representations support different cognitive operations than sentential representations. Like Zhang, they focused on representations that were equivalent in the information they provided but turned out not to be computationally equivalent in the sense that inferences that could be “drawn easily and quickly from the information given explicitly in the one” could not be drawn easily and quickly from the other. (Kulvicki [27.7] speaks in terms of information being extractable where there is a feature of a representation that is responsible for it representing a given content and nothing more specific than that. This helpfully focuses on the issue of how the representation is structured, but does not draw out the equally important point that extracting information depends on the cognitive processes that the cognizer employs.)

One of the problems Larkin and Simon investigate is the pulley problem shown in Fig. 27.2, where the task is to find the ratio of weights at which the system is in equilibrium. They developed a set of rules to solve the problem. The advantage of the pulley diagram on their analysis is that it locates information needed to apply particular rules at nearby locations in the diagram so that by directing attention to a location a person can secure the needed information. In the sentential representation the information needed for applying rules was dispersed so that the reasoner would need to conduct multiple searches. In a second example, involving a geometry proof, Larkin and Simon show how a diagram reduces both the search and recognition demands, where recognition utilizes the resources of the visual system to retrieve information. The authors also of-

fer three examples of diagrams used in economics and physics, graphs and vector diagrams, that employ not actual space but dimensions mapped to space and argue that they too provide the benefits in search and recognition.

Together, these two studies make clear that diagrams differ from other representations in terms of the cognitive operations they elicit in problem-solving situations. Most generally, diagrams as visual structures elicit pattern detection capacities whereas sentential representations require linguistic processing. Larkin and Simon note that a common response to a complex sentential description is to draw a diagram. An interesting question is whether comparable results can be obtained by mentally imagining diagrams. Pioneering studies by Shepard [27.8, 9] and Kosslyn et al. [27.10] demonstrated that people can rotate or move their attention across a mentally encoded image. But quite surprisingly Chambers and Reisberg [27.11] found that this capacity is severely limited. They presented Jastrow's duck-rabbit (Fig. 27.3) to participants sufficiently briefly that they could only form one interpretation of the figure. They then asked the participants if they could find another interpretation while imaging the figure. None were able to do so even when offered guidance. Yet, when they were allowed to draw a figure based on their mental image, all participants readily discovered the alternative interpretation.

These findings inspired numerous other investigations into the human ability to work with mental images whose results present a complex pattern. Reed and Johnsen [27.12] reached a similar conclusion as Chambers and Reisberg when they asked participants to employ imagery to determine whether a figure was contained in a figure they had previously studied. Yet when Finke et al. [27.13] asked participants to construct in imagery complex images from components, they performed well. Studies by Finke and Slayton [27.14] showed that many participants were able to generate



**Fig. 27.3** (a) The version of the duck-rabbit figure used as a stimulus in Chambers and Reisberg's experiments is shown on the left. The other two versions were drawn by participants based on their own image interpreted as a rabbit (b) and as a duck (c). From these they were readily able to discover the other interpretation, something they could not do from their mental image alone. With permission from the American Psychological Association, (after [27.11])

creative images from simple shapes in imagery (the drawings the participants produced were independently assessed for creativity). *Anderson and Helstrup* [27.15, 16] set out to explore whether drawing enhanced performance on such tasks and their conclusions were largely negative – participants produced more images, but the probability of generating ones judged creative was not increased: “These results were contrary to the initial belief, shared by most experimenters and subjects alike, that the use of pencil and paper to construct patterns should facilitate performance.”

*Verstijnen et al.* [27.17] explored whether the failure of drawing to improve performance might be due

to insufficient training in drawing. Using a task similar to that of Reed and Johnson, they compared those without formal training in drawing with design students who had 2 years of courses in drawing, and found those with training in drawing performed much better. In another study in which participants were required to create new objects from simple components, *Verstijnen et al.* [27.18] found that drawing significantly helped trained drawers create compound objects that involved restructuring the components (e.g., changing proportions within the component). One conclusion suggested by these results is that reasoning with diagrams may be a learned activity. Humans spend a great deal of time learning to read and write, and even then further education is often required to extract information from text and construct and evaluate linguistic arguments. Yet, perhaps because vision seems so natural, we assume that diagrams are automatically interpretable and except in curricula in fields like design, we provide no systematic education in constructing and reasoning with diagrams. Accordingly, it perhaps should not be a surprise that science educators have found that students often ignore the diagrams in their textbooks [27.19]. One of the challenges in teaching students how to reasoning with diagrams is identifying what cognitive operations people must perform with different types of diagrams. Cognitive scientists have begun to identify some of these operations, and I will discuss some of these in the context of data graphs and mechanism diagrams in the next two sections.

## 27.2 Reasoning with Data Graphs

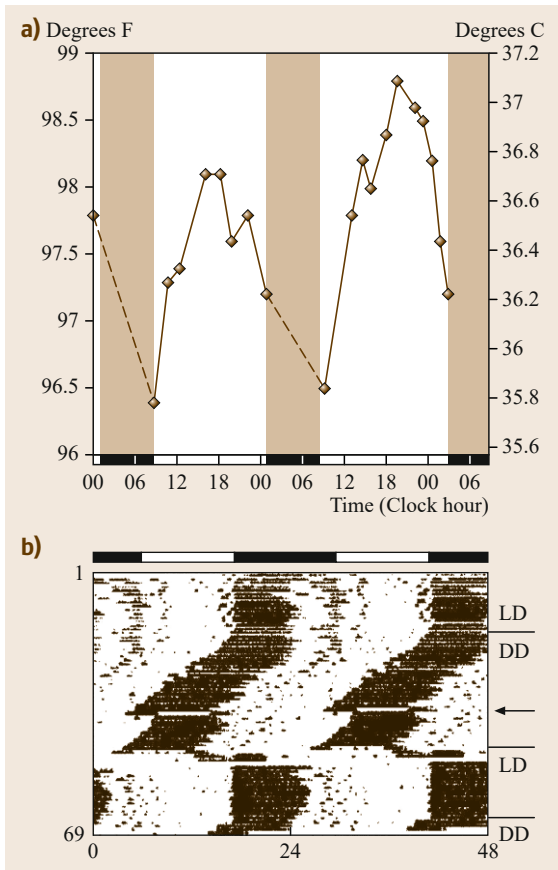
### 27.2.1 Data Graphs in Circadian Biology

By far the majority of the diagrams that figure in scientific papers are devoted to graphing data. Surprisingly, given the recognition of the roles data play both in discovering possible explanations and in evaluating them, there is little discussion in philosophy of graphing practices and how they figure in discovery and justification. Rather, the focus has been on data claims that can be represented sentimentally. Although there are common graphic formats that are highly familiar – for example, line graphs and bar graphs – in fact a wide variety of graphic formats are frequently used in science. In particular fields scientists have created their own formats, but these formats often migrate between fields. Each format elicits specific visual processing operations to identify informative patterns. In addition to different graphic formats, there are different tasks in which scientists present data. I focus on two tasks – delineating

phenomena and presenting relations between variables that are taken to be explanatory of the phenomenon.

In presenting phenomena as the target of scientific explanations, *Bogen and Woodward* [27.20] distinguish phenomena from data. They argue that phenomena, unlike data, are repeatable regularities in the world. Data provide evidence for the occurrence of phenomena. In many cases, researchers delineate phenomena by identifying patterns in data they collect. In the case of circadian rhythms, these are patterns of activity that repeat every 24 h and are often detectable by visual inspection of diagrams.

One of the most basic diagramming techniques employs a Cartesian coordinate system on which one plots values of relevant variables on the two axes. Using the abscissa to represent time and the ordinate for the value of a variable such as temperature, circadian researchers can plot each data point and then connect them by lines or a smoothed curve (Fig. 27.4a). Our visual sys-



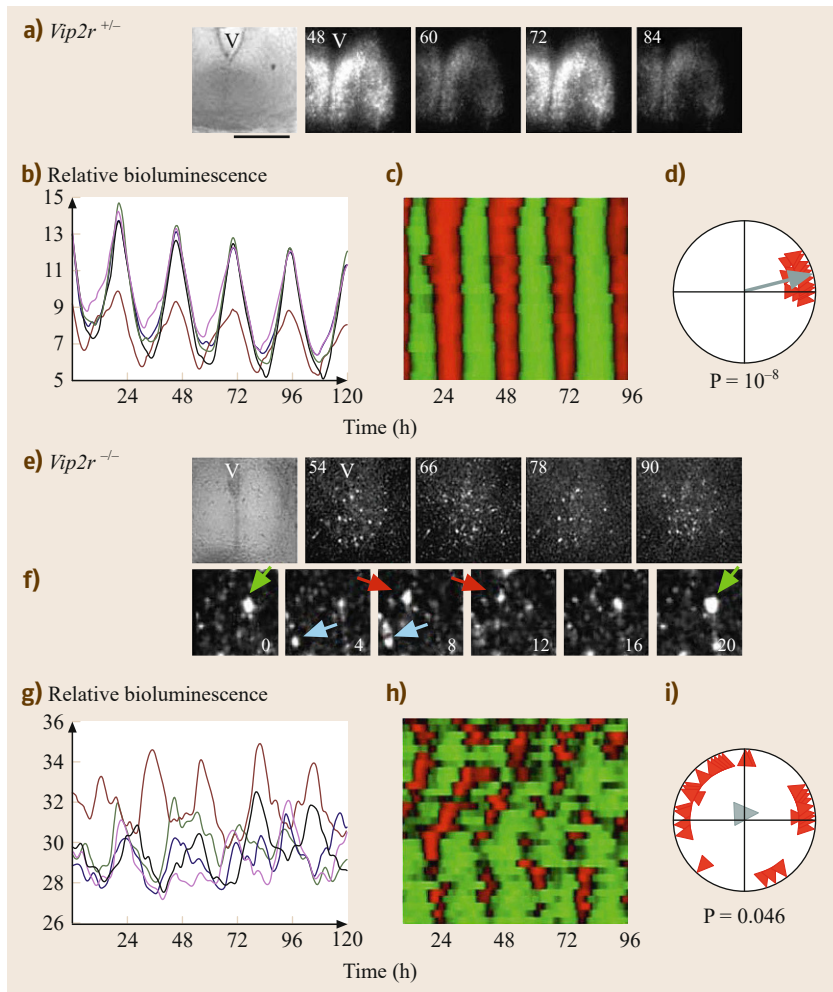
**Fig. 27.4** (a) Line graph from [27.21] showing the circadian oscillation in body temperature for one person across 48 h. (b) Example actogram showing times of running wheel activity of a wild-type mouse, reprinted with permission from Elsevier (after [27.22])

tem readily identifies the oscillatory pattern, which we can then coordinate with the bar at the bottom indicating periods of light and dark and the gray regions that redundantly indicate periods of darkness. By visually investigating the graph, one can detect that body temperature rises during the day and drops during the night, varying by about  $2^{\circ}\text{F}$  over the course of a day.

A line graph makes clear that the value of a variable is oscillating and with what amplitude, but it does not make obvious small changes in the period of activity. For this reason, circadian researchers developed actograms – a version of a raster plot on which time of each day is represented along a horizontal line and each occurrence of an activity (rotation of a running wheel by a mouse) is registered as a hash mark. Subsequent days are shown on successive lines placed below the previous one. Some actograms, such as the one shown in Fig. 27.4b, double plot the data so that each successive

24 h period is both plotted to the right of the previous 24-hour period and then again on the left on the next line. Placing adjacent times next to each other even when they wrap around a day break makes it easier to track continuous activity patterns. An actogram renders visually apparent how the phase of activity changes under different conditions such as exposure to light. In this actogram, the mouse was first exposed to a 12 : 12 light-dark cycle, as indicated by the letters LD on the right side, with the periods of light and dark indicated by the light-dark bar at the top. From day 15 to day 47, as indicated by the letters DD on the right side, the mouse was subjected to continuous darkness. On day 37, the row indicated by the arrow, the animal received a 6 h pulse of light at hour 16. It was returned to LD conditions on day 48, but returned to DD on day 67. The activity records shown on the actogram exhibit a clear pattern. During both LD periods the activity of the mouse was entrained to the pattern of light and dark so that the mouse was primarily active during the early night, with a late bout of activity late in the night (mice are nocturnal animals). On the other hand, during the DD periods the mouse began its activity somewhat earlier each day, a phenomenon known as *free running*. The light pulse reset the onset time for activity on the following day, after which the mouse continued to free run but from this new starting point. When switched back to LD the mouse exhibited a major alternation in activity the next day, but it took a couple more days to fully re-entrain to the LD pattern.

Data graphs are used not just to characterize phenomena but also to identify factors that may play a role in explaining phenomena. Figures in biological papers often contain many panels, invoking different representational formats, as part of the attempt to make visible relationship between variables that are taken to be potentially explanatory. For example, Fig. 27.5, from [27.23], employs photographs, line graphs, heat maps, and radial (Rayleigh) plots. To situate their research, in the 1970s the suprachiasmatic nucleus (SCN), a small structure in the hypothalamus, was implicated through a variety of techniques as the locus of circadian rhythms in mammals. Welsh et al. [27.24] had demonstrated that while individual SCN neurons maintain rhythmicity when dispersed in culture, they oscillate with varying periods and quickly become desynchronized. Maywood et al.'s research targeted vasoactive intestinal polypeptide (VIP), which is released by some SCN neurons, as the agent that maintains synchrony in the whole SCN or in slices from the SCN. Accordingly, they compared SCN slices from mice in which one (identified as  $\text{VIP2r}^{+/-}$ ) or both copies ( $\text{VIP2r}^{-/-}$ ) of the gene that codes for the VIP receptor are deleted. To render the rhythmicity of individual



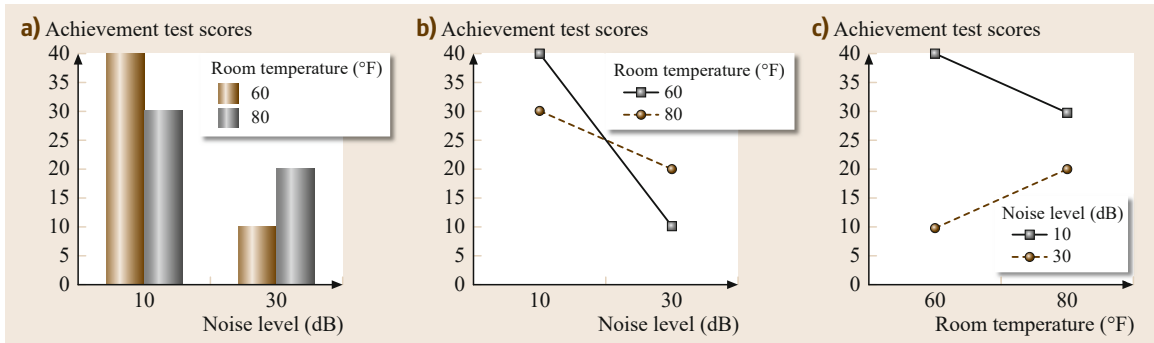
**Fig. 27.5** By using multiple graphical formats ((a,e,f) photographs of slices from the SCN, (b,g) line graphs, (c,h) raster plots (heat maps) and (d,i) radial (Rayleigh) plots) Maywood et al. make apparent in panels (a–d) that, when a receptor for VIP is present, oscillations of individual neurons are synchronized but that this is lost without VIP, panels (e–i) (after [27.23]), with permission from Elsevier

cells visible, the researchers inserted a gene coding for luciferase under control of the promoter for a known clock gene, *Per1*, so as to produce luminescence whenever PER is synthesized. The photographs in panels A are selections among the raw data. They make clear that VIP luminescence in the slice is synchronized, occurring at hour 48 and 72. Panel E reveals the lack of synchrony without the VPN receptor and panel F demonstrates that individual neurons are still oscillating without VIP but that the three neurons indicated by green, blue, and red arrows exhibit luminescence at different phases.

Although the photographs are sufficient to show that VIP is potentially explanatory of synchronous activity in the SCN, the researchers desired to characterize the relationship in more detail. They began by quantifying the bioluminescence recorded at the locus of the cell in photographs at different times. In panels B and G they displayed the results for five individual cells in each of

each type in line graphs. This makes it clear that while there is variation in amplitude, with VIP the five cells are in phase with each other while without VIP they are not. Even with five cells, though, it becomes difficult to decipher the pattern in a line graph. The raster plots in panels C and H enable comparison of 25 cells, one on each line, with red indicating periods when bioluminescence exceeds a threshold and green periods when it is below the threshold (such displays using hot and cold colors are often called *heat maps*). The raster plot enables one to compare the periodicity of individual cells more clearly, but with a loss of information about the amplitude of the oscillation at different times. The Rayleigh plots shown in panels D and I sacrifice even more information, focusing only on peak activity, but show that the peak phases are highly clustered with VIP and widely distributed without. The blue arrow shows the aggregate phase vector and indicates not only that it is oriented differently without VIP but also is





**Fig. 27.6a–c** A bar graph (a) and two line graphs (b,c), each showing the same data, but which viewers typically interpret differently (after [27.29])

extremely short, indicative of little correlation between individual neurons.

### 27.2.2 Cognitive Science Research Relevant to Reasoning with Graphs

Having introduced examples of graphs used in one field of biology, I turn now to cognitive science research that has attempted to identify aspects of the cognitive operations that figure in reasoning with graphs. *Pinker* [27.25] provided the foundation for much subsequent research on how people comprehend graphs. He differentiated the cognitive activities of creating a visual description of a graph and applying an appropriate graph schema to it. He treats the construction of a visual description as initially a bottom-up activity driven by the visual stimulus to which gestalt principles such as proximity and good continuation, among other procedures, are invoked. As explored by *Zacks* and *Tversky* [27.26], these principles differentially affect perception of bar graphs and line graphs [27.26, p. 1073]:

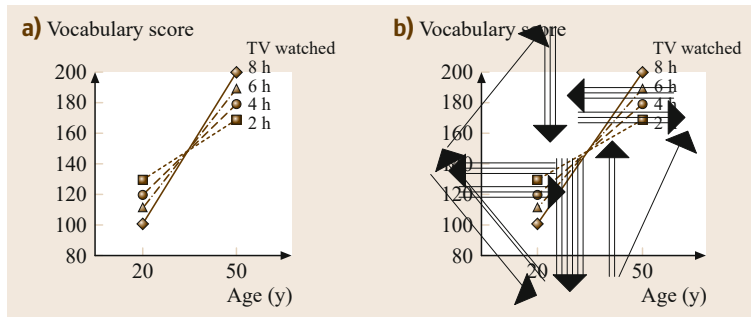
“Bars are like containers or fences, which enclose one set of entities and separate them from others. Lines are like paths or outstretched hands, which connect separate entities.”

The result, which has been documented in many studies, is that people are faster and more accurate at reading individual data points from bar graphs than line graphs but detect trends more easily in line graphs [27.27, 28]. For example, the bar graph in Fig. 27.6a makes it easy to read off test scores at different noise levels and room temperatures, and to compare test scores at the two temperatures. The line graph in Fig. 27.6b encodes the same data but the lines connecting the values at the two noise levels make that comparison more apparent. Moreover, the line graph

suggests that there are intermediate values between the two explicitly plotted. The effect is sufficiently strong that *Zacks* and *Tversky* found that when line graphs are used with categorical variables, viewers often treat them as interval variables and make assertions such as “The more male a person is, the taller he/she is” [27.26, p. 1076].

The choice of what to present on the axes also affects the information people extract. *Shah* and *Carpenter* [27.30] found that participants produce very different interpretations of the two graphs on the right of Fig. 27.6, one representing noise and the other room temperature on the abscissa. Thus, viewers of the graph in the center are more likely to notice the trend with increasing noise levels whereas those viewing the graph on the right notice the trend with increasing temperature. Further, when lines in graphs have reverse slopes, as in the rightmost graph, participants take longer to process the graph. Moreover, this difference makes the third variable, noise level, more salient since it identifies the difference responsible for the contrasting slopes.

The research reported so far focused on visual features of graphs, but one of the seminal findings about the organization of the mammalian visual processing system is that it is differentiated into two processing streams, one extracting information about the shape and identity of objects and one extracting information about location and potential for action [27.31, 32]. *Hegarty* and *Kozhevnikov* [27.33] proposed that the distinction between different processing pathways could help explain apparently contradictory results other researchers had reached about whether skill in visual imagery facilitates solving mathematics problems. They separately evaluated sixth-grade boys in Dublin, Ireland, in terms of pictorial imagery (“constructing vivid and detailed images”) and schematic imagery (“representing the spatial relationships between objects and imagining spatial transformations”). They found that good pictorial imagery was actually associated with poorer performance



**Fig. 27.7a,b** Graph (a) and superimposed eye-tracking results (b) (after [27.36]), with permission from the American Psychological Association

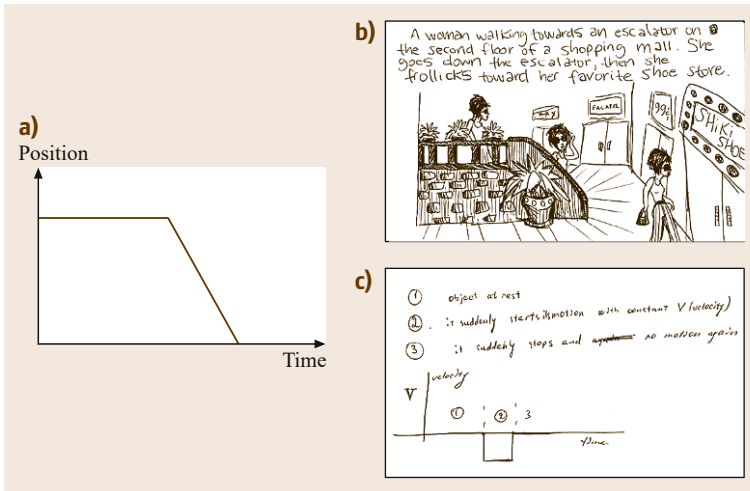
in solving mathematical problems. The following was a typical problem: *At each of the two ends of a straight path, a man planted a tree and then every 5 meters along the path he planted another tree. The length of the path is 15 meters. How many trees were planted?*

In contrast, good spatial imagery was associated with better performance. In subsequent work, Hegarty and her collaborators focused on kinetic problems involving graphs of motion and demonstrated a similar effect of pictorial versus spatial visualization. *Kozhevnikov et al.* [27.34] presented graphs such as that on the left in Fig. 27.7 to participants who, on a variety of psychometric tests, scored high or low on spatial ability. Those who scored low interpreted this graph pictorially as, for example, a car moving on a level surface, then going down a hill, and then moving again along a level surface. None of these participants could provide the correct interpretation of the graph as showing an object initially at rest, then moving at a constant velocity, and finally again at rest. On the other hand, all participants who scored high on spatial ability provided the correct interpretation. Subsequently, *Kozhevnikov et al.* [27.35] examined the differences between professionals in the arts and the sciences with respect to these graphs. They found that, except for participants who provided an irrelevant interpretation by focusing on nonpictorial features of the graph, artists tended to provide a literal pictorial interpretation of the path of movement whereas all scientists offered a correct schematic interpretation (example responses are shown on the right of Fig. 27.8).

So far I have focused on viewing a graph and extracting information from it. But an important feature of graphs in science such as those I presented in the earlier section is that they afford multiple engagements in which a user visually scans different parts of the graph seeking answers to different questions, some posed by information just encountered. *Carpenter and Shah* [27.36] drew attention to this by observing that graph comprehension is an extended activity often requiring half a minute, two orders of magnitude longer

than the time required to recognize simple patterns, including words and objects. In addition to detecting a pattern of data points along, for example, a positively sloping line, the graph interpreter must relate these points to the labels on the axes and what these represent and this is what requires processing time. Using eye tracking which participants study graphs, *Carpenter and Shah* revealed that viewers initially carve the graph into visual chunks and then cycle through focusing on different components – the pattern of the lines, the labels on the axes, the legend, and the title of the graph (Fig. 27.7). Similarly drawing attention to the prolonged engagement individuals often have with graphs, *Trickett and Trafton* [27.37] employed verbal protocols as well as eye tracking to study what people do when making inferences that go beyond what is explicitly represented in a given graph. They found that participants often employ spatial manipulations such as mentally transforming an object or extending it; they are not just passively viewing it.

Cognitive scientists have limited their focus to relatively simple graph forms such as line graphs and have not investigated the larger range of format we saw deployed in circadian research. Many of the results, however, are applicable to these other graph formats. Gestalt principles such as good continuation affect the patterns people see in actograms and raster plots (heat maps). In the actogram in Fig. 27.4, one recognizes the phase locking of activity to the light-dark cycle and daily phase advance when light cues are removed by implicitly (and sometimes explicitly) drawing a line through the starting point for each day's activity. Spatial processing is clearly important not only with the photographs in Fig. 27.5 but also with the heat map and Rayleigh plot. A skilled user of these graphs must recognize that space in the photographs corresponds to space on the slice from the brain but that space in the heat map corresponds not to physical space but an abstract space in which different cells are aligned. Finally, these diagrams are not designed to convey information in one look but rather are objects that afford shifting one's attention many times to focus on different in-



**Fig. 27.8** (a) A line graph showing an object initially at rest, then moving for a period at a constant velocity, then returning to rest that Kozhevnikov et al. [27.35] used to compare interpretations by artists and scientists. (b) A typical response from an artist, (c) response from a scientist

formation. With the Rayleigh plot, for example, one typically attends separately to the dispersal of blue arrowheads reflecting peaks of individual cells and to the vector indicating the population average. If eye tracking were performed, the pattern would likely resemble that displayed in Fig. 27.7. With panels showing the same information in multiple formats, as in Fig. 27.5, viewers are also likely to shift their focus between panels to see, for example, how the times in the line graph correspond to those in the photograph or those in the heat map. One limitation of the cognitive science studies is that the tasks participants were asked to perform were usually quite limited (e.g., interpret the graph)

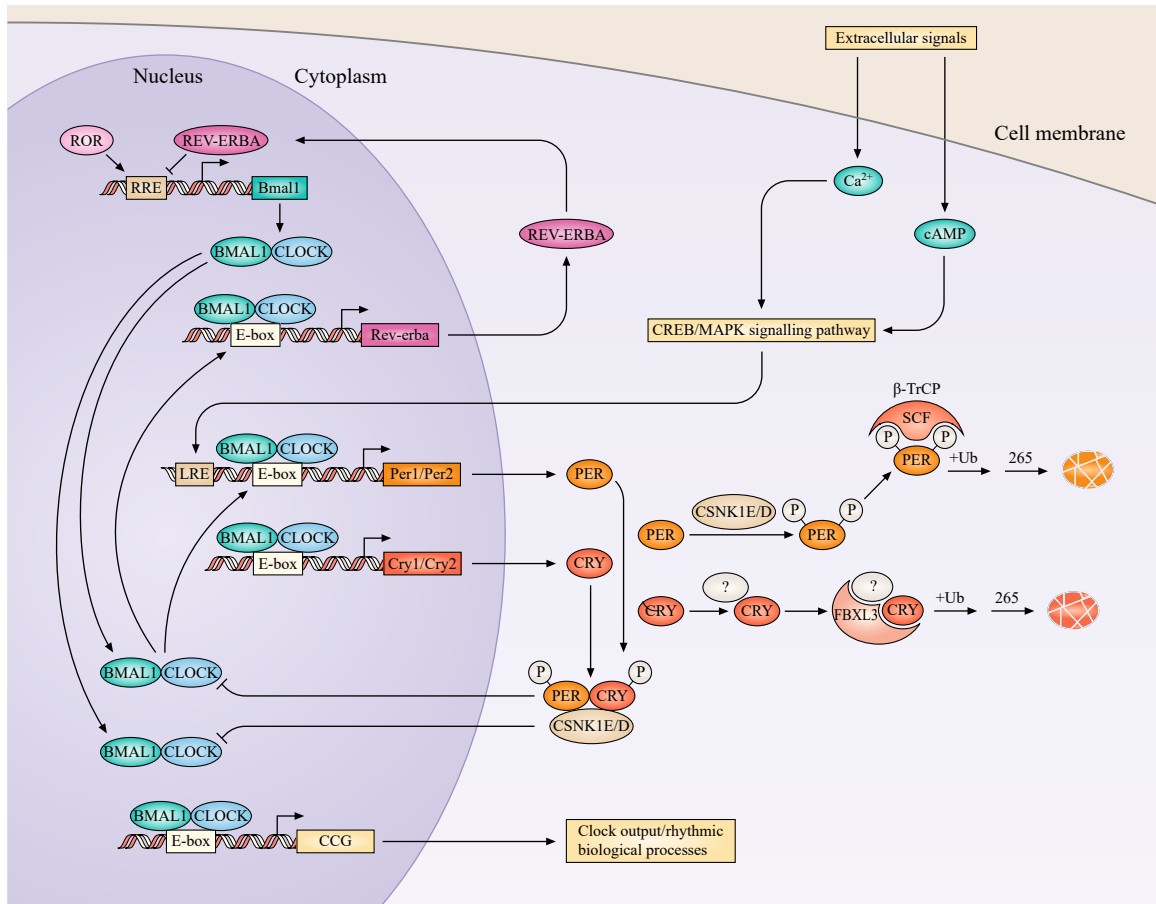
whereas scientists often use interact with graphs over multiple engagements, constructing new queries on the basis of previous ones (e.g., probing an actogram to see if the behavior really does look rhythmic or not or exploring the variability between cells revealed in a heat map). This is particularly evident when a researcher pours over a graph after producing it to determine what it means or when, in a journal club discussion, other researchers raise questions about specific features of a graph. Ultimately we need to better understand how scientists pose and address such queries over time if we are to understand the different roles graphs play in scientific reasoning.

## 27.3 Reasoning with Mechanism Diagrams

### 27.3.1 Mechanism Diagrams in Circadian Biology

Recognizing that individual activities, even if they do play a causal role in generating a phenomenon, typically do not work in isolation but only in the context of a mechanism in which they interact with other components, biologists often set as their goal to characterize the mechanism. (For discussion of the appeals to mechanism to explain biological phenomena, see [27.39–41].) The researchers' conception of the mechanism is sometimes presented in a final figure in a journal article but mechanism diagrams are even more common in review papers. Figure 27.9 is a representative sample of a mechanism diagram for the intracellular oscillator in mammalian SCN neurons. The diagram uses glyphs—"simple figures like points, lines, blobs, and arrows, which derive their meaning from geometric or

gestalt properties in context" [27.42] to represent the parts and operations of the mechanism. Tversky emphasizes the abstractness of glyphs over more iconic representations, arguing that the abstractness promotes generalization. One can abstract even more by allowing only one type of glyph (e.g., a circle) for an entity and one for an operation (an arrow), generating the sort of representations found in graph theory and used to capture general consequences of the organization of mechanisms. See Bechtel [27.43]. The parts shown in Fig. 27.9 include DNA strands, indicated by two wavy lines, on which promoter regions are indicated by lightly shaded rectangles, genes by darkly colored rectangles, and protein products, by colored ovals. Lines with arrow heads represent operations such as expression of a gene or transport of proteins to locations where they figure in other reactions, including activating gene transcription. Lines with squared



**Fig. 27.9** Takahashi et al. mechanism diagram of the mammalian circadian clock involves genes and proteins within individual cells. Reprinted with permission from Macmillan Publishers Ltd (after [27.38])

ends indicate inhibitory activity. When phosphates attach to molecules (as preparation for nuclear transport or degradation), they are shown as white circles containing a P.

The diagram is clearly laid out spatially, but only some features of the diagram convey information about spatial structures in the cell. The differentiation of the nucleus and cytoplasm is intended to correspond to these regions in the cell and lines crossing the boundary between the two parts of the cell. Beyond that, however, the distribution of shapes and arrows conveys no spatial information but only functional differentiation. The most important operations shown in this diagram are the synthesis of REV-ERBA and its subsequent transport into the nucleus to inhibit transcription of BMAL1 (shown as a loop out from and back into the nucleus in the upper left) and the synthesis of PER and CRY, the formation of a dimer, and the transport of the dimer into the nucleus to inhibit the ability of BMAL1 and

CLOCK to activate transcription of BMAL1, PER, and CRY (shown as a loop out from and back into the nucleus in the center-left of the figure). (The other operations shown are those involved in signaling from outside the cell that regulates the overall process, in the degradation of PER and CRY, and in the expression of clock-controlled genes (CCGs) that constitute the output of the clock.

For someone acquainted with the types of parts shown and the operations in which they engage, a diagram such as this provides a means of showing schematically how the various parts perform operations that affect other parts. One is not intended to take in the whole diagram at once, but to follow the operations from one part to another. To understand how the mechanism gives rise to oscillatory activity, one can mentally simulate the operations of the mechanism by starting in the middle with the *Per* and *Cry* genes. As they are expressed, more PER and CRY proteins are generated. After the proteins form a dimer and are

transported into the nucleus, they inhibit the activity of the BMAL1:CLOCK dimer and thereby stop their own expression. This reduction in expression results in reduction in their concentration and reduced inhibitory activity, which allow expression to resume. This capacity for mental animation is, however, limited, and to determine what the activity will be, especially when the other components are included, researchers often turn to computational models, generating what *Abrahamsen and I* [27.44] refer to as *dynamic mechanistic explanations*. Even here, though, diagrams provide a reference point in the construction of equations describing operation of the various parts [27.45].

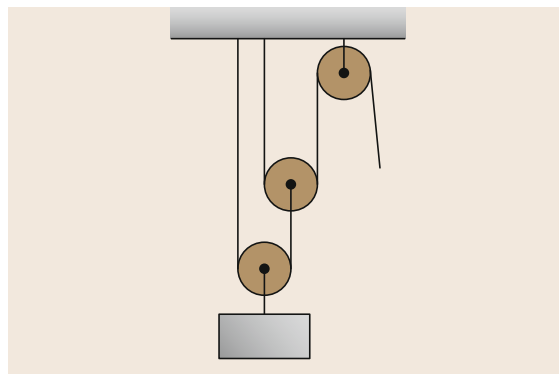
Looking carefully at the lower right side of the figure, one will see two ovals with question marks in them. This indicates that the researchers suspected that something unknown binds with CRY before and potentially mediates its binding with FBXL3, which then results in its degradation. Here it is the identity of an entity that is in doubt, but sometimes question marks are employed to indicate uncertainty about the identity of an operation. In this case the diagram is from a review paper and the question mark reflects uncertainty in the discipline. On occasions when question marks appear in mechanism diagrams presented at the beginning of a researcher paper they signal that the goal of the paper is to answer a question regarding the identity of a component or its operation.

### 27.3.2 Cognitive Science Research Relevant to Reasoning with Mechanism Diagrams

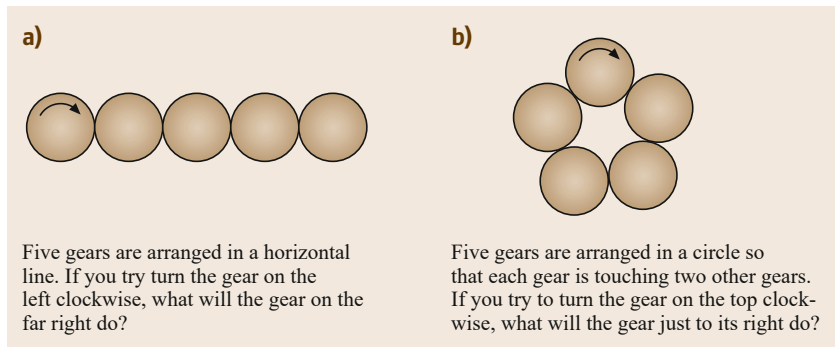
Although cognitive scientists have not explicitly focused on mechanism diagrams that figure in biology (but see *Stieff et al.* [27.46], which explores strategies used to transform diagrams of molecular structure used in organic chemistry), research on simple mechanical systems such as pulley systems (already the focus of *Larkin and Simon's* research discussed above) has highlighted one of the important cognitive activities people use with mechanism diagrams – mentally animating the operation of a mechanism when trying to figure out how it will behave. Drawing on theorists in the mental models tradition (see papers in [27.47] that explore how people answer problems by constructing and running a mental model), *Hegarty* [27.48] investigated experimentally “to what extent the mental processes involved in reasoning about a mechanical system are isomorphic to the physical processes in the operation of the system.” She measured reaction times and eye movements as participants answered questions about how various parts of a pulley system such as shown in Fig. 27.10 would behave if the rope is pulled. From the fact that

both error rates and reaction times increased with the number of operations within the mechanism the participant had to animate in order to answer the question, she inferred that people don't simulate the whole machine operating at once but rather animate individual parts in sequence. She provided further evidence for this claim by tracking the movements of the participants' eyes as they solved problems. In a follow-up experiment, *Hegarty* compared performance when participants were asked to infer the motion of a component from that of another component earlier in the causal chain or from that of a component later in the chain. Participants made more errors and required more time when they had to reason backward from events later in the chain, and still showed a preference to move their eyes forward along the causal chain.

*Schwartz and Black* [27.49] provided further insights into how people simulate mechanisms by attending to the gestures people make. In one task, shown in Fig. 27.11a, participants were asked to determine in which direction the rightmost gear would turn given the clockwise turn of the leftmost gear. They found that their participants would use their hands to indicate the direction of movement of each successive gear. (In these studies the participants never saw the diagrams but were provided with verbal descriptions of the configuration.) In this case, an alternative strategy is available: apply a simple global rule such as the parity rule: *if there are odd number of gears, the first and last will turn in the same direction* or the more local rule *if two gears are touching, they will turn in opposite directions*. *Schwartz and Black* found that as people acquired the rule, their gestures declined. But when people lack such rules or find their application uncertain, as in the gear problem in Fig. 27.11b, they again gesture. This use of gesture indicates that whatever imagery people



**Fig. 27.10** Pulley problem (after [27.48]) used to study how people employ mental animation in problem solving. With permission from the American Psychological Association



**Fig. 27.11a,b** Gear problem used by Schwartz and Black [27.49] to study when participants gesture while solving problems

employ to solve the task, it is coordinated with action. Accordingly, the researchers propose a theory of simulated doing in which [27.50]:

“the representation of physical causality is fundamental. This is because *doing* requires taking advantage of causal forces and constraints to manipulate the world. Our assumption is that people need to have representations of how their embodied ideas will cause physical changes if they are to achieve their goals.”

Animating a diagram, either mentally or with gesture, plays a crucial role in the cognitive activity of understanding how a proposed mechanism could produce the phenomenon one is trying to explain. But diagrams present not only a finished explanation of the phenomenon, they often figure in the process of discovering mechanisms. Here what matters is the ability to create and alter the glyphs and their arrangement. *Tversky* [27.42] suggests a helpful way to understand this activity – view diagrams as the “permanent traces of gestures” in which “fleeting positions become places and fleeting actions become marks and forms” [27.42, p. 500]. There is a rich literature showing how ges-

ture figures not only in communication but also in the development of one’s own understanding [27.51]. *Tversky* focuses on the activity of drawing maps, highlighting such features of the activity as selecting what features to include and idealizing angles to right angles. These findings can be extended to mechanism diagrams, which constitute a map of the functional space of the mechanism, situating its parts and operations. While *Tversky* speaks of diagrams as permanent traces and there is a kind of permanence (or at least endurance) to diagrams produced on paper or in computer files, they are also subject to revision – one can add glyphs for additional parts or alter arrows to represent different ideas of how the operations of one part affect others. In the design literature this is often referred to as *sketching*. Sketching mechanism diagrams can be motivated by evidence, but they can also be pursued in a purely exploratory manner, enabling reasoning about what would happen if a new connection were made or an existing one redirected. Sketching possible mechanisms is a common activity of scientists, and by further investigating the cognitive activities involved in this activity one can develop richer analyses of this important type of scientific reasoning.

## 27.4 Conclusions and Future Tasks

This chapter has addressed the use of diagrams by scientists in characterizing phenomena to be explained, identifying variables that figure in explaining those phenomena, and advancing proposals for mechanisms, drawing examples from circadian rhythm research. Over the last 30 years cognitive scientists have attempted to characterize cognitive activities people employ when perceiving and using diagrams in problem-solving tasks, such as making multiple scans of graphs and animating mechanical diagrams. For the most part, cognitive scientists have employed diagrams and tasks

in their studies that are simpler than those that figure in actual scientific research. But these cognitive science studies nonetheless provide insights into the cognitive processes that figure in scientists’ use of diagrams. To date the roles diagrams play in science have not figured in a major way in philosophical accounts of scientific reasoning but given the important roles diagrams play in science, there is great potential to advance our understanding of scientific reasoning by investigating further the cognitive processes involved as scientists create and use diagrams in the course of their research.

**Acknowledgments.** Research for this chapter was supported by Grant 1127640 from the US National Science Foundation and a Senior Visiting Fellowship in the Center for Philosophy of Science at the University of Pittsburgh, which are gratefully acknowledged. I also thank fellow members of the Working Group on Diagrams in Science at UCSD (Adele Abrahamsen, Daniel

C. Burnston, and Benjamin Sheredos) and John Norton and the Fellows at the Center for Philosophy of Science at the University of Pittsburgh in Fall 2014 (Joshua Alexander, Karim Bschrir, Ingo Brigandt, Sara Green, Nicholaos Jones, Raphael Scholl, and Maria Serban) for insights and suggestions.

## References

- 27.1 J.M. Zacks, E. Levy, B. Tversky, D.J. Schiano: Graphs in print. In: *Diagrammatic Representation and Reasoning*, ed. by M. Anderson, B. Meyer, P. Olivier (Springer, London 2002) pp. 187–206
- 27.2 D.C. Marr: *Vision: A Computation Investigation Into the Human Representational System and Processing of Visual Information* (Freeman, San Francisco 1982)
- 27.3 J.J. Gibson: *The Ecological Approach to Visual Perception* (Houghton Mifflin, Boston 1979)
- 27.4 J. Zhang: The nature of external representations in problem solving, *Cogn. Sci.* **21**, 179–217 (1997)
- 27.5 J. Zhang, D.A. Norman: Representations in distributed cognitive tasks, *Cogn. Sci.* **18**, 87–122 (1994)
- 27.6 J.H. Larkin, H.A. Simon: Why a diagram is (sometimes) worth ten thousand words, *Cogn. Sci.* **11**, 65–99 (1987)
- 27.7 J.V. Kulvicki: *Images*, 1st edn. (Routledge, New York 2013)
- 27.8 L.A. Cooper, R.N. Shepard: Chronometric studies of the rotation of mental images. In: *Visual Information Processing Symposium on Cognition, 8th Carnegie Mellon University*, ed. by W.G. Chase (Academic Press, New York 1973) pp. 75–175
- 27.9 R.N. Shepard, J. Metzler: Mental rotation of three-dimensional objects, *Science* **171**, 701–703 (1971)
- 27.10 S.M. Kosslyn, T.M. Ball, B.J. Reiser: Visual images preserve metric spatial information – Evidence from studies of image scanning, *J. Exp. Psychol. Human Percept. Perform.* **4**, 47–60 (1978)
- 27.11 D. Chambers, D. Reisberg: Can mental images be ambiguous?, *J. Exp. Psychol. Human Percept. Perform.* **11**, 317–328 (1985)
- 27.12 S. Reed, J. Johnsen: Detection of parts in patterns and images, *Memory Cogn.* **3**, 569–575 (1975)
- 27.13 R.A. Finke, S. Pinker, M.J. Farah: Reinterpreting visual patterns in mental imagery, *Cogn. Sci.* **13**, 51–78 (1989)
- 27.14 R.A. Finke, K. Slayton: Explorations of creative visual synthesis in mental imagery, *Memory Cogn.* **16**, 252–257 (1988)
- 27.15 R.E. Anderson, T. Helstrup: Multiple perspectives on discovery and creativity in mind and on paper. In: *Imagery, Creativity, and Discovery: A Cognitive Perspective*, Vol. 98, ed. by B. Roskos-Ewoldsen, M.J. Intons-Peterson, E.A. Rita (Elsevier, Amsterdam 1993) pp. 223–253
- 27.16 R.E. Anderson, T. Helstrup: Visual discovery in mind and on paper, *Memory Cogn.* **21**, 283–293 (1993)
- 27.17 I.M. Verstijnen, C. van Leeuwen, R. Hamel, J.M. Hennessey: What imagery can't do and why sketching might help, *Empir. Stud. Arts* **18**, 167–182 (2000)
- 27.18 I.M. Verstijnen, C. van Leeuwen, G. Goldschmidt, R. Hamel, J.M. Hennessey: Creative discovery in imagery and perception: Combining is relatively easy, restructuring takes a sketch, *Acta Psychol.* **99**, 177–200 (1998)
- 27.19 M.P. Cook: Students' comprehension of science concepts depicted in textbook illustrations, *Electron. J. Sci. Educ.* **12**(1), 1–14 (2008)
- 27.20 J. Bogen, J. Woodward: Saving the phenomena, *Philos. Rev.* **97**, 303–352 (1988)
- 27.21 W.L. Koukkari, R.B. Southern: *Introducing Biological Rhythms* (Springer, New York 2006)
- 27.22 M.K. Bunger, L.D. Wilsbacher, S.M. Moran, C. Clendenin, L.A. Radcliffe, J.B. Hogenesch, M.C. Simon, J.S. Takahashi, C.A. Bradfield: Mop3 is an essential component of the master circadian pacemaker in mammals, *Cell* **103**, 1009–1017 (2000)
- 27.23 E.S. Maywood, A.B. Reddy, G.K.Y. Wong, J.S. O'Neill, J.A. O'Brien, D.G. McMahon, A.J. Harmar, H. Okamura, M.H. Hastings: Synchronization and maintenance of timekeeping in suprachiasmatic circadian clock cells by neuropeptidergic signaling, *Curr. Biol.* **16**, 599–605 (2006)
- 27.24 D.K. Welsh, D.E. Logothetis, M. Meister, S.M. Reppert: Individual neurons dissociated from rat suprachiasmatic nucleus express independently phased circadian firing rhythms, *Neuron* **14**, 697–706 (1995)
- 27.25 S. Pinker: A theory of graph comprehension. In: *Artificial Intelligence and the Future of Testing*, ed. by R. Feedle (Psychology Press, London 1990) pp. 73–126
- 27.26 J.M. Zacks, B. Tversky: Bars and lines: A study of graphic communication, *Memory Cogn.* **27**, 1073–1079 (1999)
- 27.27 W.A. Simcox: A method for pragmatic communication in graphic displays, *Human Factors* **26**, 483–487 (1984)
- 27.28 C.M. Carswell, C.D. Wickens: Information integration and the object display: An interaction of task demands and display superiority, *Ergonomics* **30**, 511–527 (1987)
- 27.29 P. Shah, J. Hoeffner: Review of graph comprehension research: Implications for instruction, *Educ. Psychol. Rev.* **14**, 47–69 (2002)

- 27.30 P. Shah, P.A. Carpenter: Conceptual limitations in comprehending line graphs, *J. Exp. Psychol. General* **124**, 43–61 (1995)
- 27.31 L.G. Ungerleider, M. Mishkin: Two cortical visual systems. In: *Analysis of Visual Behavior*, ed. by D.J. Ingle, M.A. Goodale, R.J.W. Mansfield (MIT Press, Cambridge 1982) pp. 549–586
- 27.32 D.C. van Essen, J.L. Gallant: Neural mechanisms of form and motion processing in the primate visual system, *Neuron* **13**, 1–10 (1994)
- 27.33 M. Hegarty, M. Kozhevnikov: Types of visual-spatial representations and mathematical problem solving, *J. Educ. Psychol.* **91**, 684–689 (1999)
- 27.34 M. Kozhevnikov, M. Hegarty, R.E. Mayer: Revising the visualizer-verbalizer dimension: Evidence for two types of visualizers, *Cogn. Instr.* **20**, 47–77 (2002)
- 27.35 M. Kozhevnikov, S.M. Kosslyn, J. Shephard: Spatial versus object visualizers: A new characterization of visual cognitive style, *Memory Cogn.* **33**, 710–726 (2005)
- 27.36 P.A. Carpenter, P. Shah: A model of the perceptual and conceptual processes in graph comprehension, *J. Exp. Psychol. Appl.* **4**, 75–100 (1998)
- 27.37 S. Trickett, J. Trafton: Toward a comprehensive model of graph comprehension: Making the case for spatial cognition. In: *Diagrammatic Representation and Inference*, Vol. 4045, ed. by D. Barker-Plummer, R. Cox, N. Swoboda (Springer, Berlin, Heidelberg 2006) pp. 286–300
- 27.38 J.S. Takahashi, H.-K. Hong, C.H. Ko, E.L. McDermott: The genetics of mammalian circadian order and disorder: Implications for physiology and disease, *Nat. Rev. Genet.* **9**, 764–775 (2008)
- 27.39 W. Bechtel, R.C. Richardson: *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research* (MIT Press, Cambridge 1993)
- 27.40 W. Bechtel, A. Abrahamsen: Explanation: A mechanist alternative, *Stud. Hist. Philos. Biol. Biomed. Sci.* **36**, 421–441 (2005)
- 27.41 P. Machamer, L. Darden, C.F. Craver: Thinking about mechanisms, *Philos. Sci.* **67**, 1–25 (2000)
- 27.42 B. Tversky: Visualizing thought, *Top. Cogn. Sci.* **3**, 499–535 (2011)
- 27.43 W. Bechtel: Generalizing mechanistic explanations through graph-theoretic perspectives. In: *Explanation in Biology. An Enquiry into the Diversity of Explanatory Patterns in the Life Sciences*, ed. by P.-A. Braillard, C. Malaterre (Springer, Dordrecht 2015) pp. 199–225
- 27.44 W. Bechtel, A. Abrahamsen: Dynamic mechanistic explanation: Computational modeling of circadian rhythms as an exemplar for cognitive science, *Stud. Hist. Philos. Sci. Part A* **41**, 321–333 (2010)
- 27.45 N. Jones, O. Wolkenhauer: Diagrams as locality aids for explanation and model construction in cell biology, *Biol. Philos.* **27**, 705–721 (2012)
- 27.46 M. Stieff, M. Hegarty, B. Dixon: Alternative strategies for spatial reasoning with diagrams. In: *Diagrammatic Representation and Inference*, Vol. 6170, ed. by A. Goel, M. Jamnik, N.H. Narayanan (Springer, Berlin, Heidelberg 2010) pp. 115–127
- 27.47 D. Gentner, A.L. Stevens: *Mental Models* (Erlbaum, Hillsdale 1983)
- 27.48 M. Hegarty: Mental animation: Inferring motion from static displays of mechanical systems, *J. Exp. Psychol. Learn. Memory Cogn.* **18**, 1084–1102 (1992)
- 27.49 D.L. Schwartz, J.B. Black: Shuttling between depictive models and abstract rules: Induction and fallback, *Cogn. Sci.* **20**, 457–497 (1996)
- 27.50 D.L. Schwartz, T. Black: Inferences through imagined actions: Knowing by simulated doing, *J. Exp. Psychol. Learn. Memory Cogn.* **25**, 116–136 (1999)
- 27.51 S. Goldin-Meadow, S.M. Wagner: How our hands help us learn, *Trends Cogn. Sci.* **9**, 234–241 (2005)



## 28. Embodied Mental Imagery in Cognitive Robots

Alessandro Di Nuovo, Davide Marocco, Santo Di Nuovo, Angelo Cangelosi

This chapter is focused on discussing the concept of *mental imagery* as a fundamental cognitive capability to enhance the performance of cognitive robots. Indeed, the emphasis will be on the embodied imagery mechanisms applied to build artificial cognitive models of *motor imagery* and *mental simulation* to control complex behaviors of humanoid platforms, which represent the artificial body.

With the aim of providing a panorama of the research activity on the topic, first we give an introduction on the neuroscientific and psychological background of mental imagery in order to help the reader to contextualize the multidisciplinary environment in which we operate. Then, we review the work done in the field of artificial cognitive systems and robotics to mimic the process behind the human ability of creating *mental images* of events and experiences, and to use this process as a cognitive mechanism to improve the behavior of complex robots. Finally, we report the detail of three recent empirical studies in which mental imagery approaches were modelled through artificial neural networks (ANNs) to enable a cognitive robot with some human-like capabilities.

The work presented in this chapter takes inspiration from the human capability to build representations of the physical world in its mind. In particular, we studied the *motor imagery*, which is considered a multimodal simulation that activates the same, or very similar, sensory and motor modalities that are activated when human beings interact with the environment in real time. This is strictly related to the embodied cognition hypothesis, which affirms that all aspects of human cognition are shaped by aspects of the body.

Similarly, when artificial intelligence was moving his first steps, *Alan Turing* argued that in order to think and speak a machine may need a human-like body and that the development of robot cognitive skills might be just as simple as teaching a child [28.1]. This is the mo-

28.1	<b>Mental Imagery Research Background</b>	620
28.2	<b>Models and Approaches Based on Mental Imagery in Cognitive Systems and Robotics</b> .....	622
28.3	<b>Experiments</b> .....	624
28.3.1	The Humanoid Robotic Platform: The iCub .....	624
28.3.2	First Experimental Study: Motor Imagery for Performance Improvement .....	624
28.3.3	Second Experimental Study: Mental Training Evoked by Language ...	628
28.3.4	Third Experimental Study: Spatial Imagery .....	630
28.4	<b>Conclusion</b> .....	635
	<b>References</b> .....	635

These empirical studies exemplify how the proprioceptive information can be used by *mental imagery* models to enhance the performance of the robot, giving evidence of the embodied cognition theories in the context of artificial cognitive systems.

tivation behind a strongly humanoid design of some of the recent and most advanced robotic platforms, for example, iCub [28.2], NAO [28.3], and Advanced Step in Innovative MObility (ASIMO) [28.4]. These platforms are equipped with sophisticated motors and sensors, which replicate animal or human sensorimotor input–output streams. The sensors and actuators arrangement determine a highly redundant morphological structure of humanoid robots, which are traditionally difficult to control and, thus, require complex models implementing more sophisticated and efficient mechanisms that resemble the human cognition [28.5].

In this multidisciplinary context, improving the skill of a robot in terms of motor control and navigation capabilities, especially in the case of a complex robot

with many degrees of freedom, is a timely and important issue in current robotics research. Among the many bio-inspired mechanisms and models already tested in the field of robot control and navigation, the use of mental imagery principles is of interest in modeling mental

imagery as a complex, goal directed and flexible motor planning strategy and to go further in the development of artificial cognitive systems capable to better interact with the environment and refine their cognitive motor skill in an open-ended process.

## 28.1 Mental Imagery Research Background

Mental imagery, as the process behind the human ability of creating mental images of events and experiences, has long been the subject of research and debate in philosophy, psychology, cognitive science, and more recently, neuroscience [28.6]. Some of the main effects of mental practice on physical performance have been well established in experiments with humans as in the fields of sports science, work psychology, and motor rehabilitation. Neuropsychological research has long highlighted the complexity of brain activation in the activity of imagination. Studies have demonstrated the localization partly similar and partly different between imagery, perception, and visual memory [28.7, 8], while, in [28.9], it was demonstrated by a clinical case as visuospatial perception and imagery can be functionally separated in activating brain. In [28.10], authors proposed a revision of the constructs relevant to cognitive styles, placing them into a complex framework of heuristics regarding multiple levels of information processing, from the attentional and perceptual to metacognitive ones. These heuristics are grouped according to the type of regulatory function that they take from the automatic coding of data to the conscious use of cognitive resources. In this view, also at the cerebral area activation level, the distinction between elaboration of object properties (like shape or color) and spatial relations is better representative of a different *style* in the use of mental images than the ancient dichotomy verbal-visual [28.11].

But only quite recently a growing amount of evidence from empirical studies begun to demonstrate the relationship between bodily experiences and mental processes that actively involve body representations. This is also due to the fact that in the past, philosophical and scientific investigations of the topic primarily focused upon visual mental imagery. Contemporary imagery research has now broadly extended its scope to include every experience that resembles the experience of perceiving from any sensorial modality. From this perspective, understanding the role of the body in cognitive processes is extremely important and psychological and neuroscience studies are extremely important in this regard. *Wilson* [28.12] identified six claims in the current view of embodied cognition:

1. Cognition is situated
2. Cognition is time-pressured
3. We off-load cognitive work onto the environment
4. The environment is part of the cognitive system
5. Cognition is for action
6. Offline cognition is bodily based.

Among those six claims, the last claim is particularly important. According to this claim, sensorimotor functions that evolved for action and perception are used during offline cognition that occurs when the perceiver represents social objects, situations, or events in times and places other than the ordinary ones. This principle is reinforced by the concept of embodied cognition [28.13, 14], which affirms that the nature of intelligence is largely determined by the form of the body. Indeed, the body and every physical experience made through the body, shape the form of intelligence that can be observed in any autonomous systems. This means that even if the mind does not directly interact with the environment, it is able to apply mechanisms of sensory processing and motor control by using some innate abilities such as memory (implicit, short, and long term), problem solving, and mental imagery. These capabilities have been well studied in psychology and neuroscience, but the debate is still open on the issue of mental imagery, where mental imagery is defined as a sensation activated without sensorial stimulation.

Many evidences from empirical sciences have demonstrated the relationship between bodily experiences and mental processes that involve body representation. Neuropsychological research has demonstrated that the same brain areas are activated during seeing or recalling by images [28.15] and that areas controlling perception are needed also for maintaining mental images active in working memory. Therefore, mental imagery may be considered as a kind of biological simulation. In [28.16], author observed that the primary motor cortex M1 is activated during the production of motor images as well as during the production of active movement. Similarly, experimental studies show that neural mechanisms underlying real-time visual perception and mental visualization are the same when a task is mentally recalled [28.17]. Nevertheless, the neural

mechanisms involved in the active elaboration of mental images might be different from those involved in passive elaborations [28.18]. These studies demonstrate the tight relationship between mental imagery and motor activities (i. e., how the image in mind can influence movements and motor skills).

Recent research, both in experimental as well as practical contexts, suggests that imagined and executed movement planning relies on internal models for action [28.19]. These representations are frequently associated with the notion of internal (forward) models and are hypothesized to be an integral part of action planning [28.20, 21]. Furthermore, in [28.22], authors suggest that motor imagery may be a necessary prerequisite for motor planning. In [28.23], *Jeannerod* studied the role of motor imagery in action planning and proposed the so-called equivalence hypothesis – suggesting that motor simulation and motor control processes are functionally equivalent [28.24, 25]. These studies, together with many others (e.g., [28.26]), demonstrate how the images that we have in mind might influence movement execution and the acquisition and refinement of motor skills. For this reason, understanding the relationship that exists between mental imagery and motor activities has become a relevant topic in domains in which improving motor skills is crucial for obtaining better performance, such as in sport and rehabilitation. Therefore, it is also possible to exploit mental imagery for improving a human’s motor performance and this is achieved thanks to special mental training techniques.

Mental training is widely used among professional athletes, and many researchers began to study its beneficial effects for the rehabilitation of patients, in particular after cerebral lesions. In [28.24], for example, the authors analyzed motor imagery during mental training procedures in patients and athletes. Their findings support the notion that mental training procedures can be applied as a therapeutic tool in rehabilitation and in applications for empowering standard training methodologies. Others studies have shown that motor skills can be improved through mental imagery techniques. *Jeannerod* and *Decety* [28.18] discuss how training based on mental simulation can influence motor performance in terms of muscular strength and reduction of variability. In [28.27], authors show that imaginary fifth finger abductions led to an increased level of muscular strength. The authors note that the observed increment in muscle strength is not due to a gain in muscle mass. Rather,

it is based on higher level changes in cortical maps, presumably resulting in a more efficient recruitment of motor units. These findings are in line with other studies, specifically focused on motor imagery, which shows the enhancement of mobility range [28.28] or increased accuracy [28.29] after mental training. Interestingly, it should be noted that such effects operate both ways: mental imagery can influence motor performance and the extent of physical practice can change the areas activated by mental imagery [28.30]. As a result of these studies, new opportunities for the use of mental training have opened up in collateral fields, such as medical and orthopaedic–traumatologic rehabilitation. For instance, mental practice has been used to rehabilitate motor deficits in a variety of neurological disorders [28.31]. Mental training can be successfully applied in helping a person to regain lost movement patterns after joint operations or joint replacements and in neurological rehabilitation. Mental practice has also been used in combination with actual practice to rehabilitate motor deficits in a patient with subacute stroke [28.32], and several studies have also shown improvement in strength, function, and use of both upper and lower extremities in chronic stroke patients [28.33, 34].

In sport, beneficial effects of mental training for the performance enhancement of athletes are well established and several works are focused on this topic with tests, analysis, and in new training principles [28.35–38]. In [28.39], for example, a cognitive-behavioral training program was implemented to improve the free-throw performance of college basketball players, finding improvements of over 50%. Furthermore, the trial in [28.40], where mental imagery is used to enhance the training phase of hockey athletes to score a goal, showed that imaginary practice allowed athletes to achieve better performance. Despite the fact that there is ample evidence that mental imagery, and in particular motor imagery, contributes to enhancing motor performance, the topic still attracts new research, such as [28.41] that investigated the effect of mental practice to improve game plans or strategies of play in open skills in a trial with 10 female pickers. Results of the trial support the assumption that motor imagery may lead to improved motor performance in open skills when compared to the no-practice condition. Another recent paper [28.42] demonstrated that sports experts showed more focused activation patterns in prefrontal areas while performing imagery tasks than novices.

## 28.2 Models and Approaches Based on Mental Imagery in Cognitive Systems and Robotics

The introduction of humanoid robots had a great impact on the fast growing field of cognitive robotics, which represents the intersection of artificial cognitive modeling and robotic engineering. Cognitive robotics aims to provide new understanding on how human beings develop their higher cognitive functions. Thanks to the many potential applications, researchers in cognitive robotics are still facing several challenges in developing complex behaviors [28.43].

As exemplified in the following literature review, a key role is played by mental imagery and its mechanisms in order to enhance motor control in autonomous robots, and to develop autonomous systems that are capable of exploiting the characteristics of mental imagery training to better interact with the environment and refine their motor skills in an open-ended process [28.44]. Indeed, among the many hypotheses and models already tested in the field of cognitive systems and robotics, the use of mental imagery as a cognitive tool capable of enhancing robot behaviors is both innovative and well-grounded in experimental data at different levels.

A model-based learning approach for mobile robot navigation was presented in [28.45], where it is discussed how a behavior-based robot can construct a *symbolic process* that accounts for its deliberative thinking processes using internal models of the environment. The approach is based on a forward modeling scheme using recurrent neural learning, and results show that the robot is capable of learning grammatical structure hidden in the geometry of the workspace from the local sensory inputs through its navigational experiences.

An example of the essential role mental imagery can play in human–robot interaction was recognized by Roy et al. [28.46]. She presented a robot, called Ripley, which is able to translate spoken language into actions for object manipulation guided by visual and haptic perception. The robot maintained a dynamic *mental model*, a three-dimensional model of its immediate physical environment that it used to mediate perception, manipulation planning, and language. The contents of the robot’s mental model could be updated based on linguistic, visual, or haptic input. The mental model endowed Ripley with object permanence, remembering the position of objects when they were out of its sensory field.

Experiments on internal simulation of perception using ANN robot controllers are presented by Ziemke et al. [28.47]. The paper focuses on a series of experiments in which feedforward neural networks (FFNNs)

were evolved to control collision-free corridor following behavior in a simulated Khepera robot and predict the sensory input of next time step as accurately as possible. The trained robot is actually able to move blindly in a simple environment for hundreds of time steps, successfully handling several multistep turns.

In [28.48], authors present a neurorobotics experiment in which developmental learning processes of the goal-directed actions of a robot were examined. The robot controller was implemented with a multiple timescales recurrent neural network (RNN) model, which is characterized by the coexistence of slow and fast dynamics in generating anticipatory behaviors. Through the iterative tutoring of the robot for multiple goal-directed actions, interesting developmental processes emerged. Behavior primitives in the earlier fast context network part were self-organizing, while they appeared to be sequenced in the later, slow context part. Also observed was that motor images were generated in the early stage of development.

The study presented in [28.49] show how simulated robots evolved for the ability to display a context-dependent periodic behavior can spontaneously develop an internal model and rely on it to fulfil their task when sensory stimulation is temporarily unavailable. Results suggest that internal models might have arisen for behavioral reasons and successively exapted for other cognitive functions. Moreover, the obtained results suggest that self-generated internal states need not match in detail the corresponding sensory states and might rather encode more abstract and motor-oriented information.

Fascinatingly, in [28.50], authors explore the idea of dreams as a form of mental imagery and the possible role they might play in mental simulations and in the emergence and refinement of the ability to generate predictions on the possible outcomes of actions. In brief, what the authors propose is that robots might first need to possess some of the characteristics related to the ability to dream (particularly those found in infants and children) before they can acquire a robust ability to use mental imagery. This ability to dream, according to them, would assist robots in the generation of predictions of future sensory states and of situations in the world.

Internal simulations can help artificial agents to solve the stereo-matching problem, operating on the sensorimotor domain, with retinal images that mimic the cone distribution on the human retina [28.51]. This is accomplished by applying internal sensorimotor simulation and (subconscious) mental imagery to the

process of stereo matching. Such predictive matching is competitive to classical approaches from computer vision, and it has moreover the considerable advantage that it is fully adaptive and can cope with highly distorted images.

A computational model of mental simulation that includes biological aspects of brain circuits that appear to be involved in goal-directed navigation processes is presented in [28.52]. The model supports the view of the brain as a powerful anticipatory system, capable of generating and exploiting mental simulation for predicting and assessing future sensory motor events. The authors show how mental simulations can be used to evaluate future events in a navigation context, in order to support mechanisms of decision-making. The proposed mechanism is based on the assumption that choices about actions are made by simulating movements and their sensory effects using the same brain areas that are active during overt actions execution.

An interpretation of mental imagery based on the context of homeostatic adaptation is presented in [28.53], where the internal dynamics of a highly complex self-organized system is loosely coupled with a sensory-motor dynamic guided by the environment. This original view is supported by the analysis of a neural network model that controls a simulated agent facing sensor shifts. The agent is able to perceive a light in the environment through some light sensors placed around its body and its task is that of approaching the light. When the sensors are swapped, the agent perceives the light in the opposite direction of its real position and the control systems has to autonomously *detect* the shifting sensor and act accordingly. The authors speculate that mental imagery could be a viable way for creating self-organized internal dynamics that is loosely coupled with sensory motor dynamics. The loose coupling allows the creation of endogenous input stimulations, similar to real ones that could allow the internal system to sustain its internal dynamics and, eventually, reshape such dynamics while modifying endogenous input stimulations.

*Lalle* and *Dominey* [28.54] suggest the idea that mental imagery can be seen as a way for an autonomous system of generating internal representation and exploiting the convergence of different multimodal contingencies. That is, given a set of sensory-motor contingencies specific to many different modalities, learned by an autonomous agent in interaction with the environment, mental imagery constitutes the bridge toward even more complex multimodal convergence. The model proposed by the authors is based on the

hierarchical organization of the cortex and it is based on a set of interconnected artificial neural networks that control the humanoid robot iCub in tasks that involve coordination between vision, hand-arm control, and language. The chapter also highlights interesting relations between the model and neurophysiological and neuropsychological findings that the model can account for.

An extension of the neurocomputational model TRoPICAL (two route, prefrontal instruction, competition of affordances, language simulation) is proposed by [28.55] to implement an embodied cognition approach to mental rotation processes, a classic task in mental imagery research. The extended model develops new features that allow it to implement mental simulation, sensory prediction, as well as enhancing the model's capacity to encode somatosensorial information. The model, applied to a simulated humanoid robot (iCub) in a series of mental rotation tests, shows the ability to solve the mental rotation tasks in line with results coming from psychology research. The authors also claim the emergence of links between overt movements with mental rotations, suggesting that affordance and embodied processes play an important role in mental rotation capacities.

Starting from the fact that some evidence in experimental psychology has suggested that imagery ability is crucial for the correct understanding of social intention, an interesting study to investigate intention-from-movement understanding is presented in [28.56]. Authors' aim is to show the importance of including the more cognitive aspects of social context for further development of the optimal theories of motor control, with positive effects on robot companions that afford true interaction with human users. In the paper, the authors present a simple but thoroughly executed experiment, first to confirm that the nature of the motor intention leads to early modulations of movement kinematics. Second, they tested whether humans use imagery to read an agent's intention when observing the very first element of a complex action sequence.

A neural network model to produce an anticipatory behavior by means of a multimodal off-line Hebbian association is proposed in [28.57]. The model emulates a process of mental imagery, in which visual and tactile stimuli are associated during a long-term predictive simulation chain motivated by covert actions. Such model was studied by means of two experiments with a physical Pioneer 3-DX robot that developed a mechanism to produce visually conditioned obstacle avoidance behavior.

## 28.3 Experiments

In this section, we present three experimental studies that exemplify the capabilities and the performance improvements achievable by an *imagery*-enabled robot. Results of experimental tests with the simulator of the iCub humanoid robot platform are presented as evidence of the opportunities given by the use of artificial *mental imagery* in cognitive artificial systems.

The first study, [28.58], details a model of a controller, based on a dual network architecture, which allows the humanoid robot iCub to improve autonomously its sensorimotor skills. This is achieved by endowing the controller of a secondary neural system that by exploiting the sensorimotor skills already acquired by the robot, is able to generate additional *imaginary* examples that can be used by the controller itself to improve the performance through a simulated *mental training*.

The second study, [28.59], builds on the previous study showing that the robot could *imagine* or *mentally* recall and accurately execute movements learned in previous training phases, strictly on the basis of the verbal commands. Further tests show that data obtained with *imagination* could be used to simulate *mental training* processes, such as those that have been employed with human subjects in sports training, in order to enhance precision in the performance of new tasks through the association of different verbal commands.

The third study, [28.60], explored how the relationship between spatial mental imagery practice in a training phase could increase accuracy in sports related performance. The focus is on the capability to estimate, after a period of training with proprioceptive and visual stimuli, the position into a soccer field when the robot acquires the goal.

### 28.3.1 The Humanoid Robotic Platform: The iCub

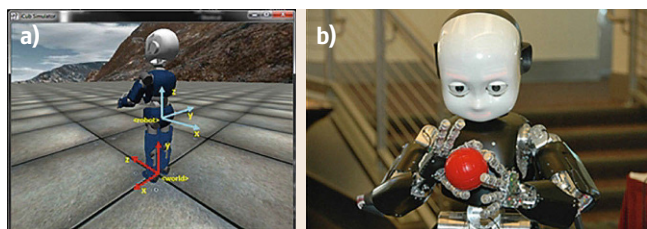
The cognitive robotic platform used for the experiments presented here is the simulation of the iCub humanoid robot controlled by artificial neural networks. The iCub

(Fig. 28.1) is an open-source humanoid robot platform designed to facilitate cognitive developmental robotics research as detailed in [28.2]. At the current state the iCub platform is a child-like humanoid robot 1.05 m tall, with 53 degrees of freedom (DoF) distributed in the head, arms, hands, and legs. The implementation used for the experiments presented here is a simulation of the iCub humanoid robot (Fig. 28.1). The simulator, which was developed with the aim to accurately reproduce the physics and the dynamics of the physical iCub [28.61], allows the creation of realistic physical scenarios in which the robot can interact with a virtual environment. Physical constraints and interactions that occur between the environment and the robot are simulated using a software library that provides an accurate simulation of rigid body dynamics and collisions.

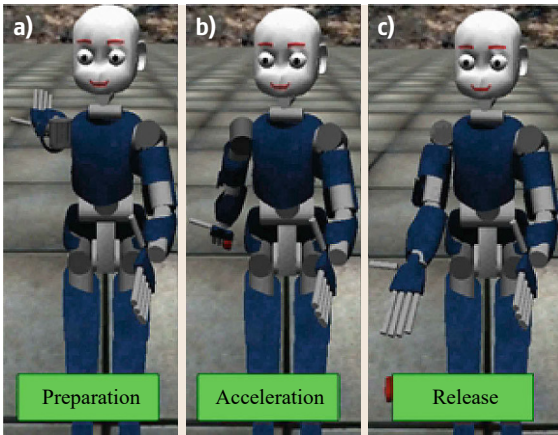
### 28.3.2 First Experimental Study: Motor Imagery for Performance Improvement

The first experimental study explored the application of mental simulation to robot controllers, with the aim to mimic the mental training techniques to improve the motor performance of the robot. To this end, a model of a controller based on neural networks was designed to allow the iCub to autonomously improve its sensorimotor skills.

The experimental task is to throw a small cube of side size 2 cm and weight 40 g as far as possible according to an externally given velocity for the movement. The task phases are shown in Fig. 28.2 and it is the realization of a ballistic action, involving the simultaneous movement of the right arm and of the torso, with the aim to throw a small object as far as possible according to an externally given velocity for the movement. Ballistic movements can be defined as rapid movements initiated by muscular contraction and continued by momentum [28.62]. These movements are typical in sport actions, such as throwing and jumping (e.g., a soccer kick, a tennis serve, or a boxing punch). In this experiment, we focus on two main features that characterize a ballistic movement: (1) it is executed by the brain with a predefined order, which is fully programmed before the actual movement realization and (2) it is executed as a whole and will not be subject to interference or modification until its completion. This definition of ballistic movement implies that proprioceptive feedback is not needed to control the movement and that its development is only based on starting conditions [28.63]. It should be noted here that since ballistic movements are



**Fig. 28.1a,b** The iCub humanoid robot platform: (a) The realistic simulator; (b) The real platform



**Fig. 28.2a–c** Three phases of the movement: **(a) Preparation:** The object is grabbed and shoulder and wrist joints are positioned at  $90^\circ$ ; **(b) Acceleration:** The shoulder joint accelerates until a given angular velocity is reached, while the wrist rotates down; **(c) Release:** the object is released and thrown away

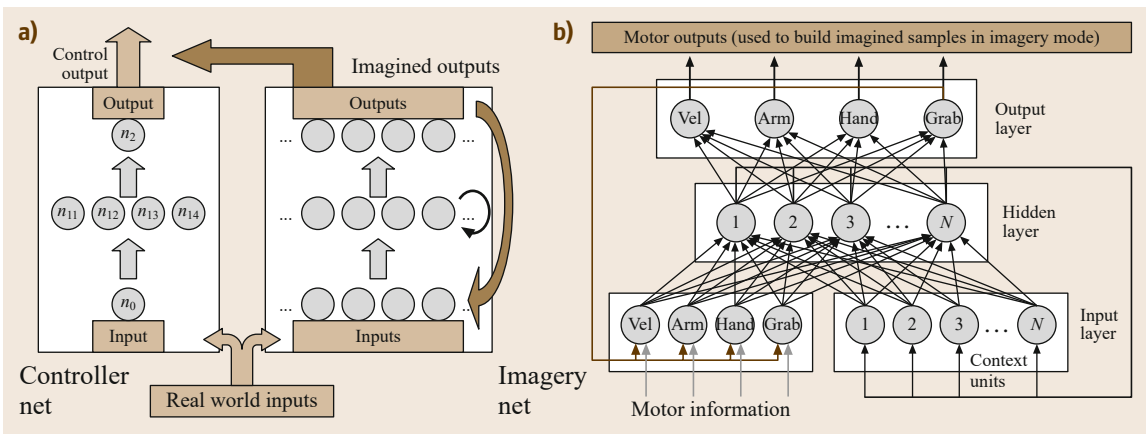
by definition not affected by external interferences, the training can be performed without considering the surrounding environment, as well as vision and auditory information.

To build the input–output training and testing sets, all values were normalized in the range  $[0, 1]$ .

To control the robot, we designed a dual neural network architecture, which can operate to improve autonomously the robot motor skills with techniques inspired by the ones that are employed with human subjects in sports training. This is achieved through two interconnected neural networks to control the robot: a FFNN that directly controls the robot’s joints and

an RNN that is able to generate additional *imaginary* training data by exploiting the sensorimotor skills acquired by the robot. The new data, in turn, can be used to generate additional *imaginary* examples that can be used by the controller itself to improve the performance through an additional learning process. We show that data obtained with artificial *imagination* could be used to simulate *mental training* to learn new tasks and enhance their performance.

The artificial neural network architecture that models the *mental training process* is detailed in Fig. 28.3a. It is a three-layer dual recurrent neural network (DRNN) with two recurrent sets of connections, one from the hidden layer and one from the output layer, which feed directly to the hidden layer through a set of context units added to the input layer. At each iteration (epoch), the context units hold a copy of the previous values of the hidden and outputs units. This creates an internal memory of the network, which allows it to exhibit dynamic temporal behavior [28.64]. It should be noted that in preliminary experiments, this architecture proved to show better performances and improved stability with respect to classical architectures, for example, *Jordan* [28.65] or *Elman* [28.66]. The DRNN comprises 5 output neurons, 20 neurons in the hidden layer, and 31 neurons in the input layer (6 of them encode the proprioceptive inputs from the robot joints and 25 are the context units). Neuron activations are computed according to (28.1) and (28.2). The six proprioceptive inputs encode, respectively, the shoulder pitch angular velocity (constant during the movement), positions of shoulder pitch and hand wrist pitch (at time  $t$ ), elapsed time, expected duration time (at time  $t$ ), and the grab/release command (0 if the object is grabbed, 1 otherwise). The five outputs encode the pre-



**Fig. 28.3a,b** Design of the cognitive architecture of the first experimental study: **(a)** The dual network architecture (FFNN + RNN). **(b)** Detail of RNN: *Brown connections* (recurrences and predicted output for the FFNN) are active only in *imagery mode*, meanwhile *light grey links* (external input–output from real world) are deactivated

dictions of shoulder and wrist positions at the next time step, the estimation of the elapsed time, the estimation of the movement duration, and the grab/release state. The DRNN retrieves information from joint encoders to predict the movement duration during the acceleration phase. The activation time step of the DRNN is 30 ms. The object is released when at time  $t_1$  the predicted time to release,  $t_p$ , is lower than the next step time  $t_2$ .

A functional representation of the neural system that controls the robot is given in Fig. 28.3a: a three-layer FFNN, which implements the actual motor controller, and of RNN. The RNN models the motor imagery and it is represented in detail in Fig. 28.3b. The choice of the FFNN architecture as a controller was made according to the nature of the problem, which does not need proprioceptive feedback during movement execution. Therefore, the FFNN sets the duration of the entire movement according to the given speed in input and the robot's arm is activated according to the duration indicated by the FFNN. On the other hand, the RNN was chosen because of its capability to predict and extract new information from previous experiences, so as to produce new imaginary training data, which are used to integrate the training set of the FFNN.

The FFNN comprises one neuron in the input layer, which is fed with the desired angular velocity of the shoulder by the experimenter, four neurons in the hidden layer, and one neuron in the output layer. The output unit encodes the duration of the movement given the velocity in input. To train the FFNN, we used a classic backpropagation algorithm as the learning process. The learning phase lasted  $1 \times 10^6$  epochs with a learning rate ( $\alpha$ ) of 0.2, without momentum (i. e.,  $\eta = 0$ ). The network parameters were initialized with randomly chosen values in the range  $[-0.1, 0.1]$ . The FFNN was trained by providing a desired angular velocity of the shoulder joint as input and the movement duration as output. After the movement duration is set by the FFNN, the arm is activated according to the desired velocity and for the duration indicated by the FFNN output.

Activations of hidden and output units  $y_i$  are calculated at a discrete time, by passing the net input  $u_i$  to the logistic function, as it is described in (28.1) and (28.2)

$$u_i = \sum_i (y_i w_{ij} - k_i) , \quad (28.1)$$

$$y_i = \frac{1}{1 - e^{-u_i}} . \quad (28.2)$$

As learning process, we used the classic backpropagation algorithm, the goal of which is to find optimal values of synaptic weights that minimize the error  $E$ , defined as the error between the teaching sequences and

the output sequences produced by the network. The error function  $E$  is calculated as follows

$$E = \frac{1}{2} \sum_{i=1}^p \|y_i - t_i\|^2 , \quad (28.3)$$

where  $p$  is the number of outputs,  $t_i$  is the desired activation value of the output unit  $i$ , and  $y_i$  is the actual activation of the same unit produced by the neural network, calculated using (28.2) and (28.3). During the training phase, synaptic weights at learning step  $n + 1$  are updated using the error calculated at the previous learning step  $n$ , which in turn depend on the error  $E$ . Activations of hidden and output units  $y_i$  are calculated by passing the net input  $u_i$  to the function, as it is described in (28.2) and (28.3). The backpropagation algorithm updates link weights and neuron biases, with a learning rate ( $\alpha$ ) of 0.2 and a momentum factor ( $\eta$ ) of 0.6, according to the following equation

$$\Delta w_{ij}(n + 1) = \eta \delta_j y_j + \alpha \Delta w_{ij}(n) , \quad (28.4)$$

where  $y_j$  is the activation of unit  $j$ ,  $\alpha$  is the learning rate, and  $\eta$  is the momentum.

To calculate the error at first step of the backpropagation algorithm, initial values of back links are initialized to one. The network parameters are initialized with randomly chosen values in the range  $[-0.1, 0.1]$ .

The experimental study is divided into two phases: in the first phase the FFNN is trained to predict the duration of the movement for a given angular velocity in input. Meanwhile, using the same movements the RNN was trained by a simple heuristic to predict its own subsequent sensorimotor state. To this end, joint angle information over time was sampled in order to build 20 input–output sequences corresponding to different directions of the movement. In addition, in order to model the autonomous throw of an object, the primitive action to grab/release was also considered in the motor information fed to the network. In the second phase, the RNN operates in *offline* mode and, thus, its prediction is made according only to the internal model built during the training phase. Normalized joint position of shoulder pitch, torso yaw, and hand wrist pitch are the proprioceptive information for input and output neurons. Another neuron implements the grab/release command, respectively, with values 1 and 0.

In this study, we intended to test the impact of mental training in action performance in a different speed range that was not experienced before. Because of this, we split both the learning and testing dataset into two subsets according to the duration of the movement:

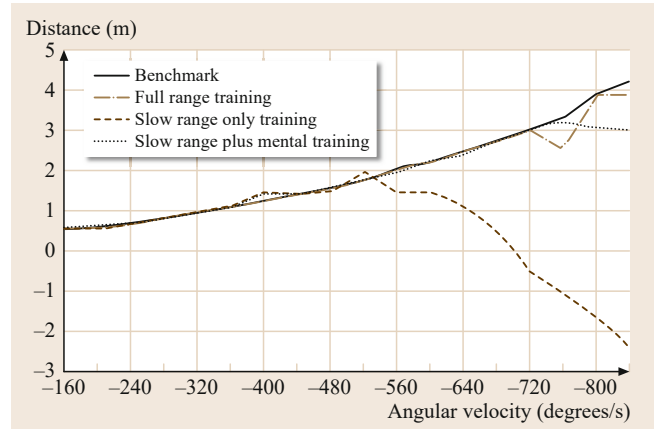


- *Fast range* subset comprises examples that last less than 0.3 s
- *Slow range* subset comprises all the others (i. e., those that last more than 0.3 s).

A reference benchmark controller (RBC) was used to build training and testing datasets for the learning phase. The training dataset comprises data collected during the execution of the RBC with angular velocity of shoulder pitch ranging from  $-150$  to  $-850^\circ/\text{s}$  and using a step of  $-25^\circ/\text{s}$ . Thus, the learning dataset comprises 22 input–output series. Similarly, the testing dataset was built in the range from  $-160$  to  $-840^\circ/\text{s}$  and using a step of  $-40^\circ/\text{s}$ , and it comprises 18 input–output series. The learning and testing dataset for the FFNN comprises one pair of data: the angular velocity of the shoulder as input and the execution time, that is, duration, as desired output. The learning and testing datasets for the RNN comprises sequences of 25 elements collected using a time-step of 0.03 s. All data in both learning and testing datasets are normalized in the range  $[0, 1]$ . Results are shown for the testing set only.

To test the mental training, we compared results on three different case studies:

1. *Full range*: For benchmarking purposes, it is the performance obtained by the FFNN when it is trained using the full range of examples (*slow + fast*)
2. *Slow range only training*: The performance obtained by the FFNN only when it is trained using only the slow-range subset. This case stressed the generalization capability of the controller when it is tested with the fast range subset
3. *Slow range plus mental training*: In this case the two architectures operate together as a single hierarchical architecture, in which first both nets are trained



**Fig. 28.4** Comparison of the distance reached by the object after throwing with the FFNN as controller and different training approaches. Negative values represent the objects falling backward

with the slow range subset, then the RNN runs in *mental imagery mode* to build a new dataset of fast examples for the FFNN that is incrementally trained this way.

As expected, the FFNN is the best controller for the task if the full range is given as training, thus, it is the ideal controller for the task (Table 28.1). But, not surprisingly, in Table 28.2 it is shown that the FFNN it is not able to generalize with the fast range when it is trained with the slow range only.

These results show that generalization capability of the RNN helps to feed the FFNN with new data to cover the fast range, simulating *mental training*. In fact, the FFNN, trained only with the slow subset, is not able to foresee the trend of duration in the fast range; this implies that fast movements last longer than needed and, because the inclination angle is over  $90^\circ$ , the object falls backward (Fig. 28.4).

**Table 28.1** Full-range training: comparison of average results of feedforward and recurrent artificial neural nets

Test	Feedforward net				Recurrent net			
	Duration		Release point		Duration		Release point	
	<i>s</i>	Error%	Degree	Error%	<i>s</i>	Error%	Degree	Error%
Slow	0.472	1.75	-30.718	-6.46	0.482	3.60	-33.345	-11.96
Fast	0.202	0.87	-31.976	-6.34	0.194	5.67	-28.088	-22.18
Full	0.307	1.21	-31.486	-6.39	0.306	4.86	-30.132	-18.21

**Table 28.2** FFNN: Comparison of average performance improvement with artificial mental training

Test	Slow range only training				Slow range plus mental training			
	Duration		Release point		Duration		Release point	
	<i>s</i>	Error%	Degree	Error%	<i>s</i>	Error%	Degree	Error%
Slow	0.474	1.38	-30.603	-7.88	0.471	1.74	-30.774	-8.18
Fast	0.247	26.92	-64.950	-111.72	0.188	7.12	-20.901	-35.89
Full	0.335	16.99	-51.593	-71.34	0.298	5.03	-24.741	-25.11

The FFNN failure in predicting temporal dynamics is explainable by the simplistic information used to train the FFNN, which seems to be not enough to reliably predict the duration time in a faster range, never experienced before. On the contrary, the greater amount of information that comes from the proprioception and the fact that the RNN has to integrate over time those information in order to perform the movement, makes the RNN able to create a sort of internal model of the robot's body behavior. This allows the RNN to better generalize and, therefore guide the FFNN in enhancing its performance.

This interesting aspect of the RNN can be partially unveiled by analyzing the internal dynamic of the neural network, which can be done by reducing the complexity of the hidden neuron activations through a principal component analysis. Figure 28.5a, for example, presents the values of the first principal component at the last timestep, that is, after the neural network has finished the throwing movement, for all the test cases, both slow and fast showing that the internal representations are

very similar, also in the case in which the RNN is trained with the slow range only. Interestingly, the series shown in Fig. 28.8 are highly correlated with the duration time, which is never explicitly given to the neural network. The correlation is 97.50 for the full-range training, 99.41 for slow only and 99.37 for slow plus mental training. This result demonstrates that the RNN is able to extrapolate the duration time from the input sequences and to generalize when operating with new sequences never experienced before.

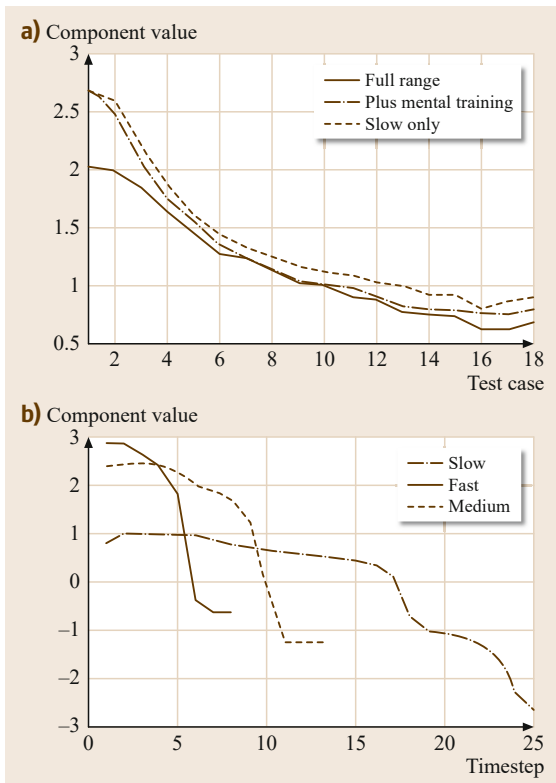
Similarly, Fig. 28.5b shows the first principal component of the RNN in relation to different angular velocities: slow being the slowest test velocity, medium the fastest within the slow range, and fast the fastest possible velocity tested in the experiment. As can be seen, the RNN is able to uncover the similarities in the temporal dynamics that link slow and fast cases. Hence, it is finally able to better approximate the correct trajectory of joint positions also in a situation not experienced before.

### 28.3.3 Second Experimental Study: Mental Training Evoked by Language

In this experimental study, we dealt with motor imagery and how verbal instruction may evoke the ability to imagine movements, already seen before or new ones obtained by combination of past experiences. These imagined movements either replicate the expected new movement required by verbal commands or correspond in accuracy to those learned and executed during training phases. Motor imagery is defined as a dynamic state during which representations of a given motor act are internally rehearsed in working memory without any overt motor output [28.67].

This study extends the first experimental study presented above, by focusing on the integration of auditory stimuli in the form of verbal instructions, to the motor stimuli already experienced by the robot in past simulations. Simple verbal instructions are added to the training phase of the robot, in order to explore the impact that linguistic stimuli could have in its processes of mental imagery practice and subsequent motor execution and performance. In particular, we tested the ability of our model to use imagery to execute new orders, obtained combining two single instructions. This study has been inspired by embodied language approaches, which are based on evidence that language comprehension is grounded in the same neural systems that are used to perceive, plan, and take action in the external world.

Figure 28.6 presents pictures of the action with the iCub simulator, which was commanded to execute the four throw tasks according to the verbal command is-



**Fig. 28.5** (a) Hidden units' activation analysis. Lines represent final values of the first principal component for all test cases; (b) hidden units' activation analysis. Lines represent the values of first principal component for a slow velocity, a medium velocity and a fast velocity

sued. The basic task is the same presented in Fig. 28.2; in this case the torso is also moving to obey to the verbal commands *left* and *right*.

The neural system that controls the robot is a three-layer RNN with the architecture proposed by *Elman* [28.66]. The Elman RNN adds in the input layer a set of *context units*, directly connected with the middle (hidden) layer with a weight of one (i. e., directly copied). At each time step, the input is propagated in a standard feedforward fashion, and then a learning rule is applied. The fixed back connections result in the context units always maintaining a copy of the previous values of the hidden units (since they propagate over the connections before the learning rule is applied). This creates an internal state of the network, which allows it to exhibit dynamic temporal behavior. To model mental imagery the outputs related with the motor activities are redirected to corresponding inputs.

Similarly to the previous experimental study, after the learning phase in which *real* data collected during simulator execution was used to train the RNN and for comparison with *imagined* data, we tested the ability of the RNN architecture to model *mental imagery*. As before, this was achieved by adding other back connections from motor outputs to motor inputs; at the same time connections from/to joint encoders and motor controllers are deactivated. This setup is presented in Fig. 28.7, where red connections are the ones active only when the *imagery mode* is on, while green connections are deactivated, including the motor controller. Specific neurons, one for each verbal instruction, were included in the input layer of the RNN in order for it to take into account these commands, while the sensorimotor information is directed to the rest of the neurons in the input layer. The RNN architecture implemented, as presented in Fig. 28.7, has 4 output units, 20 units in the hidden layer, and 27 units in the input layer, 7 of

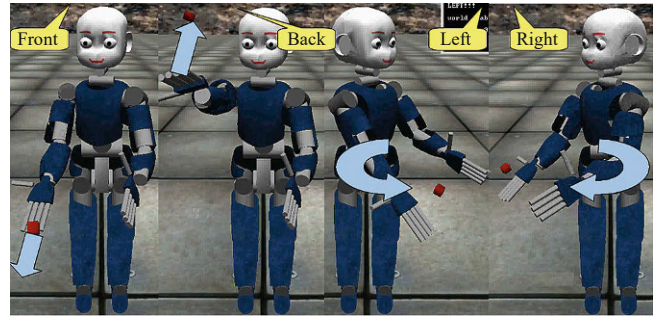


Fig. 28.6 Examples of the iCub simulator in action: pictures of the execution of throw tasks

them encode the proprioceptive inputs from the robot's joints and 20 are the context units, that is, are back links of hidden units, they only copy the value from output of upper unit to the input of lower unit. The learning algorithm and parameters are the same as the second experiment.

As proprioceptive motor information, we take into account just the following three joints, shoulder pitch, torso yaw, and hand wrist pitch. In addition, in order to model the throw of an object, the primitive action to grab/release was also considered in the motor information fed to the network. Visual information was not computed and speech input processing was based on standard speech recognition systems.

Using the iCub simulator, we performed two experiments:

- The first experiment aimed to evaluate the ability of the RNN to model artificial mental imagery. It was divided into two phases: in the first phase the network was trained to predict its own subsequent sensorimotor state. The task was to throw in different directions (forward, left, right, back) a small object that was placed in the right hand of the robot,

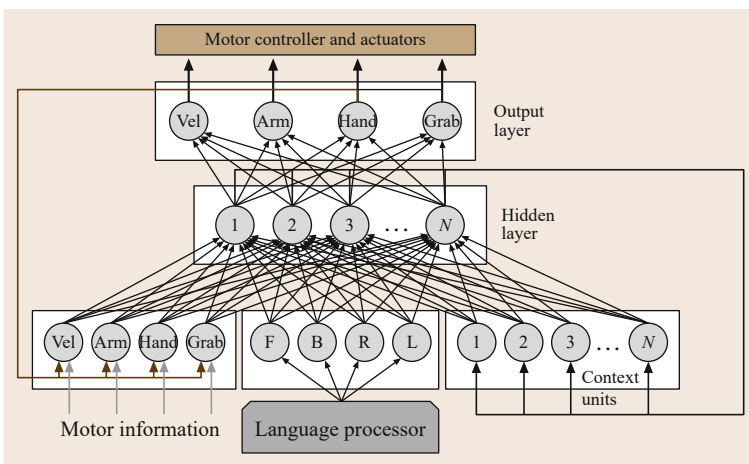


Fig. 28.7 Recurrent neural network architecture used in the second experimental study. *Brown* connections are active only in *imagery mode*, meanwhile *light grey* connections are deactivated

which is able to grab and release it. To this end the RNN was trained using the proprioceptive information collected from the robot. The proprioceptive information consisted of sensorimotor data (i. e., joint positions) and of verbal commands given to the robot according to directions. In the second phase, we tested the ability of the RNN to model mental imagery providing only the auditory stimulus (i. e., the verbal commands) and requiring the network to obtain sensorimotor information from its own outputs.

- The goal of the second experiment was to test the ability of the RNN to imagine how to accomplish a new task. In this case we had three phases: in the first phase (*real training*), the RNN was trained to throw front and just to move left and right (with no throw). In the second phase (*imagined action*), the RNN was asked to imagine its own subsequent sensorimotor state when the throw command is issued together with a side command (left or right). In the final phase (*mental training*), the input/output series obtained are used for an additional *mental training* of the RNN. After the first and third phase, experimental tests with the iCub simulator were made to measure the performance of the RNN to control the robot.

In this experiment, we tested the ability of the RNN to recreate its own subsequent sensorimotor state in absence of external stimuli. In Fig. 28.8, we present a comparison of training and imagined trajectories of learned movements according to the verbal command issued:

1. Shows results with the FRONT command
2. With the BACK command
3. With the RIGHT command
4. For the LEFT command.

Imagined trajectories are accurate with respect to the ones used to train the robot only in Fig. 28.8b, we notice a slight difference between imagined and training positions of the arm. This difference can be attributed to the fact that the BACK command is the only one that does not require the arm to stop early in throwing the object. In other words, the difference is related to the timing of the movement rather than to the accuracy. Results show that the RNN is able to recall the correct trajectories of the movement according to the verbal command issued. The trajectories are the sequence of joint positions adopted in the movements.

The second test was conducted to evaluate the ability of the RNN to build its own subsequent sensorimotor states when it is asked to accomplish new tasks not experienced before. In this case, the RNN was trained

only to throw front and to move right and left (without throwing). To allow the RNN to generalize, training examples were created using an algorithm that randomly chose joint positions not involved in the movement, that is, when throwing, the torso joint had a fixed position that was randomly chosen. The same was true for arm joints when moving right and left.

Test cases, presented in Fig. 28.9, were composed using two commands (e.g., *throw* together with *right* or *left*). In our experiments, we tested two different approaches in the language processing. In this test two commands were computed at the same time, so that input neurons associated with throw and right (or left) were fully activated at the same time with value 1.

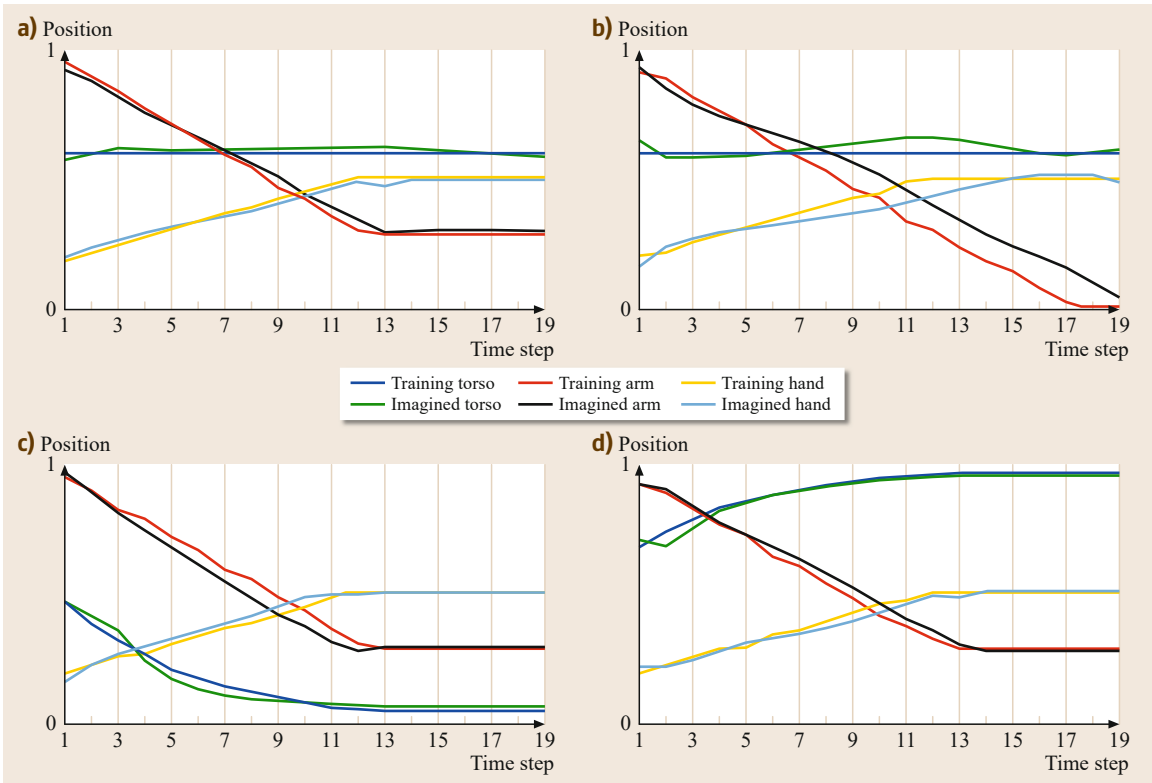
Results of the mental training experiment are presented in Fig. 28.10, which show the error of torso and arm joint position with respect to the ideal ones. The *before mental training* column presents the results of tests made without additional training, the *after mental training* column shows results after the simulated mental training, the *imagined action only* column refers to totally imagined data (i. e., when the RNN predicts its own subsequent input). Comparing results before and after mental training an improvement in precision of dual command execution could be noticed, this should be accounted to the additional training that helps the RNN to operate in a condition not experienced before.

It should be noticed also that the throw right task has worse performance compared to that of throw left with iCub simulations, but the same result is not achieved in imagined only action mode. This could be mainly explained by the influence of real proprioceptive information coming from robot joints that modifies the ideal behavior expected by the RNN, as evidenced by the comparison between imagined only and the real tests. In fact, we noticed that when a right command is issued the robot torso is initially moved on the left for few time-steps and then it turns right. Since the total time of the movement is due to the arm movement to throw, the initial error for the torso could not be recovered.

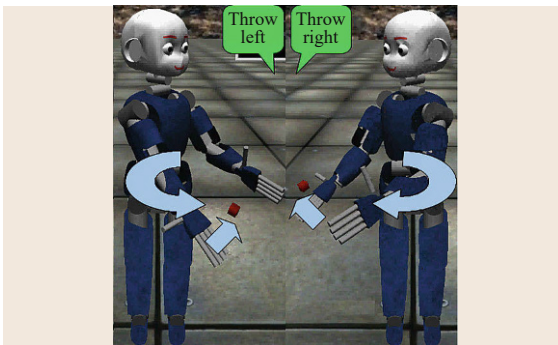
### 28.3.4 Third Experimental Study: Spatial Imagery

For this experiment the environment is a square portion of a soccer field, whose length and width are both 15 m. At one end is placed a goal 1.94 m wide, as can be seen in Fig. 28.11b, which is represented all in blue to contrast with the background and to be easily recognized. The robot can be positioned anywhere in this square and, as starting position, the ball is placed in front of his left foot (Fig. 28.11a).

The neural system that controls the robot is a fully connected RNN with 16 hidden units, 37 input units,

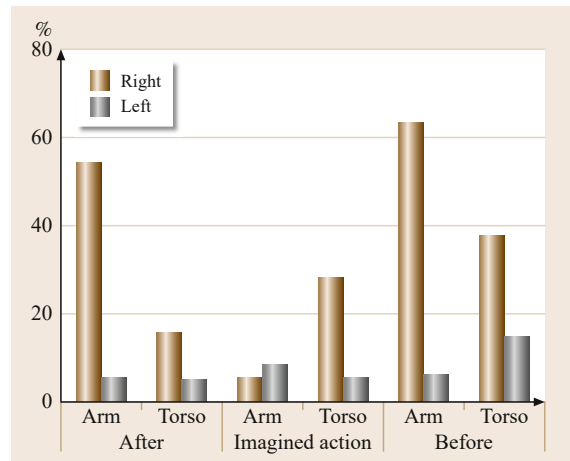


**Fig. 28.8** First test: A comparison of training and imagined trajectories of learned movements



**Fig. 28.9** Pictures present the execution of the two composed tasks: throw left and throw right. As test cases for autonomous learning with iCub simulator, the actions to throw left or right are now obtained by learning and then combining the basic actions of throw and move left or right

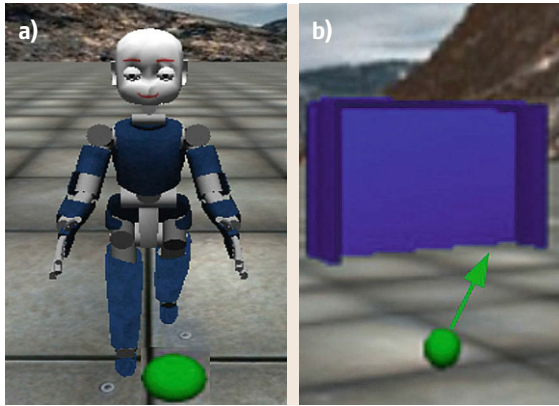
and 6 output units. The main difference between a standard feedforward neural network and the RNN is that in the latter case, the training set consists in a series of input–output sequences. The RNN architecture allows the robot to learn dynamical sequences of actions as they develop in time. The goal of the learning process is to find optimal values of synaptic weights that minimize the error, defined as the error between the teaching



**Fig. 28.10** Second test: Autonomous learning of new combined commands, error of torso and arm joint positions with respect to ideal ones

sequences and the output sequences produced by the network. Figure 28.12 presents the neural system. Input variables are as follows:

- The angle of, respectively, the neck (left–right movement), the torso (left–right movement), and



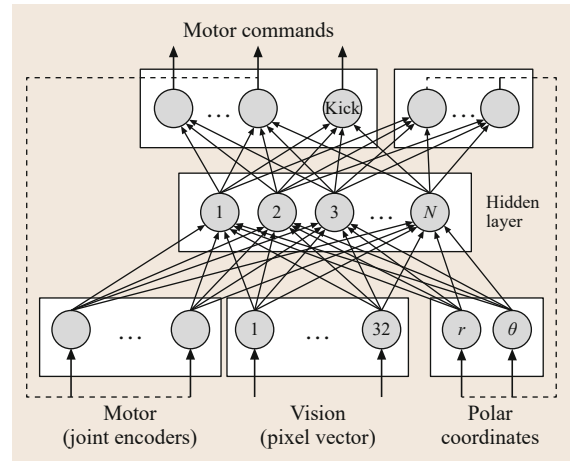
**Fig. 28.11a,b** iCub robot simulator: kicking the ball (a), the goal (b)

the body rotation joint (a joint attached to the body that allows the robot to rotate on its vertical axis).

- Visual information is provided through a vector of 32 bits, which encodes a simplified visual image obtained by the eyes and allows the robot to locate the goal in its visual field.
- Polar coordinates of the robot with respect to the goal (represented by the radius and the polar angle).

The six outputs are, respectively, the polar coordinates of the robot with respect to the goal, the desired angle of the neck, the torso and the body rotation joint, and an additional binary output that makes the robot kick the ball when its value is 1. In the learning phase, this output was set to 1 only when the motion ends and the robot must kick the ball to the goal. All input and output variables are normalized in  $[0, 1]$ . Angle ( $\theta$ ) ranges from  $-90^\circ$  to  $90^\circ$ , while radius ( $r$ ) could vary from 0 to 15 m. As angle and radius values are normalized in  $[0, 1]$ , the  $0^\circ$  (i. e., when the goal is in front of the robot) equals to 0.5.

For training the neural network, we used the *backpropagation through time algorithm* (BPTT), which is typically used to train neural networks with recurrent nodes for time-related tasks. This algorithm allows a neural network to learn the dynamical sequences of input–output patterns as they develop in time. Since we are interested in the dynamic and time-dependent processes of the robot–object interaction, an algorithm that allows to take into account dynamic events is more suitable than the standard backpropagation algorithm [28.68]. For a detailed description of the BPTT algorithm, see also [28.69]. The main difference between a standard backpropagation algorithm and the BPTT is that in the latter case the training set consists in a series of input–output sequences, rather than in a single input–output pattern. The BPTT allows the robot to learn sequences of actions. The goal of the learning pro-



**Fig. 28.12** Neural network architecture for spatial position estimation. *Dotted lines* (recurrences from output to input) are active only when the network operates in imagery mode

cess is to find optimal values of synaptic weights that minimize the error  $E$ , defined as the error between the teaching sequences and the output sequences produced by the network. The error function  $E$  is calculated as follows

$$E = \sum_s \sum_t \sum_i \left( (y_{its}^* - y_{its}) (y_{its} - (1 - y_{its})) \right)^2, \quad (28.5)$$

where  $y_{its}^*$  is the desired activation value of the output unit  $i$  at time  $t$  for the sequence  $s$  and  $y_{its}$  is the actual activation of the same unit produced by the neural network, calculated using (28.1) and (28.2). During the training phase, synaptic weights at learning step  $n + 1$  are updated using the error  $\delta_i$  calculated at the previous learning step  $n$ , which in turn depend on the error  $E$ , according to (28.4).

To build the training set, we used the iCub simulator primitives to position the robot in eight different locations of the field. For each position the robot rotated to acquire the goal by means of a simple search algorithm. During the movement, motor and visual information were sampled in input–output sequences of 20 time steps. The command to kick the ball is issued after the acquisition of the target (the middle of the goal). The kicking movement was preprogrammed. Then, the neural network was trained to predict the next sensory state (excluding the visual input) by means of a backpropagation through time algorithm for 50 000 epochs, after which the mean squared error (MSE) error in estimating the six output variables was 0.0056.

In a preliminary study the robot was first controlled by the same algorithm used for collecting the train series (*controlled condition*), the neural network was not

controlling the robot, but used only to estimate the position coordinates. In a second trial the robot was fully moved by means of the neural network (*autonomous* condition), which directly controlled the joints (i. e., network outputs were sent to neck, torso, and the body rotation actuators) as well as the kick command (moves the leg down). The result of this preliminary study shows that the *autonomous* condition performs better (i. e., 5% average error less) than the *controlled* one in terms of position estimation. For this reason, for the main experiment we employed the *autonomous* condition only, which is also more in line with the theoretical background presented so far.

The aim of the experiment was to test the generalization performance of the neural network and to evaluate the use of visual and proprioceptive information for the estimation of the robot position with respect to the goal. As testing phase, the robot was positioned in the same positions used for training (*learning set*) to verify the quality of the learning and, then, in eight new positions (*testing set*), that it did not experienced before, to evaluate the generalization capability of the model.

The structure of the experiment is divided into two phases: in the first phase the network is trained to predict its own subsequent sensorimotor state. In the second phase the network is tested on the robot, in interaction with the environment.

Table 28.3 shows the percentage error between imagined positions and real position. The error is evaluated as the percentage with respect to the real positions using the polar coordinates. All values of the imagined positions in Table 28.3 were obtained with *autonomous* control. Analyzing errors in testing position, we can see that the error is very high for position 8, this is because the robot fails to acquire the target after it makes a wrong move at the beginning and the goal goes out of sight. It should be said that the robot has not been trained to find the goal when it is not at least in part in

its visual field. The same happened when the robot was in position 2 and 7 of the learning set. The robot misses the 50% of the scores, but it is worth to mention that errors were mostly made when the goal was very distant (i. e., more than 8 m) and even a little error in the position leads the ball out of the goal.

Figures 28.13 and 28.14 graphically summarize the results showing the environment with the 8 real and imagined positions of the train set and the test set, respectively. As the figures show, overall the robot is able to estimate its position in the environment to a good extent.

Table 28.4 report the distance, evaluated using Cartesian coordinates, between the real positions and the first and last estimated positions in the imagined series. This evaluation gives further evidence that the failures on some positions are due to the fact the robot is not trained to find the target when it is out of sight. Indeed, the first imagined position is quite good, but after the wrong movement the robot is no longer able to see the goal and the visual input becomes all zeros, thus, it has no way to recover.

Figures 28.15 and 28.16 reports the entire *imagined path*, along with markers for first and last imagined positions and the actual position. The *imagined path* is the fictitious path that is composed of all positions imagined according to the movements made. The *imagined paths* for positions with very high error are not depicted to avoid confusion. It can be noted that the accuracy of imagined positions gradually improves while the robot performs the movement to aim the center of the goal and shoot. The average improvement is 0.43 m for learning set and 0.56 m for testing set. According to this result, the use of proprioceptive motor information, coming from the autonomous body movements, influences the robot imagination, and it often helps to better estimate its position in the field.

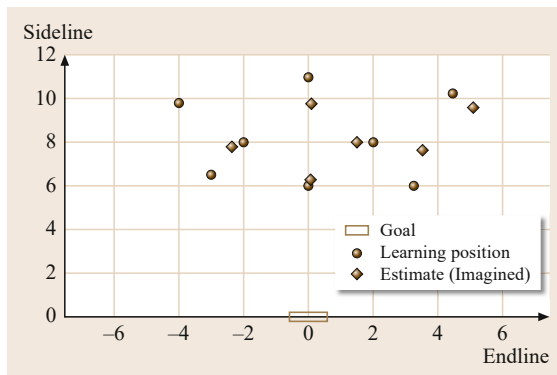
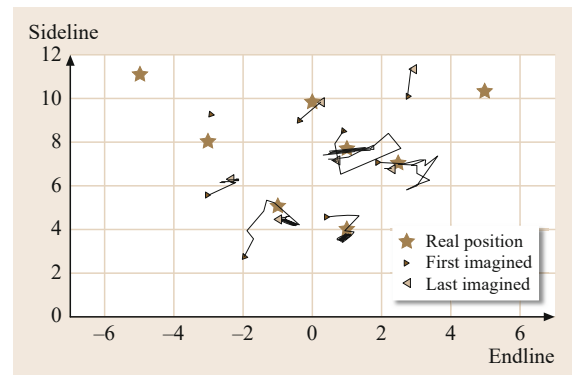
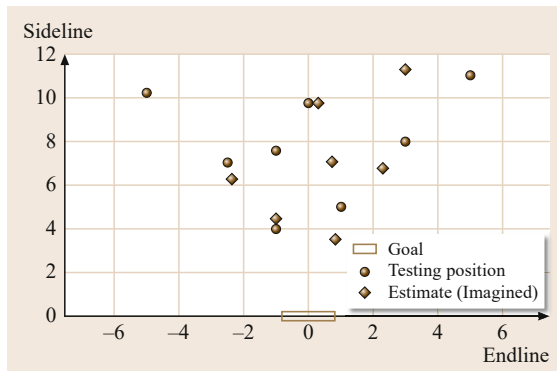
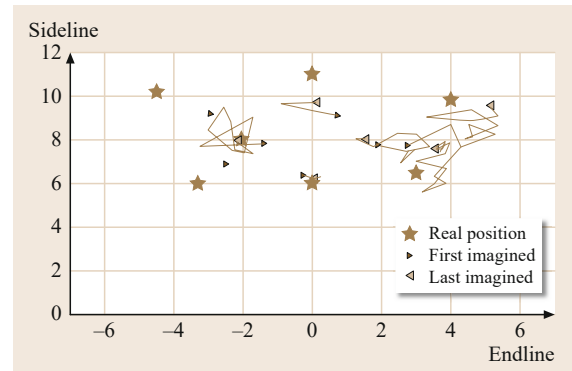
**Table 28.3** Real versus imagined positions for learning and testing sets (normalized polar coordinates) and error percentages

N	Learning					Testing					
	Actual		Imagined		Error%	Actual		Imagined		Error%	
	Angle	Radius	Angle	Radius		Angle	Radius	Angle	Radius		
1	0.50	0.400	0.50	0.415	2.07	0.58	0.275	0.57	0.242	11.95	
2	0.34	0.457	0.76	0.905	122.59	0.44	0.34	0.43	0.301	11.57	
3	0.64	0.477	0.64	0.560	11.18	0.61	0.496	0.61	0.476	4.32	
4	0.58	0.550	0.56	0.545	4.04	0.39	0.57	0.39	0.447	21.56	
5	0.42	0.550	0.41	0.543	3.38	0.54	0.511	0.53	0.476	7.35	
6	0.62	0.706	0.66	0.724	10.48	0.5	0.653	0.51	0.652	2.74	
7	0.37	0.743	0.76	0.905	130.32	0.65	0.757	0.58	0.78	20.24	
8	0.50	0.733	0.50	0.648	11.47	0.36	0.806	0.77	0.904	125.53	
Average					36.94	Average					25.66
Excluding positions 2 & 7					7.1	Excluding position 8					11.39

**Table 28.4** Distance (in meters) of imagined positions with respect to actual ones, with improvement from first to last estimate

$N$	First estimate static vision only		Last estimate vision and motor information		Improvement with body movement	
	Learning	Testing	Learning	Testing	Learning	Testing
1	0.48	0.78	0.23	0.49	0.25	0.29
2	1.18	2.46	13.66	0.59	-12.48	1.87
3	1.28	0.59	1.24	0.32	0.04	0.27
4	0.28	2.46	0.45	1.84	-0.17	0.62
5	0.64	0.86	0.14	0.56	0.5	0.3
6	2.39	0.95	1.16	0.27	1.23	0.68
7	1.84	2.22	14.52	2.3	-12.68	-0.08
8	1.98	2.73	1.27	14.99	0.71	-12.26
Avg	1.26	1.63	4.09	2.67	0.43 <sup>a</sup>	0.56 <sup>a</sup>

<sup>a</sup> Average without positions: 2 and 7 (learning)/8 (testing), see text for details

**Fig. 28.13** The eight real and imagined positions in the field for the learning set**Fig. 28.15** Learning set: *Imagined paths*, with first and last position estimates, compared to real locations in the field**Fig. 28.14** The eight real and imagined positions in the field for the training set**Fig. 28.16** Training set: *Imagined paths*, with first and last position estimates, compared to real locations in the field



## 28.4 Conclusion

Despite the wide range of potential applications, the fast-growing field of cognitive robotics still poses several interesting challenges, both in terms of mechanics and autonomous control. Indeed, in the new humanoid platforms, sensor and actuator arrangements determine a highly redundant system, which is traditionally difficult to control, and hard-coded solutions often do not allow further improvement and flexibility of controllers. Letting those robots free to learn from their own experience is very often regarded as the unique real solution that will allow the creation of flexible and autonomous controllers for humanoid robots in the future.

In this chapter, we presented the work done so far to explain the concept of *motor imagery* and *mental simulation* as a fundamental capability for cognitive models, based on artificial neural networks, which allow the humanoid robot iCub to autonomously improve its sensorimotor skills via simulated *mental imagery* mechanisms.

Three experimental studies with the iCub platform simulator were presented to show that the application of imagery inspired mechanisms can significantly improve the cognitive behaviors of the robot, even in ranges not experienced before. The results presented, in conclusion, allow imagining the creation of novel algorithms and cognitive systems that implement even better and with more efficacy the concept of artificial mental training. Such a concept appears very useful in robotics, for at least two reasons: it helps to speed up the learning process in terms of time resources by reducing the number of real examples and real movements performed by the robot. Besides the time issue, the reduction of real examples is also beneficial in terms of costs, because

it similarly reduces the probability of failures and damages to the robot while keeping the robot improving its performance through mental simulations. In the future, we speculate that imagery techniques might be applied in robotics not only for performance improvement, but also for the creation of safety algorithms capable to predict dangerous joints' positions and to stop the robot's movements before that critical situation actually occurs.

From a technological point of view, this chapter aims to support the better understanding of mental imagery as a potential breakthrough for cognitive robot engineering principles. Such principles can be applied to go further in the development of artificial cognitive systems capable to better interact with the environment and refine their cognitive motor skill in an open-ended process. These robots will be able to reason, behave, and interact in a human-like fashion, thanks to the integration of the capabilities to mentally represent the physical and social world, resemble experiences, and simulate actions. The imagery-enabled cognitive robotic agents will be able to handle and manipulate objects and tools autonomously, to cooperate and communicate with other robots and humans, and to adapt their abilities to changing internal, environmental, and social conditions.

**Acknowledgments.** This work was partially supported by UK EPSRC Project BABEL and the European Commission FP7 Projects: POETICON++ (ICT-288382) within the Cognitive Systems, Interaction, Robotics unit (FP7 ICT Challenge 2), ROBOT-ERA (ICT-288899) within the ICT for Health, Ageing Well, Inclusion and Governance unit (FP7 ICT Challenge 5).

## References

- 28.1 A.M. Turing: Computing machinery and intelligence, *Mind* **59**, 433–460 (1950)
- 28.2 G. Metta, L. Natale, F. Nori, G. Sandini, D. Vernon, L. Fadiga, C. von Hofsten, K. Rosander, M. Lopes, J. Santos-Victor, A. Bernardino, L. Montesano: The iCub humanoid robot: An open-systems platform for research in cognitive development, *Neural Netw.* **23**, 1125–1134 (2010)
- 28.3 D. Gouaillier, V. Hugel, P. Blazevic, C. Kilner, J. Monceaux, P. Lafourcade, B. Marnier, J. Serre, B. Maisonnier: Mechatronic design of NAO humanoid, *Proc. IEEE Int. Conf. Robot. Autom.* (2009) pp. 764–774
- 28.4 Y. Sakagami, R. Watanabe, C. Aoyama, S. Matsunaga, N. Higaki, K. Fujimura: The intelligent ASIMO: System overview and integration, *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Vol. 3 (2002) pp. 2478–2483
- 28.5 R. Pfeifer, J. Bongard, S. Grand: *How The Body Shapes The Way We Think: A New View of Intelligence* (MIT Press, Cambridge 2007)
- 28.6 S.M. Kosslyn: *Image and Brain: The Resolution of The Imagery Debate* (MIT Press, Cambridge 1996)
- 28.7 S. Gardini, C. Cornoldi, R. De Beni, A. Venneri: Cognitive and neuronal processes involved in sequential generation of general and specific mental images, *Psychol. Res.* **73**, 633–643 (2009)
- 28.8 S.D. Slotnick, W.L. Thompson, S.M. Kosslyn: Visual memory and visual mental imagery recruit common control and sensory regions of the brain, *Cogn. Neurosci.* **3**, 14–20 (2012)
- 28.9 A.Z.J. Zeman, S. Della Sala, L.A. Torrens, V.E. Gountouna, D.J. McGonigle, R.H. Logie: Loss of imagery phenomenology with intact visuo-spatial task per-

- formance: A case of blind imagination, *Neuropsychologia* **48**, 145–155 (2010)
- 28.10 M. Kozhevnikov: Cognitive styles in the context of modern psychology: Toward an integrated framework of cognitive style, *Psychol. Bull.* **133**, 464–481 (2007)
- 28.11 M. Kozhevnikov, S. Kosslyn, J. Shephard: Spatial versus object visualizers: A new characterization of visual cognitive style, *Mem. Cogn.* **33**, 710–726 (2005)
- 28.12 M. Wilson: Six views of embodied cognition, *Psychon. Bull. Rev.* **9**(4), 625–636 (2002)
- 28.13 A. Clark, D. Chalmers: The extended mind, *Analysis* **58**(1), 10–23 (1998)
- 28.14 L. Munari: How the body shapes the way we think – A new view of intelligence, *J. Med. Pers.* **7**, 110–111 (2009)
- 28.15 A. Ishai, L.G. Ungerleider, G.V. Haxby: Distributed neural systems for the generation of visual images, *Neuron* **28**(3), 379–390 (2000)
- 28.16 M. Jeannerod: The representing brain. Neural correlates of motor intention and imagery, *Behav. Brain Sci.* **17**(2), 187–245 (1994)
- 28.17 J. Decety, M. Jeannerod, C. Prablanc: The timing of mentally represented actions, *Behav. Brain Res.* **34**(1/2), 35–42 (1989)
- 28.18 M. Jeannerod, J. Decety: Mental motor imagery: A window into the representational stages of action, *Curr. Opin. Neurobiol.* **5**(6), 727–732 (1995)
- 28.19 G. Hesselrow: The current status of the simulation theory of cognition, *Brain Res.* **1428**, 71–79 (2012)
- 28.20 D.M. Wolpert: Computational approaches to motor control, *Trends Cogn. Sci.* **1**(6), 209–216 (1997)
- 28.21 X. Skoura, A. Vinter, C. Papaxanthis: Mentally simulated motor actions in children, *Dev. Neuropsychol.* **34**(3), 356–367 (2009)
- 28.22 B. Steenbergen, M. van Nimwegen, C. Crajé: Solving a mental rotation task in congenital hemiparesis: Motor imagery versus visual imagery, *Neuropsychologia* **45**(14), 3324–3328 (2007)
- 28.23 M. Jeannerod: Neural simulation of action: A unifying mechanism for motor cognition, *Neuroimage* **14**, S103–S109 (2001)
- 28.24 J. Munzert, B. Lorey, K. Zentgraf: Cognitive motor processes: The role of motor imagery in the study of motor representations, *Brain Res. Rev.* **60**(2), 306–326 (2009)
- 28.25 R. Ramsey, J. Cumming, D. Eastough, M.G. Edwards: Incongruent imagery interferes with action initiation, *Brain Cogn.* **74**(3), 249–254 (2010)
- 28.26 K.D. Markman, W.M. Klein, J.A. Suhr: *Handbook of Imagination and Mental Simulation* (Psychology, New York 2009)
- 28.27 G. Yue, K.J. Cole: Strength increases from the motor program: Comparison of training with maximal voluntary and imagined muscle contractions, *J. Neurophysiol.* **67**(5), 1114–1123 (1992)
- 28.28 T. Mulder, S. Zijlstra, W. Zijlstra, J. Hochstenbach: The role of motor imagery in learning a totally novel movement, *Exp. Brain Res.* **154**(2), 211–217 (2004)
- 28.29 L. Yáguiez, D. Nagel, H. Hoffman, A.G. Canavan, E. Wist, V. Hömberg: A mental route to motor learning: Improving trajectorial kinematics through imagery training, *Behav. Brain Res.* **90**(1), 95–106 (1998)
- 28.30 M. Takahashi, S. Hayashi, Z. Ni, S. Yahagi, M. Favilla, T. Kasai: Physical practice induces excitability changes in human hand motor area during motor imagery, *Exp. Brain Res.* **163**(1), 132–136 (2005)
- 28.31 P.L. Jackson, M.F. Lafleur, F. Malouin, C. Richards, J. Doyon: Potential role of mental practice using motor imagery in neurologic rehabilitation, *Arch. Phys. Med. Rehabil.* **8**, 1133–1141 (2001)
- 28.32 J.A. Verbunt, H.A. Seelen, F.P. Ramos, B.H. Michielsen, W.L. Wetzelaer, M. Moennekens: Mental practice-based rehabilitation training to improve arm function and daily activity performance in stroke patients: A randomized clinical trial, *BMC Neurol.* **8**(C), 7 (2008)
- 28.33 S.J. Page, P. Levine, A. Leonard: Mental practice in chronic stroke: Results of a randomized, placebo-controlled trial, *Stroke A J. Cereb. Circ.* **38**(4), 1293–1297 (2007)
- 28.34 D.M. Nilsen, G. Gillen, A.M. Gordon: Use of mental practice to improve upper-limb recovery after stroke: A systematic review, *Am. J. Occup. Ther.* **64**(5), 695–708 (2010)
- 28.35 A.A. Sheikh, E.R. Korn: *Imagery in Sports and Physical Performance* (Baywood, Amityville 1994)
- 28.36 B.S. Rushall, L.G. Lippman: The role of imagery in physical performance, *Int. J. Sport Psychol.* **29**(1), 57–72 (1998)
- 28.37 T. Morris, M. Spittle, A.P. Watt: *Imagery in Sport* (Human Kinetics, Champaign 2005)
- 28.38 R. Weinberg: Does imagery work? Effects on performance and mental skills, *J. Imag. Res. Sport Phys. Act.* **3**(1), 1–22 (2008), <http://www.degruyter.com/view/jjirspsa.2008.3.1/jirspsa.2008.3.1.1025/jirspsa.2008.3.1.1025.xml>
- 28.39 S.A. Hamilton, W.J. Fremouw: Cognitive-behavioral training for college basketball free-throw performance, *Cogn. Ther. Res.* **9**(4), 479–483 (1985)
- 28.40 D. Smith, P. Holmes, L. Whitmore, D. Collins, T. Devonport: The effect of theoretically-based imagery scripts on field hockey performance, *J. Sport Behav.* **24**(4), 408–419 (2001)
- 28.41 A. Guillot, E. Nadrowska, C. Collet: Using motor imagery to learn tactical movements in basketball, *J. Sport Behav.* **32**(2), 27–29 (2009)
- 28.42 G. Wei, J. Luo: Sport expert's motor imagery: Functional imaging of professional motor skills and simple motor skills, *Brain Res.* **1341**, 52–62 (2010)
- 28.43 M. Asada, K. Hosoda, Y. Kuniyoshi, H. Ishiguro, T. Inui, Y. Yoshikawa, M. Ogino, C. Yoshida: Cognitive developmental robotics: A survey, *Auton. Ment. Dev. IEEE Trans.* **1**(1), 12–34 (2009)
- 28.44 A. Di Nuovo, V.M.D. De La Cruz Marocco: Special issue on artificial mental imagery in cognitive systems and robotics, *Adapt. Behav.* **21**(4), 217–221 (2013)
- 28.45 J. Tani: Model-based learning for mobile robot navigation from the dynamical systems perspective, *IEEE Trans. Syst. Man. Cybern.* **26**(3), 421–436 (1996)

- 28.46 D. Roy, K.-Y. Hsiao, N. Mavridis: Mental imagery for a conversational robot, *Syst. Man. Cybern. Part B Cybern. IEEE Trans.* **34**(3), 1374–1383 (2004)
- 28.47 T. Ziemke, D.-A. Jirenghed, G. Hesslow: Internal simulation of perception: A minimal neuro-robotic model, *Neurocomputing* **68**(0), 85–104 (2005)
- 28.48 R. Nishimoto, J. Tani: Development of hierarchical structures for actions and motor imagery: A constructivist view from synthetic neuro-robotics study, *Psychol. Res. PRPF* **73**(4), 545–558 (2009)
- 28.49 O. Gigliotta, G. Pezzulo, S. Nolfi: Evolution of a predictive internal model in an embodied and situated agent, *Theory Biosci.* **130**(4), 259–276 (2011)
- 28.50 H. Svensson, S. Thill, T. Ziemke: Dreaming of electric sheep? Exploring the functions of dream-like mechanisms in the development of mental imagery simulations, *Adapt. Behav.* **21**(4), 222–238 (2013)
- 28.51 A. Kaiser, W. Schenck, R. Möller: Solving the correspondence problem in stereo vision by internal simulation, *Adapt. Behav.* **21**(4), 239–250 (2013)
- 28.52 F. Chersi, F. Donnarumma, G. Pezzulo: Mental imagery in the navigation domain: A computational model of sensory-motor simulation mechanisms, *Adapt. Behav.* **21**(4), 251–262 (2013)
- 28.53 H. Iizuka, H. Ando, T. Maeda: Extended homeostatic adaptation model with metabolic causation in plasticity mechanism – toward constructing a dynamic neural network model for mental imagery, *Adapt. Behav.* **21**(4), 263–273 (2013)
- 28.54 S. Lalle, P.F. Dominey: Multi-modal convergence maps: From body schema and self-representation to mental imagery, *Adapt. Behav.* **21**(4), 274–285 (2013)
- 28.55 K. Seepanomwan, D. Caligiore, G. Baldassarre, A. Cangelosi: Modelling mental rotation in cognitive robots, *Adapt. Behav.* **21**(4), 299–312 (2013)
- 28.56 D. Lewkowicz, Y. Delevoye-Turrell, D. Bailly, P. Andry, P. Gaussier: Reading motor intention through mental imagery, *Adapt. Behav.* **21**, 315–327 (2013)
- 28.57 W. Gaona, E. Escobar, J. Hermosillo, B. Lara: Anticipation by multi-modal association through an artificial mental imagery process, *Conn. Sci.* **27**, 68–88 (2015)
- 28.58 A. Di Nuovo, D. Marocco, S. Di Nuovo, A. Cangelosi: Autonomous learning in humanoid robotics through mental imagery, *Neural Netw.* **41**, 147–155 (2013)
- 28.59 A. Di Nuovo, V.M. De La Cruz, S. Di Nuovo, D. Marocco, A. Cangelosi: Mental practice and verbal instructions execution: A cognitive robotics study, *Proc. Int. Jt. Conf. Neural Netw. (IJCNN)* (2012) pp. 2771–2776
- 28.60 A. Di Nuovo, D. Marocco, S. Di Nuovo, A. Cangelosi: A neural network model for spatial mental imagery investigation: A study with the humanoid robot platform iCub, *Proc. Int. Jt. Conf. Neural Netw. (IJCNN)* (2011) pp. 2199–2204
- 28.61 V. Tikhonoff, A. Cangelosi, P. Fitzpatrick, G. Metta, L. Natale, F. Nori: An open-source simulator for cognitive robotics research: The prototype of the iCub humanoid robot simulator, *Proc. 8th Workshop Perform. Metr. Intell. Syst. (PerMIS)* (2008) pp. 57–61
- 28.62 R. Bartlett: *Introduction to Sports Biomechanics: Analysing Human Movement Patterns* (Routledge, London 2007)
- 28.63 E.P. Zehr, D.G. Sale: Ballistic movement: Muscle activation and neuromuscular adaptation, *Can. J. Appl. Physiol.* **19**(4), 363–378 (1994)
- 28.64 M.M. Botvinick, D.C. Plaut: Short-term memory for serial order: A recurrent neural network model, *Psychol. Rev.* **113**(2), 201–233 (2006)
- 28.65 M.I. Jordan: Attractor dynamics and parallelism in a connectionist sequential machine, *Proc. 8th Annu. Conf. Cogn. Sci. Soc.* (1986) pp. 531–546
- 28.66 J.L. Elman: Finding structure in time, *Cogn. Sci.* **14**(2), 179–211 (1990)
- 28.67 J. Decety: The neurophysiological basis of motor imagery, *Behav. Brain Res.* **77**, 45–52 (1996)
- 28.68 D.E. Rumelhart, J.L. McClelland: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (MIT Press, Cambridge 1986)
- 28.69 P.J. Werbos: Consistency of HDP applied to a simple reinforcement learning problem, *Neural Netw.* **3**(2), 179–189 (1990)

# Dynamical M

## 29. Dynamical Models of Cognition

Mary Ann Metzger

Models of cognition address properties of the mind by formulating cognitive processes such as memory, perception, inference, and comprehension of language. Dynamical models of cognition ascribe importance to time and complexity, both of which bring context to behavior. Temporal processes bring into the moment the possibility of memory, feedback, the effects of nonlinear recursion, and the generation of expectation. Complexity brings the possibility of stable patterns of coordination emerging from interaction of subprocesses.

In some models, time and complexity have provided a bridge between thought and action, a basis by which to characterize thought and action as inextricably combined. These models hold that action is a component of perception, or that thought and action are inseparable, or that thought and action act in concert, two sides of the same coin serving to reduce the uncertainty about the nature of events.

This chapter provides a review of several models of cognition in terms of their dynamical features, including models not generally included in the dynamical tradition, such as ART and ACT-R. It focuses on the manner in which each model treats time and complexity, thought, and action. It provides a glimpse into the methods of model development and analysis associated with the various approaches to modeling cognitive processes.

<b>29.1 Dynamics</b> .....	639
29.1.1 Time and Complexity .....	640
29.1.2 Cognition and Action .....	641
<b>29.2 Data-Oriented Models</b> .....	641
29.2.1 Methods .....	641
29.2.2 Example: Motor Coordination .....	642
29.2.3 Example: Decision Under Risk .....	643
29.2.4 Summary .....	644
<b>29.3 Cognition and Action Distinct</b> .....	644
29.3.1 Recognition Memory Model .....	644
29.3.2 Adaptive Control of Thought – Rational .....	645
29.3.3 Artificial Neural Networks Methods .....	646
29.3.4 Adaptive Resonance Theory .....	647
29.3.5 Summary .....	648
<b>29.4 Cognition and Action Intrinsically Linked</b> .....	648
29.4.1 Methods .....	648
29.4.2 Embodied Cognition .....	650
29.4.3 Motor Theory .....	651
29.4.4 Simulation Theory .....	651
29.4.5 Free Energy Theory .....	652
29.4.6 Evolution of Cognitive Search .....	653
29.4.7 Summary .....	653
<b>29.5 Conclusion</b> .....	653
<b>References</b> .....	655

### 29.1 Dynamics

The properties of the mind are formulated and quantified by models of cognition. Examples of these will be summarized here from the dynamic point of view, which will bring in some of their points of contact with fields other than psychology, from philosophy to engineering. Models will be described in ways which highlight the manner in which they address dynamical features of mental activity.

The dynamic point of view involves two properties of cognitive models, one methodological the other substantive. The methodological aspect centers on matters of time and complexity, the question being in what way does the model incorporate concepts and methods which address temporal change as formulated in nonlinear dynamic systems theory and the science of complex systems. The substantive aspect involves the manner in

which the models address the relationship between the mental and the physical, cognition and action.

### 29.1.1 Time and Complexity

Philosophical properties of dynamical models have been considered in comparison to models which have been called computational. The former models are distinguished by elements of timing and complexity, as might characterize the production and comprehension of speech. The latter by stability, for which timing is arbitrary, as in associations among mental symbols, or application of rules of logic. Another way to express the difference is that for dynamical models of complex processes, several subprocesses act simultaneously and interactively, while for computational models, the subparts of a modeled process act as independent modules, each of which communicates with other modules by taking input prepared by a previously active module, processing it, and making the result available to the next module.

Similar distinctions have appeared in philosophical debates using related terminology as in emergent versus reductionist, dynamical versus generative, and connectionist versus artificial intelligence. Advocates for dynamical models have argued for dynamical models being superior in the sense of being more complete, perhaps more general, than their computational counterparts, as in the following quote, which refers to the computational approach as Hobbesian [29.1]:

“[...] Every cognitive process unfolds in continuous time, and the fine temporal detail calls out for scientific accounting. Moreover, many cognitive structures are essentially temporal: like utterances, they exist only as change in time. Often, getting the timing right is critical to the success of cognitive performance; this is especially so when in direct interaction with surrounding events.

Hobbesian computational models have made a bet that cognitive phenomena can be described in a way that abstracts away from the full richness of real time, replacing it with discrete orderings over formal states.”

The emphasis on continuous time is a stringent requirement for research strategies in psychology. Incorporating time as an essential component of data is something easily done, but in many areas of psychology this means not continuous time, but time-sampling, in which the process in question is measured at intervals, resulting not in a continuous signal, but in discrete time, a *time series* of observations. Much of the terminology characteristic of research related to dynamical models arises from the requirement of de-

scribing and drawing conclusions from dynamical time series.

With dynamical time-series data, the researcher can apply graphical techniques to succinctly lay out the time-course or *trajectory* of the process, the appearance of both stable patterns, called *attractors*, and the paths, *transients*, to and from attractors. The graphical techniques have mathematical counterparts and additional mathematical techniques are available to quantify and summarize features of the time series. In this way the dynamics of the process, its *rule of evolution*, may be quantified, classified, and understood.

Accepting time and time series as fundamental requirement of data emphasizes a focus not only on cognitive *entities*, symbols and rules of manipulation, but also on cognitive *performances*, *perceiving*, *remembering*, *conversing*. From this point of view, the scientific objective should be one of describing the time series of processes and correlates of cognitive behavior and discovering the rule of evolution by which the cognitive performance unfolds over time [29.2].

In an example from the production and comprehension of speech, dynamical defines the basic unit of data to be a continuous linguistic signal, while generative defines it as discrete phonetic segments. A summary of the implication of the two approaches, continuous-dynamical versus segmented generative is given in the following quotation [29.3]:

“[...] a fundamental mistake of the generative paradigm is its assumption that phonetic segments are formal symbol tokens. This assumption permitted the general assumption that language is a discrete formal system. This bias forced generative phonology to postulate a phonetic space that is closed and contains only stable symbolic objects. We show that theories of phonetics satisfying these constraints have little to no support from phonetic evidence and thus that the formal-language assumption is surely incorrect.”

There is an implication that dynamical models of cognition enjoy superior status, its associated laws being deeper and more generally applicable to behavior. That assertion has been philosophically evaluated by examining dynamical models considered examples of greater and lesser laws. Greater laws are defined to be more widely applicable than lesser. Lesser laws might be descriptive accounts of particular mental dynamics, or they might rely on concepts for which temporal factors are negligible. Dynamical models of cognition were shown to exemplify both greater and lesser laws, leading to the conclusion that time and complexity alone are not sufficient to distinguish greater versus lesser laws [29.4].

### 29.1.2 Cognition and Action

The dynamical versus computational distinction has been characterized as competing answers to the mind-body problem. In this view, dynamical theories address mind-body as a single phenomenon with cognition and action being two faces of the same process. In contrast, computational models have the dualist view that mind and body are different entities; the mind manipulates information to formulate goals and plans for the body to execute. The following is an elaboration of the dynamical view [29.5]:

“All that we know we have constructed within ourselves from the unintelligible fragments of energy impacting our senses as we move our bodies through the world. This process of intention is transitive in the outward thrust of the body in search of desired future states; it is intransitive in the dynamic construction of predictions of the states in the sensory cortices by which we recognize success or failure in achievement. The process is phenomenologically experienced in the action-perception cycle. Enactment is through the serial creation of neurodynamic activity patterns in brains, by which the self of mind-brain-body comes to know the world first by shaping the self to an approximation of the sought-for input, and then by

assimilating those shapes into cognition and meaning.”

In discussions of dynamical versus computational models of cognition, illustrations have been examples of action, for which timing of interacting subsystems is an essential feature versus cognition, for which timing enters as a secondary or negligible feature. The contrast between dynamical and computational might be illustrated as the contrast between a musical performance and the music sheet, or between a conversation and a transcript, in each case the former being action, the latter cognition. Hence, there are two conflicting points of view on the relationship between action and cognition, they are either sequentially cooperating subsystems, or they are integrated, communicating subsystems.

This chapter examines several theories of cognition focusing where possible on ways in which they characterize cognition and action and on ways in which they incorporate timing and complexity. The topic is divided into three major categories. The first concerns methods in data-oriented models, which make few general claims, but rather summarize a phenomenon or a special purpose model. The second concerns general models of cognition for which cognition and action are treated as distinct and separate processes, that is, knowledge can be developed and transformed without reference to action. The third concerns models for which cognition and action are intrinsically linked.

## 29.2 Data-Oriented Models

Data analytic methods for detecting and illustrating specific dynamical properties of a process set the stage for development of more general theories. The methods are generally derived from complexity science and nonlinear dynamic systems theory. Complexity science is the study of processes comprised of interacting subprocesses. Nonlinear dynamic systems theory is comprised of mathematical methods, usually based on differential or difference equations, used for inference about features of the trajectory of nonlinear recursive processes. A nonlinear recursive process is one for which exponential powers of previous events in the process (*feedback*) systematically determine subsequent events. In practical application, the distinction between the two methods of analysis is not always useful, since they characterize and quantify phenomena of the dynamical processes in compatible ways. Data-oriented studies aim to identify and quantify dynamical phenomena which appear during the unfolding of an objective or theoretical process. Their goal is to set data into a context which

highlights dynamical properties, thus establishing requirements for theory.

Dynamical properties of a process are revealed in a sequence of measurements on its important variables, a time series which can be used to describe and draw inferences about the trajectory of the process. Of the numerous graphical and mathematical approaches to understanding the dynamics of a process, this chapter will consider just a few central concepts related to constructing features and drawing conclusions from time series and dynamical models. More details, definitions, methods, and objectives can be found in [29.6–8].

### 29.2.1 Methods

Part of the appeal of dynamical models is they allow the modeling of complex-appearing processes, with simple, deterministic rules of evolution which generate periods of patterned behavior interwoven with transitional periods of random-seeming variation. Other dynamical

models generate simple patterns from complex systems. Three of the many methods associated with dynamical modeling will be of interest here, quantifying attractors, concepts of potential, and scale invariance.

### Quantifying Attractors

Attractors are persistent patterns which arise in a dynamical process. As an example, a whirlpool is an attractor in a flowing stream. When and how attractors are exhibited by a dynamical process often depends on the value of a quantity analogous to energy of the process. That quantity is represented by the *control parameter*. One of the ways to formalize a dynamical system is to identify the attractors, then specify manner in which the control parameter governs the trajectory of the process into and out of attractors. In the example of the flowing stream, the control parameter could be the rate of flow, when the stream flows faster or slower old whirlpools might disappear and new ones come into being. The idea is to incorporate the control parameter and nonlinear feedback into the rule of evolution.

Attractors are the most perceptible aspect of a dynamical process primarily due to their duration. Other aspects of the process are fleeting and ephemeral, but attractors can last long enough to be noticed and named. Take, for example, the process underlying the reversals of an optical illusion in which the attractors are the two possible perceptions of a staircase, rising or descending. Most of the time the perception is of one or the other, each of which is easy to describe. The time of transition between attractors is rapid and the transient ephemeral. During the transient, the staircase is neither rising nor descending, and is not easily described or named. Transient and attractor intuitively may seem two different processes, yet the dynamics of process in the visual system are presumed to be the same whether the perception is of the staircase rising, falling, or transient. It is that constant process which is captured in its rule of evolution. The assumption of dynamical systems analysis is that all phenomena arise from a rule of evolution which remains unchanged as the system evolves through attractors and transients. As the process continues through time, it visits the attractors, whose persistence constitutes the phenomenological experience of the process. While in an attractor, the system generates and maintains a pattern of increased predictability, equated with lower surprise, sometimes equated with emergent phenomena.

### Potential

Dynamical processes may be drawn into attractors and may stay in or escape from an attractor based on energetic properties. Systematic energetic properties are measured by a potential function  $V$  the basic parameter

of which is the control parameter. Thus, the dynamics, motion, and phenomenology of the system is governed by the control parameter, via the potential function at any point in time. There is a component of optimization, that is, of the system having a quality of always moving toward a lower potential.

A method for quantifying the relation between the control parameter and the attractors of a dynamical process starts with a formula for potential  $V$  associated with each point in the space possibly lying on a trajectory of the system, that is, in a space defined by the axes which correspond to the system variables and the control parameter. The rule is that systems always move in a direction to reduce  $V$ . The attractors will therefore be located at points for which the potential is at a local or global minimum which might be conceived, respectively, as a shallow or deep valley. When the system reaches a point for which  $V$  is a minimum, it tends to remain there in an attractor to be moved out of that attractor only by random or systematic fluctuation both of which might be increased by additional energy, hence the persistence of attractors. Some energy is required to keep the process in operation and drive it to its deepest attractor, but too much energy can make the system less predictable by causing it to jump in and out of attractors.

It is the essence of attractors that they persist over time and so they are perceived as patterns or entities. Local attractors last for a time comparatively shorter than global attractors. Thus, global attractors have the effect of staving off the effects of time. That is, the system remains patterned and therefore more predictable over an extended period. Comparing to transients, transients happen in time whereas attractors happen over time.

### Scale Invariance

Nonlinear dynamic systems analysis can often reveal that a process or object has the same form whether it is viewed overall or microscopically. When such a relationship occurs, it is described by the equivalent terms of fractal, self-similar, or scale invariant. Examples of scale invariance occur widely in nature, often being visual examples of trees or landscapes, but the concept of scale invariance also applies to the mathematics of the underlying processes, so a model without a visual counterpart may be described as scale invariant.

### 29.2.2 Example: Motor Coordination

Applied to bilateral motor coordination, a data-oriented method of complexity science, *coupled oscillators*, was applied specifically to bilateral coordination of hand motion [29.9]. The analysis of the purely the motor

phenomenon has served as inspiration for applying the same method of analysis to cognitive phenomena. In the hand-motion study, a participant with hands on table is instructed to start moving both index fingers in a given pattern at a given rate, then on cue, speed up or slow down. The model relies on a fundamental variable  $\phi$ , the observed phase difference between the two fingers. When in phase (toward and away from each other),  $\phi = 0$ . When out of phase (both point the same direction),  $\phi = \pi$ . With the application of complexity theory and the theory of coupled oscillators, it was possible to derive specific predictions of amplitudes and frequencies critical for the system. That is, the theory predicted accurately for individuals at what frequency in-phase or out-of-phase patterns would appear, and additionally what changes in amplitude of finger motion occurred as the process moved between out-of-phase and in-phase. The fundamental equation used to quantify the locations and depths of the attractors of the system was derived and given as

$$V = -a \cos \phi - b \cos 2\phi, \quad (29.1)$$

where potential  $V$  is a function of the phase difference  $\phi$ , and constants  $a$  and  $b$ . The trajectory of coordination of hand motion occurs predominantly in a direction which minimizes  $V$ , with small fluctuations. The function  $V$  has two distinct minima, a shallow, local minimum, and a deep global minimum. The values of  $\phi$  at which these minima occur can be found by taking the derivative of  $V$  with respect to  $\phi$ , setting it equal to zero and solving for  $\phi$ . These are the values of  $\phi$  associated with two attractors of the process. When in operation, the system generates only two attractors, a strong one at  $\phi = 0$ , associated with a global minimum of  $V$ ; a weak one at  $\phi = \pi$  associated with a local minimum. Transition between phases is governed by the control parameter, the speed of oscillation, and shows unidirectional hysteresis. That is, it is more likely to fall into the global attractor than to escape it. While the process is in an attractor, it appears patterned, either in phase or out of phase. When the trajectory takes the system between attractors, it is in a transient and appears uncoordinated. For any participant, the precise frequencies at which the two attractors occur and the ease of moving between attractors is captured by the relationship of  $|b|$  to  $|a|$ , a quantity which can independently be determined for each participant. The description so derived has been shown to be consistent with details of coordination of bilateral hand movement.

Although the application obviously concerns only coordinated motor systems, the method of coupled oscillators has come into play, at least metaphorically in theories of infant cognitive development. The develop-

mental applications rely on the emergence of patterns of phase coordination which arise naturally out of properties of motor structures. A pattern, however, gives the impression of independent existence. Some theories of cognitive development claim some of the phenomena of the infant's cognitive development, ideas about the permanence of objects, for example, may simply be emergent properties wholly a product of interaction of motor systems.

### 29.2.3 Example: Decision Under Risk

A second example of a data-oriented approach offers an integration of several theories and results in the field of decision under risk [29.10]. To summarize, using a modification of terminology: Established results begin with the concept of the subjective value ( $y$ ) of a gamble, a quantity which can be determined from a person's choices among gambles. Within a wide range, for a rational person, subjective value of a gamble is equal to the objective expected value ( $x$ ), negative for losses, positive for gains. It might be called Rational Value Theory (RVT). But much research has shown people often act as if  $y$  has been modified by additional subjective evaluations of winning and losing and other features of the gambling experience, yielding a measure of utility. When  $y$  equals utility for each value of  $x$ , it can be said choice of gambles is governed by Expected Utility Theory (EUT). In a third theory, acknowledging that losses and gains are often relative,  $y$  of the gamble can be profoundly affected, even reversed, by the context in which the gamble is presented (*framing*). When  $y$  can be reversed from a preference to an aversion by, for example, changing the framing from a context of gain to a context of loss, it is an example of Prospect Theory (PT). Each theory yields a characteristic pattern in the graph of  $y$  against  $x$ . The patterns RVT, EUT, and PT can all be observed within an individual. Deviations from RVT have been associated with emotional involvement, where least emotional involvement described by RVT, moderate by EUT, and high by PT, critical facts for dynamical systems analysis.

Dynamical systems analysis of the choices among gambles begins with the assumption that the three graphs of  $y$  against  $x$  constitute the observed attractors of the process of choice among gambles. For RVT, the graph would be a 45° line,  $y = x$  through origin. For EUT the graph is ogive shaped with varying degrees of steepness, often asymmetric. For PT, over the mid-range of  $x$ , the graph is S-shaped, also somewhat asymmetric. Taking just the upper and lower arms of the S,  $y$  is then a two-valued function of  $x$ , representing reversal of preferences due to framing. The goal



is to embed the three attractors on a surface in three dimensions by introducing a third axis, the control parameter  $a$ .

The control parameter  $a$  is defined as an indicator of amount of emotional involvement. It is included as a third dimension, placing the three  $y$  by  $x$  graphs on a surface in three-dimensional space and in a specific relation to each other. Graphs for RVT, EUT, and PT are arranged, respectively, from linear to S-shape, at  $a = 0$ ,  $a = \text{moderate-value}$ , and  $a = \text{high-value}$ . As is characteristic in nonlinear dynamic systems analysis, the control parameter is hypothesized to govern movement between the attractors; the least emotional involvement associated with linear RVT, moderate involvement with ogival EUT, and high involvement with discontinuities of PT. When the three curves are arranged along the third dimension,  $a$ , their graphs suggest the cross-sections of the folded surface of a cusp catastrophe, for which a standard quantification exists.

For processes the trajectories of which lie on a cusp-catastrophe surface, (29.2) applies. It gives the formula for finding the potential ( $V$ ) associated with any triple of points ( $y, x, a$ ). To obtain the formula for the surface, it is only necessary to find the values of ( $y, x, a$ ) which make  $V$  a minimum. This is a straightforward process of taking the partial derivative of  $V$  with respect to  $y$ , setting the derivative equal to zero and solving the result to express  $y$  as a function of  $x$  and  $a$ . The graph of the resulting function will be the cusp catastrophe surface upon which the graphs of the three theories can be fit

$$V = y^4 + ay^2 + xy. \quad (29.2)$$

The value of this analysis by cusp catastrophe is not only that it represents several types of betting choice on a single surface, describing all three with a single formula but also that it lays out a dynamics which might

take place within an agent. This suggests time-series studies of a single agent with varying levels of emotional involvement over might profitably address dynamical movement among attractors. Such an approach might reveal individual differences, intermediate attractors, a bifurcation point at which the curve becomes 2-valued, and possibly hysteresis in the 2-valued range, thus posing challenges for continued development of theories of decision under risk.

#### 29.2.4 Summary

The examples illustrate two of the many dynamical approaches of data-oriented research. In the first, the approach quantifies coordination of limb movement, applying the model of coupled oscillators, well-understood in complexity science. The model enabled predictions about both phenomenological and quantitative details of coordination of hand movement, that is, about the nature of attractors and transients. This adds to theory by clearly delineating an approach to characterize the nature of emergent phenomena. The second approach is to begin with known attractors, as in the three styles of risk-taking characterizing subjective value as a function of objective value, then propose to place the functions in a space, arranged along a dimension of a control parameter. The approach envisions a single surface on which the three attractors lie. The resulting relationships add to theory by delineating the forms of other possible attractors of the process and processes which might go on within an individual. The first theoretical analysis concerns only action. In the second, study concerns only cognition. In both types of approaches, the goal is to end up with an understanding of the attractors the nature of the transients, and the control parameter, using methods of complexity science and nonlinear dynamic systems theory.

### 29.3 Cognition and Action Distinct

The philosophical contrast between dynamical and computational theories of cognition rests to some degree on their respective theoretical link between cognition and action. For computation, the link is proposed to be a minimal relation between modules. Mental modules manipulate information and issue goals for a motor module to carry out. The claim that modular theories are not dynamical deserves examination. Consider three models, each apparently exclusively concerned with mental processes only very loosely linked to performance. The first model is of limited applicability and serves mostly to illustrate possible but unrealized links

between cognition and action. The remaining two are models that might quite reasonably be described as very large and comprehensive. Both have impressive records of empirical application and test.

#### 29.3.1 Recognition Memory Model

A theory with apparently only cognitive aspects appears in a model of recognition memory [29.11] proposed to be dynamical. The model is closely tied to a particular experimental paradigm. For a typical experiment on visual recognition memory, an agent might view either

a familiar picture or a novel one and be instructed to respond familiar or new. The model characterizes the process as a sequence of re-perceptions of the picture or re-looks at it, each look providing more information to update positive and negative accumulators for familiarity. When the positive and negative accumulators for familiarity reach sufficiently small rates of change, a judgment is generated and a response is initiated. Although the model is not presented in this way, the rates of change of the accumulators are analogs of a potential function to be minimized by re-looks. When the potential function has reached a minimum, the cognitive system has reached an attractor, namely a persistent judgment of familiar or unfamiliar. Temporal features of accumulating and looking are hypothesized to affect response time.

In the matter of the relation between cognition and action, some parts of the model are undefined. The process of re-looking, for example, is not explicitly defined. It is not clear whether re-looking is a motor function or mental re-perception without motor involvement. It is not clear whether looking is at the service of the accumulators. It is possible an accumulating module might send out a goal to a looking module, and the looking module then produce some sensations for the accumulating module to work on. Alternatively, it could be that looking and accumulating proceed in mental flux until the process reaches an attractor and initiates a response. The first would be more like a computational model, the second, dynamical.

There are additional questions about the relation of cognitive to motor. The ocular system is not the only motor function in the experimental setup. The agent must also record his judgment with a word or a press of a button. After the process has reached an attractor, presumably a command would be issued to a motor module for this purpose. So this model has the possibility of both an integrated motor process for looking and a separate motor module for executing the response. It illustrates concepts that come into play when analyzing modular models in terms of dynamic and computational features.

### 29.3.2 Adaptive Control of Thought – Rational

Adaptive Control of Thought – Rational (*ACT-R*) is a theory of cognition designed to incorporate what is known as brain function into an architecture which operates to solve problems using symbols and rules of deductive and inferential logic. Because its operations are based on symbols and rules, it can be classified as a computational model of cognition. It has focused on

higher level cognition and problem solving rather than perception or action [29.12].

In any application, ACT-R produces a simulation of the operation of a system consisting of modules of a multifaceted brain, a perceptual motor system, a goal system, a declarative memory, and a procedural system. Each module produces information in a form useful to one or more other modules. This has some dynamical aspects since there is an empirically determined characteristic timing of the operations applied to separate modules. Although it deals with symbols and rules, timing for operations of a module is not necessarily arbitrary nor altogether ignored. This model has complexity of the mental system without the additional complexity of the motor system to which it issues goals [29.12]:

“[. . .] the critical cycle in ACT-R is one in which the buffers hold representations determined by the external world and internal modules, patterns in these buffers are recognized, a production fires, and the buffers are then updated for another cycle. The assumption in ACT-R is that this cycle takes about 50 ms to complete this estimate of 50 ms as the minimum cycle time for cognition has emerged in a number of cognitive architectures [. . .]. Thus, a production rule in ACT-R corresponds to a specification of a cycle from the cortex, to the basal ganglia, and back again. The conditions of the production rule specify a pattern of activity in the buffers that the rule will match, and the action specifies changes to be made to buffers. The architecture assumes a mixture of parallel and serial processing.”

The feature of parallel processing does not imply unlimited capacity since there are two limited-capacity features built into the system. The first is the limitation on buffer contents. A buffer can hold only a single item from memory or perception. The second is a limitation on production rules, only a single one can be selected on each cycle.

ACT-R has been used along with brain imaging to identify certain brain structures with which aspects of cognition can be associated. Brain activity was imaged for participants while they learned a new artificial algebraic system, manipulated its equations, and keyed in answers, in an experiment which lasted over several days. Imaging yielded a measure of activity, the blood oxygenation level-dependent (*BOLD*) function in brain structures over the time course of the experiment and related the measure to theoretical account of steps in problem solution with the following results [29.12]:

- “1. The motor area tracks onset of keying. Otherwise, the form of the BOLD function is not sensitive to cognitive complexity or practice.
2. The parietal area tracks transformations in the imagined equation. The form of the BOLD function is sensitive to cognitive complexity but not practice.
3. The prefrontal area tracks retrieval of algebraic facts. The form of the BOLD function is sensitive to cognitive complexity and decreases with practice.
4. The caudate tracks learning of new procedural skill. The BOLD function is not sensitive to cognitive complexity and disappears with practice.”

These four results illustrate the role of timing and complexity in ACT-R theory. Although the theory can be categorized as computational, it is clear from this example that the theory addresses the dynamics of brain processes which accompany learning procedural skills of varying difficulty. Although time comes in as an index of practice, measured in days, the focus is on the end-effect of practice, to identify which brain modules and transmitter substances might be involved in developing an understanding of the algebra and skill at executing a sequence of steps to solve an algebraic problem. The objective is to match brain function and modules with their theoretical counterparts. That is, the modules and information flow represented in the architecture of the theory are matched with activities in specific brain modules, allowing the function of the brain modules to be inferred and described in terms of the theory. So, too, with complexity, which is here not taken to refer to the complexity of complexity science, but rather to the difficulty of the artificial algebra and the problems given for solution, and is also a variable to be related to the function of brain modules. Thus, ACT-R has timing and complexity, but does not address the process using methods of complexity science or nonlinear dynamic systems theory.

### 29.3.3 Artificial Neural Networks Methods

Artificial neural networks (ANNs) are a large number of theories addressing many phenomena including phenomena of cognition. ANN theories share the fundamental component of the artificial neuron ( $N$ ) an element with some similarity to physical neurons. Ns are usually lined up in layers. Artificial neurons in the lowest layer receive stimulation from an external source (*input*). The remaining layers receive weighted stimulation from other neurons usually from the next lower layer. The weights reflect variations in the strength of the connection to one N from another. The firing pattern of the highest layer (*output*) is readable in meaningful terms by some other system. Intermediate layers are

generally referred to as hidden, and their firing patterns, determined by the weights, are not necessarily easily interpreted.

Each N accumulates the weighted stimulation and transforms it according to a nonlinear function usually acting as a threshold. When the transformed value reaches the threshold, the N fires, usually stimulating at least one N in the next higher level of its next higher neighboring artificial neurons. The word *usually* appears often in the description of ANNs because the structure and operation of an ANN is subject to the ingenuity of its designer. The amount of stimulation received by one N from the firing of another is governed by their connection weight which is taken to be a measure analogous to synaptic strength. From these few properties and their numerous variations, structures can be developed which when in operation simulate thought processes associated with many brain activities thought to underlie learning, concept formation, and rational thought.

For an example of the operation, input might be of a new pattern, expressed in ones and zeros. The weighted and reweighted elements of the pattern are passed through one or more layers of artificial neurons, of each which applies a nonlinear transformation. When it reaches the output layer, the output might be ones and zeros representing a category into which the ANN has determined the new pattern falls.

Whatever the network and task, connection weights which optimize performance must be found. This is usually accomplished by minimizing a cost function during a training procedure in which the ANN learns the optimal connection weights to perform well on a particular task. Considering only supervised pattern learning, any given input pattern can be associated with a desired output, a correct classification for example. With the goal of minimizing the error over all patterns, the weights over the entire network can systematically be adjusted in a direction which reduces the cost function, which will over several trials incrementally bring the cost function to a local or global minimum. Minimizing error is similar to minimizing potential  $V$  of (29.1) and (29.2). A typical cost function is the least-squares minimization given as

$$V = \sum_i (o_i - d_i)^2, \quad (29.3)$$

where  $V$  is the cost,  $i$  indicates the  $i$ th output N,  $o$  is the output value, and  $d$  is the desired value. Minimizing the cost function is a process of reducing errors in classification by adjusting connection weights throughout the network, usually through a technique called *back-propagation*. Unlike the earlier examples, the local and

global minima for  $V$  are not usually completely known and the process of minimizing  $V$  can settle at either type of minima, arriving at a local or global attractor for the set of weights. It is a process of parameter estimation which has some analogy to the characterization of the learning process. Normally the process of parameter fitting is not part of the cognitive model, but in the case of ANNs, as the parameters are being changed trial by trial the process of parameter estimation mimics the learning process, that is, changes synaptic weights. Such changes in synaptic weights are thought to characterize learning and adaptation in biological systems.

One of the strengths of ANN models is that they can often easily incorporate substructures with known properties. ANNs are well suited to do real tasks such as those based on identification of a limited number of patterns. For example two ANNs, each of which is itself a theory of adaptation are the ANN model for Hebbian learning and the Hopfield Network for learning to classify patterns without supervision. These are examples of the power of neural network modules which can be used in sections of an ANN model of cognitive and behavioral functions.

### Hebbian Learning

In Hebbian learning when a neuron  $N$  fires, all of its connections with other  $N$ s in the layer below are affected, in particular, any  $N$  whose firing stimulated it. Each of these connections is increased in effectiveness, that is, will transmit a greater effect in the future. Hebbian learning is a dynamical model of neural modification during learning. In a Hebbian, ANN parts of the system during their normal operation create a useful learned configuration. It is a dynamical model of adaptive brain changes, which instantiates the Hebbian theory of neural correlates of learning.

### Hopfield Network

An ANN can simulate other cognitive functions, for example, the Hopfield network can infer good patterns from samples without any supervision, that is, without feedback on correctness of its performance while learning to classify patterns during training. The Hopfield Network can also remove noise from imperfect patterns, and is known as a technique for self-addressed memory.

## 29.3.4 Adaptive Resonance Theory

Adaptive Resonance Theory (*ART*) is a global theory of cognition using related models of brain and neuronal properties to assemble ANNs to simulate cognitive and other brain functions. Adaptive resonance is analogous to energy produced by a pattern match based on

processes of stored patterns (*top-down*) and of input management (*bottom-up*). When a match is achieved the system sends the information to the next module, by ANN routes, for further processing. ART has by these means addressed and simulated the theoretical processes that underlie numerous results from the literature of experimental psychology and neuroscience.

The fundamental unit of ART incorporates the dynamics of adaptive resonance and forms the common basis for many related models. Models have been formulated with different objectives but all with *ART* in the acronym, indicating that adaptive resonance is a fundamental feature of every model so derived. Adaptive resonance is the fundamental theoretical process of cognition.

Adaptive resonance entails a temporal process that unfolds during the creation of a meaningful perception from an input. The input may be a pattern to be identified or categorized, or it may be part of a temporal sequence such as speech to be comprehended. Having received input, ART initiates a comparison process is with the objective of maximizing a resonance function. In the comparison process, the bottom-up input signal interacts with top-down previously learned patterns, prototypes, and expectations. The interaction consists of repeated cycles of directing attention to combinations of bottom-up features deemed significant by a matching process and suppressing bottom-up features deemed irrelevant. In repeated cycles, the resonance function is optimized and the information can be passed on to the artificial neurons in the next module for further processing.

ART is currently silent when it comes to generating overt responses, although some indication has been given of a proposed approach to the problem. For ART, the perceptual system derives resonance from achieving a match. For the motor response, the system issues a goal and leaves it at that. The suggestion has been put forward that the motor response might be shaped by a *complementary* process, a sort of mirror image of resonance. In the proposed process, the complementary energy analog of resonance would be generated not by a match, but by a mismatch between actual and desired action, the mismatch indicating the desired goal has not yet been reached. Motor functions are addressed and characterized thus [29.13]:

“The START model proposes how adaptively timed inhibition of the hippocampal orienting system [...] and adaptively timed disinhibition of cerebellar nuclear cells [...] may be coordinated to enable motivated attention to be maintained on a goal while adaptively timed responses are released to obtain a valued goal. [...] Biological learning in-

cludes both perceptual/cognitive and spatial/motor processes. Accumulating experimental and theoretical evidence show that perceptual/cognitive and spatial/motor processes both need predictive mechanisms to control learning. Thus there is an intimate connection between learning and predictive dynamics in the brain. However, neural models of these processes have proposed, and many experiments have supported, the hypothesis that perceptual/cognitive and spatial/motor processes use different types of predictive mechanisms to regulate the learning that they carry out. [...] The need for different predictive mechanisms is clarified by accumulating theoretical and empirical evidence that brain specialization is governed by computationally complementary cortical processing streams that embody different predictive and learning mechanism.”

### 29.3.5 Summary

Both ACT-R and ART characterize dynamics of mental processes theoretically observable through brain imaging. Both models characterize cognition and performance as separate processes. Cognition may be

required for a performance, but cognition and performance are two different processes. Behavior is taken as a window on mental processes, that is, at the completion of the cognitive process, a goal may be issued to a motor system. Behavior, then, indicates what goal was set. Separating cognition and action is, however, not a necessary feature of these types of models, an integration of the two is an explicit goal for both ACT-R and ART.

With regard to complexity, both ACT-R and ART have interrelated communicating subsystems, modules which act in cooperation, rather than in concert, that is, usually sequentially, not simultaneously. As such, they do not invite the techniques of complexity science. The fact that ACT-R specifically addresses symbols and rules of their manipulation does not prohibit application of nonlinear dynamic systems analysis, as there has been developed a method of *symbolic dynamics* [29.14].

Not time, but timing is an essential feature of both models. That is, what is known from brain imaging about the active areas of the brain and about the times required for particular brain activities is incorporated into both models. These are reflected in temporal restrictions on sequencing and when passing information is permissible between modules.

## 29.4 Cognition and Action Intrinsically Linked

### 29.4.1 Methods

In discussions of theories for which cognition and action are intrinsically linked, it will be useful to reserve the word *model* to refer to properties of the agent. The word is commonly used in two senses, each an example of some kind of entity for prediction of events, evaluation of evidence, and revision of beliefs. The first type is the scientific model using various techniques to generate predictions, guide the formulation of experiments, and prescribe routines for evaluating results. These are the models to which the title of this chapter refers. For clarity and brevity in the presentation of models which intrinsically link cognition and action such scientific models will be referred as *theories*. The second usage of the word includes models which instantiate a set of beliefs held by the agent about the state of the external world based on patterns of sensations. They include beliefs in models of processes in the external world, such as the belief that a convivial friend will enjoy the party. They also include beliefs held by the agent about his actions and the consequences thereof in the external world, such as the belief that opening the cupboard will reveal a tasty snack. Beliefs held by the agent will be called models in the remainder of this chapter. Thus,

theories may contain hypotheses about the existence and properties of models. This is especially so for theories which assert cognition and action to be intrinsically linked. Such theories are usually associated with methods by which to characterize the processes by which the agent’s beliefs are created and altered. Two such methods will be described next, Bayesian multiprocess models and particle filters.

#### Bayesian Multiprocess Models

According to the theory of multiprocess models of cognition and behavior [29.15], the agent generates expectations for what will happen next from each of numerous mental models pertaining to the experience. These expectations take the form of a prediction for the next event and assignment of a probability from each model to each possible outcome of the event or, equivalently, to each possible error of prediction. The models themselves may be of any sort, that is, they may contain features of complexity, nonlinearity, and feedback, or they may be simple stochastic models. The word multiprocess indicates that each of several models may be characterized as a belief about the dynamics of some event. The relative amount of belief in each model is expressed as its probability. The models are required to

be mutually exclusive and jointly exhaustive as far as beliefs are concerned.

The important aspect of mental models is they make a specific probabilistic prediction for the very near future, that is, they attach a probability to every event which might possibly happen next. After an event occurs, each model is re-evaluated according to the support each receives from the evidence of the event. For example, consider an agent who has two models for a coin, that it is either fair or biased. The fair coin model,  $z_{50}$ , is expressed as  $P(H) = 0.5$  and  $P(T) = 0.5$ . The biased coin model,  $z_{80}$ ,  $P(H) = 0.8$  and  $P(T) = 0.2$ . Suppose before the coin is tossed he believes 60-40 that the coin is fair, that is, his *prior* amount of belief in each model is expressed in the probabilities  $P(z_{50}) = 0.6$  and  $P(z_{80}) = 0.4$ . The two models and his relative degree of belief in each form the context for his interpretation of the subsequent event, which will be the outcome of a toss of the coin. For each model the agent generates a prediction and a probability for any other outcome, for  $z_{50}$  he might predict heads  $H$  and note a 0.5 probability of error, namely  $T$ . For  $z_{80}$  he will predict  $H$ , but with a 0.2 probability of error. The coin is tossed and comes up  $H$ . The question is how does this outcome affect his degree of belief in each model to yield revised or *posterior* probabilities? The answer is given by Bayes rule, in the odds form, applied to each model separately in (29.4)

$$\frac{P(z_{50} | x_H)}{P(z_{80} | x_H)} = \frac{P(x_H | z_{50}) P(z_{50})}{P(x_H | z_{80}) P(z_{80})}, \quad (29.4)$$

where  $x_H$  is the event of having  $H$  occur after predicting  $H$ , that is, having predicted  $H$  accurately. Although prediction from each model is accurate, that is, prediction error is zero for each, the probability of a zero error is different for each model. Evaluating (29.4) along with the condition that the models are jointly exhaustive, yields the effect of the event on the belief in each of the two models, posterior beliefs of  $P(z_{50} | x_H) = 0.48$  and  $P(z_{80} | x_H) = 0.52$ . Given the same prediction  $H$ , if the outcome of the coin toss had been  $T$ , the result would have been different with  $P(x_T | z_{50}) = 0.5$  and  $P(x_T | z_{80}) = 0.2$ . Then the result would have been the posterior beliefs of  $P(z_{50} | x_T) = 0.79$  and  $P(z_{80} | x_T) = 0.21$ .

In the theory of multiprocess models of cognition, the models invoked by the agent arise from both internal sources and experience. The set of models together with their degrees of belief form the context by which the agent understands events. In any situation, the prediction error is used to revise the degree of belief in each model. The effect of repeated application of Bayes rule is to alter beliefs to reduce errors of prediction.

In the dynamics of multiprocess models, the prediction error serves the same purpose  $V$  does for complex systems, that is, the system of beliefs forming the context of the experience, evolves in such a way always to reduce errors of prediction by systematically bringing the successive priors closer to their respective posteriors. In this way, ongoing experience produces and refines the context of the experience.

### Particle Filters

Particle filters provide a method for applying Bayes rule in an environment of constant flux where the agent has both beliefs about his actions as they affect his environment and beliefs derived from evidence about the state of the environment. By incorporating beliefs about the effects of action, new models may be introduced to the set, providing a flux of models to accommodate the flux of the environment.

Particle filters give a best-guess approximate solution to the problem of finding prior beliefs about otherwise unknown states of an agent moving in a changing world. The method uses such prior beliefs, a *motion model*, input to a *map model*, and current stimulation to generate updated beliefs about current state [29.16]. Its application can be illustrated with a simple example of an agent moving in the dark around the furniture in a familiar room.

The *particles* in question are simple hypotheses, such as a statement about location. One particle might claim *You are here at A*. another *You are here at B*. and so forth. Having started into the room from the threshold and taken three steps, he consults his motion model and believes he has arrived at one of three points,  $A, B, C$  with respective probabilities 0.2, 0.3, 0.5. To further delineate his location, he reaches out and finds that he touches a table, providing evidence  $x$ . He then consults his mental model of a map of the familiar room and determines the probability of touching the table from  $A, B, C$ , is respectively, 0.7, 0.5, 0.2. These yield posterior probabilities for  $A, B, C$ , given  $x$ , of 0.36, 0.38, 0.26. At this point, the filter reweights each particle according to the degree to which touching the table was a surprise for it. The weight for each particle is calculated as the ratio of its posterior according to the map to its prior according to the motion model. The particles are then re-sampled with replacement according to the weights to yield a new probability density function (*pdf*) for final belief of location. The posteriors for  $A, B, C$ , respectively are 0.50, 0.36, 0.14. He can use these posteriors as priors in the motion model, for his next steps, then once again reach out, then consult his map about the result. At this time it is quite likely a new particle  $D$  might come into the picture, justifying the re-weighting and

re-sampling of the posteriors for his next input to the motion model.

The particle filter is a method for continuously bootstrapping probabilities which cannot simply be dragged from one situation to another because the question and the environment are continually changing. The question is not simply *Where am I?*, but rather *Now that I have taken three steps from one of three places I was at with varying probabilities, where am I?* It gives weight to the previous beliefs from models, but sheds them when new circumstances and other models come into play.

Although it is not stated explicitly in terms of prediction errors, by the procedure just described, particles with smaller prediction error become more likely to be sampled to estimate the pdf, while particles with greater errors tend to drop out. This keeps neighboring particles more influential and distant particles less. The dynamical feature of this model is that cognition changes always in a way to reduce a potential function, in this case, the quantity minimized is surprise, the discrepancy between the prior and posterior beliefs.

### 29.4.2 Embodied Cognition

Theories of embodied cognition hold that phenomena of cognition are all, or in large part, emergent from coordination of components of a complex motor system. The psychology of infant development has been a source of examples of the theory, taking advantage of the fact that infants change over time in correlated cognitive and motor abilities. The theory overview is that over a period of growth, motor subsystems settle into new patterns of coordination in a manner analogous to coupled oscillators, a feature of increasing motor size, strength, and complexity. The patterns give the appearance of new cognitive abilities but the theory holds in certain cases the cognitive property exists only in new patterns of the motor coordination. Thus eye movements and limb movements may become coordinated in a pattern which gives the appearance of a new level of belief about properties of objects, but that emergent property is not independent of motor coordination. The embodiment theory of cognitive development relies on concepts from both complexity science and nonlinear dynamic systems theory [29.17], Chap. 30.

An illustration of the dynamical approach to cognitive development concerns a prototypical phenomenon known as the *A-not-B* error. To demonstrate the developmental difference in *A-not-B* the infant is shown a toy, which is then hidden at location *A* within easy reach. The infant reaches to uncover and retrieve the toy at *A*. This is repeated several times, then the toy is shown being hidden at location *B* also within easy reach. After a short delay, the infant is allowed to reach

for the toy. If the infant is 10 months old he will reach for *A* not *B*, apparently making an error of cognition in the sense that the infant appears to believe that the toy hidden at *B* nevertheless will still appear at *A*. At 12 months, the infant will correctly reach to *B*, appearing to have reached a new concept that toys put someplace will not move from there on their own. This consistently reproducible error had previously been interpreted as a sign that between 10 and 12 months, the infant develops an understanding of object permanence, that is, develops a belief that the object will not magically jump from *B* to *A*. The dynamical model takes issue with this interpretation proposing the same intellectual factors enter into the two types of responses and the same process but the process is complex, made up of two motor subprocesses with different timing at different ages.

The first process is a motor memory of reaching for *A*, made strong by the initial repetitions. The second is memory for the looking where the toy was recently hid, *B*. For a younger infant, the memory of looking at *B* is hypothesized to decay faster and by the end of the delay the memory of reaching for *A* comes to dominate. For older, a different time course of decay for the two memories leaves the memory of looking dominant at the end of the delay. This dynamical interpretation has been successful in several tests using variations of conditions intended to differently affect the time-courses of decay of the two memories. In this way, the relation between the time-courses of decay for two memories, each of the result of its respective motor process, gives rise to an apparent cognitive advance, object permanence. The cognitive advance is thereby an emergent property of the system, created from more elementary subsystems. Instead of the motor function obeying the cognitive command, the cognition arises as an epiphenomenon of motor coordination.

Outside of developmental psychology, the theory of embodied cognition has been applied to mental imagery. The theory of embodied mental imagery holds that even when there is no overt behavior representing the cognition, it nevertheless is body based. Abstractions such as mental imagery, working memory, episodic memory, implicit memory, reasoning, and problem solving which operate in the absence of overt behavior may be called *off-line* cognition. The process from which off-line cognition derives, is linked to decoupling of mental processes from overt behavior, a process by which mental processes which formerly accompanied overt behavior have decoupled from the behavioral aspect and go forward on their own [29.18].

As a matter of principle, the off-line theory of embodied cognition holds that decoupled, abstract cognitive activities are remnants of bodily activities. By this

reasoning, mental processes which originally only accompanied veridical perception and action, have been adapted to operate off-line, without requiring input or output, but nevertheless retain features characteristic of perception and action [29.18]:

“Off-line aspects of embodied cognition, in contrast, include any cognitive activities in which sensory and motor resources are brought to bear on mental tasks whose referents are distant in time and space or are altogether imaginary. These include symbolic off-loading, where external resources are used to assist in the mental representation and manipulation of things that are not present, as well as purely internal uses of sensori-motor representations, in the form of mental simulations. In these cases, rather than the mind operating to serve the body, we find the body (or its control systems) serving the mind.”

Almost in contrast to the off-line view is a theory of embodiment which holds that the mind extends into the physical world. For example, this theory asserts that mental activity includes actions such as using paper and pencil to solve an arithmetic problem or write a sentence. The rules of manipulation of symbols, and the symbols themselves are extensions of the mind into the physical world. In this way, the theory asserts the mind increases its capacity for memory and rule application [29.18].

### 29.4.3 Motor Theory

Motor theories of perception link action to cognition. Understanding or perceiving an observed action by others is accomplished by the covert production of that action. The early example of such a theory is the motor theory of speech perception. The motor theory holds that speech is perceived through the listener’s covert production of the same speech [29.19]:

“As for speech perception, there is now evidence that perceiving speech involves neural activity of the motor system. Two recent studies involving the use of transcranial magnetic stimulation of the motor cortex have demonstrated activation of speech-related muscles during the perception of speech.”

Concepts in the motor theory of speech perception have been extended to a more general theory, applying to any observed action by others. The general motor theory of perception has received much support from studies of neuronal activation and brain imaging which shows perceiving actions involve the same neuronal, brain, and motor system as producing the action. Complexity in production, complexity in perception, and

complexity in coupling the two are central features of motor theories of perception.

### 29.4.4 Simulation Theory

Simulation theory [29.20] holds, in the large, that mental life, particularly imagination, consists of processes by which the brain synthesizes sensations and perceptions which do not arise from the external world and actions which do not affect the external world. To achieve this, the brain activates the same pathways used in veridical sensation, perception, and action, but interrupts their contact with the external world. In this way, a replica, or simulation of external events and processes can be experienced without requiring the presence of their veridical counterparts. Simulation can be either of a stable perception, or of a dynamic unfolding of a process of simulated perception, action, and anticipation of the consequences of action.

Simulation theory requires that the brain contain structures which can accommodate perception, action, and anticipation in the absence of external input, as contrasted with veridical perception and action. The resulting motor stimulation is stopped short of execution, resulting in simulated behavior [29.20]:

“Saying that behaviour can be simulated here means nothing more than that the signal flow from the prefrontal cortex via the premotor areas may occur even if it is interrupted before it activates the primary motor cortex and results in overt behaviour. A simulated action is thus essentially a suppressed or unfinished action.”

Simulation theory explains imagination and anticipation, both arising from sensations in the absence of external input. It is characterized as a variety of perception involving the same parts of the brain which occur with veridical perception. In addition to being stimulated by input from the external world and transmitting the stimulation to higher parts of the brain, the parts of the brain that produce sensation can also be stimulated by a retrograde flow from higher parts to sensation. Such sensations may be distinguished from veridical sensations by the fact that they occur in an episode during which overt action is being suppressed.

Anticipation is characterized as the imagined consequences of suppressed action. From experience, the agent learns to anticipate consequences of action and can fold this information into the simulated experience. Anticipation thus consists of models of the effects of actions. It returns the effects of the imagined actions as imagined sensations. In accordance with simulation theory, after some amount of veridical experience, perception, action, and consequence, all



may occur without concurrent contact with the external world.

### 29.4.5 Free Energy Theory

Free Energy Theory connects beliefs with actions as two sides of the same coin. The theory is presented with rigor, definitions, and distinctions in [29.21]. The theory addresses cognition and action with a notational system which accommodates interfaces between the external world and the sensations it produces for both cognition and action.

On the cognition side, it distinguishes between brain states and *causes*. Brain states are analogous to the models of Bayesian multiprocess models, while causes are part of the external world. *The recognition density is a probabilistic representation of what caused a particular sensation* [29.21, p. 128]. It is the aim of minimization of free energy to have the brain states match the causes, thereby reducing occasions of surprise. On the action side, the theory offers a particular kind of model, labeled *m*, which links an action to its consequences, changes in the external world.

It is the aim of action to alter sensations to more closely match beliefs associated with brain states, which it can accomplish most effectively by turning up evidence confirming the most likely model. Thus cognition and action are each part of comprehending the sensations which arise from experience, the former by identifying its causes, the latter by seeking confirmation of the most likely causes by manipulating the environment. Both cognition and action are governed by the same principle, the minimization of free energy.

The purpose of the following illustration is to explain how causes and models can be connected to free energy. Certain conditionals have been dropped in the interest of clarity and brevity. Thus, in (29.5)–(29.7) the symbol *z* should be understood as a *belief about the external causes of sensation given the brain state*. There can be many *z*s, each with some initial degree of belief, expressed as  $Q(z)$ . The quantity *x* is used here to represent external states or events given the sensations, the causes, and the model of action. These representations of *z* and *x* are similar but not identical to the models of Bayesian multiprocess theory and simulation theory. In particular, free energy theory explicitly expresses *x* as the result of a filtering process, not the unobservable external process itself ( $\varphi$ ), but an estimate of it, instantiated in neural activity, in a time series of vectors of sensations.

In the simplified example, assume the agent observes an event, *x*, for which he has prior probabilities,  $Q(z)$ , for each *z*, yielding an approximation  $Q$  to the optimum posterior probability  $P(z|x)$ . The goal is to

come to a conclusion in which the discrepancy between  $Q$  and  $P$  is minimal. This can be accomplished by minimizing a measure of the discrepancy, the Kullback–Liebler (*KL*) divergence  $D(Q||P)$ , expressed in (29.5)

$$D(Q||P) = \sum_z Q(z) \ln \frac{Q(z)}{P(z|x)}, \quad (29.5)$$

which expands to

$$D(Q||P) = \sum_z Q(z) \ln \frac{Q(z)}{P(z,x)} + \ln(p(x)). \quad (29.6)$$

On the right-hand side of (29.6) the second term is the negative of the information content of the event *x*. The first term is referred to as free energy. Rearranging the terms of (29.6) shows free energy  $F$  is equal to the KL divergence plus the information in *x* represented in (29.7) which suggests free energy may be minimized by changing the discrepancy between  $Q$  and  $P$  or by changing the event *x*

$$F = D(Q||P) - \ln(p(x)). \quad (29.7)$$

When the conditionals are included, Free Energy Theory expresses the  $F$  of (29.7) as  $F(s, \mu)$ , where *s* stands for sensory states and  $\mu$  stands for brain states or, equivalently, neural processes representing links between *s* and its causes  $\varphi$ . In addition, the set of models, *m*, link actions to changes in *s* by their effect of altering the external world. There are two ways to minimize  $F$ , both central to free energy theory, the first by changing brain states  $\mu$ , the second by changing sensations *s* through action on the environment. The role of changing brain states is one of coming to some degree of belief in new ideas of causal relationships between the external world and the accompanying sensations. In the case of action, free energy can be reduced most effectively by a particular sort of action, namely action which changes the environment to uncover evidence (event *y*) favoring the most likely cause of *s* for a given brain state. The models *m* determine which action *a* will likely uncover event *y* which has the minimizing effect,  $F(z, y) < F(z, x)$ . Free Energy Theory predicts the organism will act in a way to uncover evidence consistent with the most likely cause, because doing so most effectively minimizes  $F$ .

A model in *m* might be of any variety, for example, a simulation or a simple probability. The only requirement is it link actions to their predicted consequences. In this way, the action is conceived, not as carrying out orders to obtain certain goals, but rather as being a feature of the way the brain interacts with the physical

world to increase predictive accuracy and, equivalently, lessen surprise.

As an example of the role of action, suppose an actor cannot read a word written in poor handwriting. Assume the uncertainty is high because, considering only the word in isolation, there are several equally likely possibilities. The actor can take action to reduce the prediction error by altering the input, namely reading what he can of the words in the surrounding context. The context can alter the respective probabilities of the possibilities for the problem word. That is, action changes the sensory input. When reading the context, the action is quite directed, via the action model  $m$ . The actor is not flipping pages or scanning the room, he is acting in a way to confirm his most likely hypothesis, that this word is part of a meaningful sentence. His strongest hypothesis guides his action. Alternatively, he could instead, or in addition, have tried an approach not involving changing the environment. For example, he could have made guesses about handwriting quirks of the writer. This would introduce new mental models for the letters, thereby changing perception via changing brain state  $\mu$ . Either of these approaches can be called an effort to reduce free energy.

Free Energy Theory characterizes mental activity as a complex process, guided by optimization. In addition, free energy theory has been asserted to apply to mental activity both large and small scale. The examples apply it to mental dynamics on the scale of human activity and perception, such as finding one's way in the dark. In extension, the same concepts and formulas apply it to brain activity on a neuronal scale, as revealed by brain imaging and single-neuron techniques. The methods of nonlinear dynamic systems analysis provide a formalization of such scale invariance.

### 29.4.6 Evolution of Cognitive Search

The evolutionary theory of cognitive search proposes that processes of attention and memory include search strategies analogous to behavioral strategies used in the search of physical space for objects of value. On an evolutionary scale, cognitive search has arisen di-

rectly from physical search [29.22] from behavioral to cognitive, physical to mental. According to the theory, analogous processes occur in cognitive and physical search. For example, both have strategies of remaining in one area (*patch*) until its resources have been fully exploited or otherwise depleted, only then exploring for a new patch to be exploited. Both also are associated with the same areas of brain activities and transmitter substances. Thus, the evolutionary theory of cognitive search asserts important aspects of the relation between cognition and behavior have been developed over a long time scale. The relation between cognition and action is not one of the behavior creating the illusion of cognitive entities, but rather brain functions which govern behavioral search acting as a paradigm for brain activity governing cognitive search.

### 29.4.7 Summary

Where cognition and action are intrinsically linked, theories range from denying the existence a category of cognition separate from the complex coordination of behavioral subsystems, to a complex system of cognition and action combined, to a brain system created by analogy to a complex behavioral system. Complexity is a feature of theories that link cognition and action, almost by definition, since there are applications of complexity science to motor processes even when they are not linked to cognition. The time associated with coordination of subsystems and unfolding of processes, also enters requiring concepts of nonlinear dynamic systems theory.

For free energy theory, the goal does not direct the search, but the search policy inevitably uncovers things of value. For the theory of evolution of cognitive search, behavior and cognition are connected, but in a different way. Cognitive search, in particular, is seen as internalized replica of behavioral search. There is no single way in which all theories of linkage claim cognitive and behavioral functions are linked, but there is great agreement among them that the linkage is there. The dynamical features of these models follow directly from that link.

## 29.5 Conclusion

Unequivocally, theories which combine mental and motor into a unified process are dynamical theories of cognition. Elements of the integrated process are clearly quantified by measures of complexity, real-time coordination, and time series data, which invites application of methods of complexity science and nonlinear dynamic systems theory, the methodology of dynamical models.

From the sampling of cognitive models presented here, incorporating action appears to be a sufficient, but not a necessary condition for a cognitive model to be called dynamical.

Application of a dynamical model to a data set does not necessarily require a comprehensive theory. This is evident in the examples of data-based studies.

Breadth of theory is not a necessary starting point. What is required is instead a commitment to the ideas can be generated by a coherent application of dynamical methodology.

One of the central features of dynamical methodology is optimization of an expression of potential. The system proceeds in a way that minimizes some expression of potential, such as formulated by (29.1), (29.2), (29.3), or (29.7). Such optimization can reveal how it is that the system settles into an attractor and how it moves from one attractor to another. Minimizing potential gives direction and predictability to complex processes. Identification and formalization of a potential  $V$  of a dynamical system paves the way for understanding its stable states and transients.

Dynamical models may emphasize the role of optimization both as a method of determining important quantities in the theory and as a fundamental principle of cognitive processes. This is evident in Free Energy Theory, in which optimization is the organizing principle of thought and behavior. According to free energy theory, the optimizing process occurs in a scale-invariant way at every level from brain structures to cognition–action complex. At every level, optimization in the form of minimization of free energy brings the organism to a state of decreased surprise.

Optimization of a potential function does not have the same overarching role in theories for which cognition and action appear as separate systems. In ART, maximizing adaptive resonance is an optimizing principle at one stage of processing, designed to direct top-down and bottom-up procedures to identify the input pattern or create a new pattern in a way that is useful for further processing. Adaptive resonance and the changes that take place during training of an ART network are examples of dynamic aspects of the ART. These features are usually embedded in a network of modules to which the methodology associated with complexity is not explicitly applied.

Both theories, ART and ACT-R, address the interaction of the symbols, patterns, and rules of cognition and the relationships of these to brain function. Major objectives of ART and ACT-R are to explain how symbols and rules are manipulated and combined in the operations of pattern recognition, classification, and inference, both deductive and inductive. These operations are of central interest to both ART and ACT-R, especially along with their correlation with brain areas

and functions as revealed primarily through brain imaging. When ART or ACT-R is analyzed in contrast to dynamical theories, it is symbols and their rules of manipulation that underlie the contrast.

Symbols and rules of their manipulation are the durable, stable products of mental activity. According to dynamical models of cognition, an alphabet, for example, is the product of dynamical forces of cognition and behavior. One theory of embodiment claims that symbols are physical extensions of mind. They exist as long as general usage and culture keeps them far from equilibrium, the increased entropy of which would turn them to dust. Thus, symbols and rules have the features of attractors in a dynamical system. This is emphasized by the data-analytic method of symbolic dynamics. From this point of view, computational models address relationships among attractors, while dynamical models address the process that brings attractors into existence and govern the transients among them.

Philosophical questions arise from the dynamical point of view concerning the status of the contents and products of mind. What is the philosophical status of an internal model such as those of simulation theory or free energy theory? How are they the same as or different from scientific theories? If brains have brain-state models, and action models, do they not also have symbols? How is a model or a brain state different from a symbol? What is the status of the durable external products of mental activity; alphabets, numerals, rules of inductive and deductive logic, art and engineering, novels and history books? Are these extensions of mind?

The benefits of a dynamical point of view to modeling cognition, in addition to its intuitive appeal, are the wealth of finely developed methods, some of which have been described here. Complexity science introduces new methods which make it possible to address questions which have previously been inaccessible for formalizing and quantifying mental processes. Non-linear dynamic systems theory introduces systematic methods for characterizing the phenomena of nonlinear recursive processes using new concepts, such as formulation of trajectories and their features, including attractors and measures of predictability or dimension of the process. To preserve and further develop benefits such as these, the associated methodologies would be a valuable addition to the curriculum of psychology for the study of the mechanics and properties of mind.

## References

- 29.1 T. van Gelder: The dynamical hypothesis in cognitive science, *Behav. Brain Sci.* **21**, 615–665 (1998)
- 29.2 T. van Gelder, R. Port: It's about time. In: *Mind as Motion: Dynamics, Behavior, and Cognition*, ed. by R. Port, T. van Gelder (MIT Press, Cambridge 1995) pp. 1–44
- 29.3 R.F. Port, A.P. Leary: Against formal phonology, *Language* **81**(4), 927–964 (2005)
- 29.4 C. Zednik: The nature of dynamical explanation, *Philos. Sci.* **78**(2), 238–263 (2011)
- 29.5 W.J. Freeman: Nonlinear brain dynamics and intention according to aquinas, *Mind Matter* **6**(2), 207–234 (2008)
- 29.6 J.A.S. Kelso: *Dynamic Patterns: The Self-Organization of Brain and Behavior* (MIT Press, Cambridge 1995)
- 29.7 S.J. Guastello, M. Koopmans, D. Pincus (Eds.): *Chaos and Complexity in Psychology: Theory of Nonlinear Dynamical Systems* (Cambridge Univ. Press, New York 2009)
- 29.8 S.J. Guastello, R.A.M. Gregson (Eds.): *Nonlinear Dynamical Systems Analysis for the Behavioral Sciences Using Real Data* (CRC, Boca Raton 2011)
- 29.9 H. Haken, J.A.S. Kelso, H. Bunz: A theoretical model of phase transitions in human hand movements, *Biol. Cybern.* **51**, 347–356 (1985)
- 29.10 T.A. Oliva, S.R. McDade: Catastrophe model for the prospect-utility theory question, *Nonlinear Dyn, Psychol. Life Sci.* **12**, 261–280 (2008)
- 29.11 G.E. Cox, R.M. Shiffrin: Criterion setting and the dynamics of recognition memory, *Top. Cogn. Sci.* **4**, 135–150 (2012)
- 29.12 J.R. Anderson, D. Bothell, M.D. Byrne, S. Douglass, C. Lebiere, Y. Qin: An integrated theory of the mind, *Psychol. Rev.* **111**(4), 1036–1060 (2004)
- 29.13 S. Grossberg: Adaptive resonance theory: How a brain learns to consciously attend, learn, and recognize a changing world, *Neural Netw.* **37**, 1–47 (2013)
- 29.14 R.A.M. Gregson, S.J. Guastello: Introduction to nonlinear systems analysis. In: *Nonlinear Dynamical Systems Analysis for the Behavioral Sciences Using Real Data*, ed. by S.J. Guastello, R.A.M. Gregson (CRC, Boca Raton 2011) pp. 3–15
- 29.15 M.A. Metzger: Multiprocess models of cognitive and behavioral dynamics. In: *Mind as Motion: Dynamics, Behavior, and Cognition*, ed. by R. Port, T. van Gelder (MIT Press, Cambridge 1994) pp. 491–526
- 29.16 K. Hsiao, H. de Plinval-Salgues, J. Miller: *Particle Filters and Their Applications* (2005)
- 29.17 L.B. Smith, E. Thelen: Development as a dynamic system, *Trends Cogn. Sci.* **7**(8), 343–348 (2003)
- 29.18 M. Wilson: Six views of embodied cognition, *Psychon. Bull. Rev.* **9**(4), 625–636 (2002)
- 29.19 B. Galantucci, C.A. Fowler, M.T. Turvey: The motor theory of speech perception reviewed, *Psychon. Bull. Rev.* **13**(3), 361–377 (2006)
- 29.20 G. Hesselow: The current state of simulation theory of cognition, *Brain Res.: The Cogn. Neurosci, Thought* **1428**, 71–79 (2012)
- 29.21 K. Friston: The free-energy principle: A unified brain theory?, *Nat. Rev. Neurosci.* **11**, 127–138 (2010)
- 29.22 T.D. Hills, R. Dukas: The evolution of cognitive search. In: *Cognitive Search: Evolution, Algorithms, and the Brain*, ed. by P. Todd, T. Hills, T. Robbins (MIT Press, Cambridge 2012) pp. 11–24

## 30. Complex versus Complicated Models of Cognition

Ruud J.R. Den Hartigh, Ralf F.A. Cox, Paul L.C. Van Geert

As humans, we continuously adapt our behavior to changes in our environment, and our cognitive abilities continuously develop over time. A major question for scientists has been to discover the (cognitive) mechanism that underlies the control of human behavior in real time, as well as cognitive development in the long term. This chapter will discuss two kinds of general approaches, which we shall refer to as the reductionist approach and the complex dynamic systems (CDS) approach. Roughly speaking, the reductionist approach assumes that separate cognitive components, such as brain areas or processing mechanisms, are primarily responsible for behavior and development, by processing (and responding to) specific environmental cues. The CDS approach assumes that cognition and thereby the control of behavior and development are distributed over the brain, body, and environment, which continuously interact over time. The aim of this chapter is to compare the two approaches in terms of their assumptions, research strategies, and analyses. Furthermore, we will discuss the extent to which current research data in the cognitive domain can be explained by the two different approaches. Based on this review, we conclude that the CDS approach, which assumes a *complex* rather than a *complicated* model of cognition, provides the most plausible approach to cognition.

30.1	<b>Current Views on Cognition</b> .....	658
30.1.1	Central Control versus Self-Organization .....	658
30.1.2	Static versus Dynamic Models .....	659
30.2	<b>Explaining Cognition</b> .....	660
30.2.1	Research Strategies and Complicated Models.....	660
30.2.2	Research Strategies and Complex Models.....	660
30.2.3	Analyses to Untangle Cognition Based on Complicated Models.....	661
30.2.4	Analyses to Capture the Complexity of Cognition .....	661
30.3	<b>Is Cognition Best Explained by a Complicated or Complex Model?</b> .	662
30.3.1	Explaining Real-Time Cognitive Performance .....	662
30.3.2	Explaining Long-Term Cognitive Development .....	663
30.4	<b>Conclusion</b> .....	666
	<b>References</b> .....	666

At present, two classes of approaches are used to explain cognition. The first class proceeds from the idea that human behavior is controlled by separate cognitive (processing) components, which we refer to in this chapter as the reductionist approach. The second class assumes that cognition can be considered as a complex, dynamic set of components, and that human behavior is an emergent consequence. We shall refer to the latter as the complex dynamic systems (CDS). The first part of this chapter starts with an overview of the two approaches and the explicit and implicit as-

sumptions they make (Sect. 30.1). In Sect. 30.2, we discuss the research strategies and analyses applied by researchers proceeding from a reductionist or CDS approach. Then, in Sect. 30.3 we demonstrate the extent to which complicated (related to the reductionist approach) and complex models (related to the CDS approach) fit with research data on real-time cognitive performance and long-term cognitive development. Finally, in the concluding section, we discuss which kind of model seems to explain human cognition best (Sect. 30.4).

## 30.1 Current Views on Cognition

On 30 September, 2014, President Obama announced that the White House would make new investments in the BRAIN initiative. President Obama explained that:

“As humans we can identify galaxies light years away, we can study particles smaller than an atom, but we still haven’t unlocked the mystery of the three pounds of matter that sits between our ears.”

Not only researchers, philosophers, but also world leaders and their scientific advisors are fascinated by the question why humans behave the way they do, and more specifically *what controls* human behavior. Human adults appear to have the most evolved prefrontal cortex, neocortex, and temporal lobes of all creatures. Explicitly mentioned or implicitly assumed, for many scholars the mind – or its physical instantiation, the brain – is a key to explain human’s superior cognitive and behavioral capacities. It is generally believed that our abilities for language, abstract reasoning, problem solving, learning and memory, interacting with other people, and using tools ultimately reside in our brain. The prevailing notion today, and throughout a large part of the modern history of psychology, is that fairly localized structures (or *modules*) in the brain play specialized and identifiable roles in how we perceive, how we act, what and how we (can) learn, and even in our emotions and personality [30.1]. Cognition, as amalgam of many such distinct *cognitive* functions and subfunctions, is a mechanistic apparatus consisting of specialized modules linked together in a linear causal chain. This premise has directed the research attention to localizing these *modules* or components and the functions they perform. We shall refer to this approach, in which cognition and the explanation of human behavior is reduced to localized functions, as the reductionist approach.

Obviously, the environment also contributes to how we behave and learn. In the reductionist view, the role of the environment is rather dissociated, that is, it provides input to cognitive processes. More specifically, environmental cues are cognitively processed, after which the *best* subsequent action can be computed, and the situational input can be cognitively stored in order to respond optimally the next time a comparable situation is encountered [30.2–4].

In the past decades, several researchers have criticized the above-mentioned point of view, in particular that human cognition can be compared with a very complicated computational machine [30.1, 5–7]. The computational requirements to perform the most optimal actions would be too high to be feasible in the

context of a natural, changing environment in which humans are acting [30.8]. Rather, behavior would emerge from interactions among various nonspecific interacting processes across the brain, body, and environment (see [30.9–14] for empirical demonstrations), which we shall refer to in this chapter as the CDS approach. According to this approach, cognition is thus *distributed* across (changing) processes of brain, body, and environment, which are intertwined [30.15] (see also related discussions in cognitive sciences on *component dominant* dynamics versus *interaction dominant* dynamics [30.16]; *computationalism/cognitivism* versus *embodiment* [30.2]; and *dissociation* view versus *dynamic* view [30.17]).

Throughout this chapter, we will discuss the reductionist and CDS approaches using illustrations of various domains that the scientific study of cognition pertains to, such as sports [30.17–20], language development [30.21–23], and scientific development [30.24, 25]. We will start with an overview of assumptions that the two classes of approaches proceed from.

### 30.1.1 Central Control versus Self-Organization

One key assumption that the reductionist approach proceeds from is that specific mechanisms are responsible for the way humans behave and learn. In this sense, some environmental stimulus is represented in the mind, and based on algorithms performed by the internal cognitive components the behavioral output is produced [30.3, 26–29]. For instance, in a sports context a football player perceives the positions of his teammates and the opponents, and cognitively computes the best next move [30.30]. Expert football players would better master this skill given their extensive knowledge base, or *software*, of previous encounters with different kinds of situations. This approach implicitly takes the computer as a metaphor in order to explain behavior, and typically conceives of the mind, or brain, as a central computing agent that encodes the environmental inputs and controls subsequent behavior (see the review of *Van Gelder* [30.7] for an extensive discussion of this view). This entails that the brain, comprising the different component processes, is considered as the *central controller* of human behavior.

The idea that the brain controls our behavior, and that the body and environment provide (only) input to the brain, was first challenged by Gibson’s ecological approach [30.31–33]. He proposed that action possibilities are not cognitively computed, but are directly guided by information from a structured environment

to which an organism attunes its actions. In other words, behavior is guided by the direct information and action exchange between the organism and its environment. Gibson's theory has laid the foundation for what has come to be known as *ecological psychology*, which shares its major assumptions with the CDS approach. In Gibson's words: "Control lies not in the brain, but in the animal-environment system" [30.33], which suggests that Gibson conceived cognition as a set of interacting components, distributed across the brain, body, and environment [30.1, 15, 34–39]. Accordingly, the CDS approach considers cognition as a dynamic process characterized by the self-organization of interacting component processes. For instance, the most appropriate next action of a football player would not be computed in terms of a sequence of computational steps, but rather take the form of an ongoing loop of flexible attunement of the action, which occurs more or less simultaneously with changes in the task constraints [30.40, 41]. Hence, the control of human behavior is not centralized, but rather an emergent process.

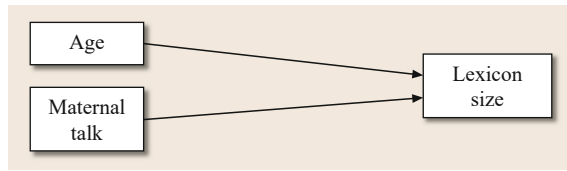
### 30.1.2 Static versus Dynamic Models

The reductionist assumption that a given behavior or psychological state is generated by one or a few components or determinants is usually associated with the construction and application of static models, in which the levels of some set of dependent variable(s) ( $y_i$ ) are directly and uniquely related to (or caused by) the levels of some other set of independent variable(s) ( $x_i$ )

$$y_i = f(x_i) . \quad (30.1)$$

In this functional description, any possible set of values of  $x_i$  generates a corresponding value for the dependent variables  $y_i$ . In other words, if we know the values of  $x_i$ , we can predict the values of  $y_i$ . An implicit assumption here is that the operating causal variables remain stable for the duration of the behavior or psychological state they would cause.

Take, as an illustration, the development of a child's lexicon. A typical *static* study would consist of assessing the maternal talk to children of different ages, for instance of 1, 2, and 3 years old. The size of the lexicon can then be predicted by explanatory variables such as age, maternal talk, or a combination of these two variables (Fig. 30.1). Note that, although age is in fact a continuously changing temporal parameter, it is used in a typically static way as a sequence of values (ages), similar to the way maternal talk is treated as a series of static values. In line with the reductionist view, the implicit assumption here is that the child's cognitive



**Fig. 30.1** A simple illustration of a reductionist explanation of a child's lexicon size

language-processing *device* is the underlying mechanism through which maternal talk affects the lexicon size [30.42, 43].

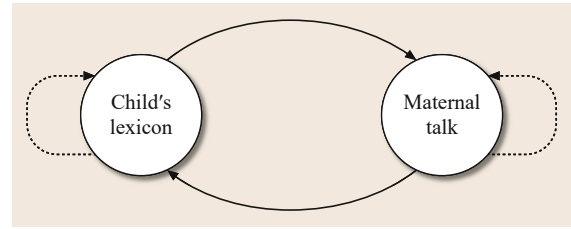
In Fig. 30.1, age and maternal talk are thus treated as the independent variables. However, in order to increase the explained variance in lexicon size, the model can be made more complicated by assuming – and statistically showing – that age is also affecting maternal talk (for all other conditions remaining equal, mothers of older children tend to produce richer maternal talk than mothers of younger children). It is even possible to also draw an arrow between age and the arrow connecting maternal talk and lexicon size, implying that age moderates the effect of maternal talk on the lexicon size (older children can profit more from the same maternal talk than younger children). Another way of increasing the complicatedness is by adding additional variables, such as socioeconomic status (SES), the mother's intelligence quotient (IQ), time spent at the day care center and so forth. Thus, generally speaking, the reductionist approach proceeds from the idea that the explanatory power should be increased by increasing the number of specific factors (i. e., components) involved and the links between them. As already mentioned, temporal change, represented by age in the example, is treated as a factor like any other factor in the model.

According to the CDS approach, however, the causal principle of behavioral or psychological change does not lie in the values of some variables or components at a certain moment in time. As noted earlier, cognition could be envisioned as a dynamic *process*, which entails that *time* is an essential aspect to take into account. More specifically, the change in the cognitive system is a function of its preceding state

$$y_{t+1} = f(y_t) , \quad (30.2)$$

where  $y_{t+1}$  corresponds to the state of the system at time  $t + 1$ , which is a function of state  $y$  at the previous time point  $t$ . Hence, the CDS approach proposes models of change that involve recursive relationships ( $y_t$  leads to  $y_{t+1}$ , which leads to  $y_{t+2}$ , and so forth) [30.44]. Returning to the explanation of a child's lexicon, a simple explanation would be that learning new words at time  $t$  depends (among possible other things) on the words the

child already knows and on the words that are spoken by the mother with whom the child communicates at time  $t$  (Fig. 30.2). This also entails that the child's lexical change is *embedded* in the environment, referring to the real-time events taking place, such as the maternal talk. Consequently, and in line with the concept of self-organization (Sect. 30.1.1), the properties of the system change over time as the underlying components interact with each other. That is, as the child's lexicon develops, the mother will change the way she speaks to the child, which feeds back to the development of the child's lexicon, and so forth (Fig. 30.2; see also the work of Van Dijk and colleagues [30.45]).



**Fig. 30.2** A simple illustration of a dynamic explanation, in which the child's lexicon at a particular moment provides the basis for the lexicon at the next time point(s), while this change is also shaped by the dynamic interactions (literally and metaphorically) with the mother

## 30.2 Explaining Cognition

In the previous sections, the terms *complicated* and *complex* have been mentioned a few times. In scientific models, these terms refer to different explanatory conceptions [30.46]. The reductionist approach generally provides a *complicated* model of cognition, whereas the CDS approach proceeds from a model of *complexity*. In the following subsections, we will explain and contrast these different kinds of modeling, and how they naturally fit with different kinds of research strategies and analyses to capture cognition.

### 30.2.1 Research Strategies and Complicated Models

In a complicated model of cognition, many components are involved, which can be studied in isolation, and the resulting psychological and performance states can be understood when knowing the contributions of the individual components (Fig. 30.1). In case researchers are interested in cognitive expertise of athletes in real time, for example, they would design a study in which cognitive measures can be obtained in a standardized environment. As an illustration, Williams and colleagues [30.30] showed skilled and less skilled football players sequences of attacking game plays in a research lab. After a brief delay, participants watched another set of sequences and indicated which sequences they already viewed before. Two variables were measured: response time and response accuracy. The authors found that skilled players responded quicker when they were shown game plays that they viewed earlier. Williams and colleagues proposed that these results indicate that expert soccer players have more refined (stored) knowledge structures of soccer game plays. Because the skilled players can access these modules quickly, they would be better able to respond rapidly [30.2].

It is assumed to be likely that this information processing mechanism facilitates decision making on the football field, which results in choosing the most appropriate (next) action (remember the computer metaphor in Sect. 30.1.1).

In order to further untangle the full richness of cognition, researchers keep increasing the complicatedness of their models by basically adding more underlying explanatory modules or variables and links between them (see also the explanation of static models in Sect. 30.1.2). For instance, referring back to the study of Williams and colleagues [30.30], they also found that skilled participants better recalled the relational information between players (how they moved relative to each other). To explain this finding, the authors added the concept of (cognitive) *motion integrators* as a possible explanation for this result, which would be integral to skilled pattern recognition [30.26].

### 30.2.2 Research Strategies and Complex Models

According to the CDS paradigm, real-time cognitive performance can be explained by a model of *complexity*. Importantly, complexity is not reflected in the number of components that are involved in cognitive performance (the number of separate cognitive modules that process information, the number of environmental variables that influence people's cognitive processing, etc.). On the contrary, a complex cognitive system is typically characterized by continuous *dynamic interactions* and multicausality between various intrapersonal and environmental components, from which the (changing) state of the cognitive system emerges (Fig. 30.2) [30.37, 47, 48]. Thus, in contrast to a complicated model of cognition, a complex model proceeds from the idea that it



is *not* possible nor feasible to reduce cognition to separate, fairly isolated components (recall the concept of self-organization again in Sect. 30.1.1).

Referring back to the example of football, according to the CDS approach the actions of a (attacking) player are emergent from the underlying self-organization dynamics (changing positions of players, the ball, etc.) [30.18, 40]. Of primary interest to researchers is therefore the unfolding of footballers' actions in real-time. A typical study would focus on the emergence of action patterns that are continuously shaped by the way the system components attune to each other, involving for instance the attackers' and defenders' relative distance to each other [30.49, 50] and to the goal [30.51], or more generally the size of the field [30.52]. As an illustration, *Headrick* and colleagues [30.51] revealed that the distance between the defender and the ball stabilized at higher values – indicating low risk-taking behavior of the defender – when the defender–attacker duel occurred close to the goal, than when it occurred relatively far from the goal.

### 30.2.3 Analyses to Untangle Cognition Based on Complicated Models

Given the different assumptions of the reductionist and CDS approaches, not only the research strategy (Sect. 30.2.2), but also the applied analyses are different [30.53]. In the reductionist approach, the analysis is focused on finding the linear associations at the level of isolated variables. Researchers therefore typically apply the so-called control of variables strategy, which in standard accounts of the scientific method is seen as the quintessential way of explaining the nature of reality, namely to disentangle the variables and control the variables separately to see what changes in those variables actually do. The way one variable controls another variable is assumed to be a property that can be isolated from other properties and other variables. The reasoning is that the most general way in which a variable can control another one is the way in which a variable co-varies with another one over the entire population. Hence, the study of the way a variable controls another one is based on samples that are big enough to be a good representative of that population. This also points to the importance of the generalizability issue, which is understood as the degree to which the statement based on a sample is true of the population that the sample is intended to represent (see also *Hasselmann* and colleagues [30.54] for a discussion of theorizing in cognitive science).

In the example of the relationship between maternal talk and a child's lexicon (Sect. 30.1.2), a researcher may analyze the effect of the quantity and sophisti-

cation of maternal lexical input on the child's lexicon [30.43, 44, 55]. The outcome is framed in terms of the variance in the child's lexicon that can be explained by (co-varies with) the variance in the maternal input variables. In other words, the researcher attempts to find a linear relationship between the lexical input of the mother and a child's lexicon (the output variable). The relationship between maternal input and a child's lexicon as it is found across a sample of mother–child dyads, is implicitly assumed to govern the process of language learning at the level of individual children [30.56].

### 30.2.4 Analyses to Capture the Complexity of Cognition

According to the CDS approach, the associations between variables as they are observed at the sample level, cannot be used as valid approximations of the dynamic relations that govern the process. More specifically, if we assume that components change over time, influence each other reciprocally, which gives rise to (changing) patterns of behavior, analyzing associations between variables in large samples cannot tell us how the process actually works (cf. the ergodicity problem as described by *Molenaar* and colleagues [30.57, 58]). According to CDS theorists, if a researcher is interested in *why* and *how* actual change occurs, the process of interest should be studied over time [30.44, 53, 56, 59, 60]. Therefore, researchers often apply time series analyses, and they focus on particular signatures of the time series, as well as on the underlying dynamic rules that may explain the dynamics of the time series.

*Van Geert* and colleagues have conducted several studies on language development from a CDS perspective [30.21, 22, 61–63]. The authors consistently found discontinuities in individual children's language development, which provided valuable information about lexical change. For example, *Bassano* and *Van Geert* [30.21] studied early language development among French children, and they showed that the discontinuities in the time series mark the transition from a one-word to a combinational mode, and from single combinations to more abstract syntactic modes of language. In CDS terms, the language modes can be considered as *attractors*, that is, states or patterns toward which the system tends to converge [30.36, 48, 64–66]. Thus, the increase in variation signals the transition to another attractor, and thereby to another milestone in children's language development (see also the work of *Van Dijk* and *Van Geert* [30.62]).

Interestingly, whereas variation patterns carry highly valuable information about the cognitive process

according to the CDS approach, variation is typically considered as *random* error according to the reductionist approach. In classical repeated measures designs, for instance, variation around the (linear) tendency over time is considered as error variance. On the contrary, periods of variation have been consistently found to be markers of a transition stage to another attractor in a variety of cognition-related (dynamical) research, not only in studies on language development (see the example above), but also on cognitive reasoning [30.67, 68], perception [30.69], and motor control [30.70].

Finally, returning to the study of *Bassano* and *Van Geert* [30.21], they proposed a mathematical model, defined as a dynamic growth model in which the growth of the language modes, and how they mutually influence each other, could be reliably modeled for the

individual children. This means that the authors were not concerned with providing a model of the average language development across the population of children, which is typical for the reductionist approach, and which would probably result in an unrealistic picture of what individual language development may look like (many children do not develop according to the statistically average child). Rather, *Bassano* and *Van Geert* proposed a dynamic model that could also be *generalized* to the individual. In other words, the authors provided insights into the lawful mechanisms, or CDS principles, underlying language development over time (for a comparable example of model building in mother–child linguistic interactions and the associated developmental process in the child, see the recent work of *Van Dijk* and colleagues [30.45]).

### 30.3 Is Cognition Best Explained by a Complicated or Complex Model?

Currently, the majority of researchers in behavioral and social sciences apply the reductionist approach, whereas a relatively small group of researchers applies the CDS approach. Ideally, researchers should proceed from the kind of approach that fits best with the research question and/or topic under study. For example, if a researcher was interested in the effect of maternal smoking during pregnancy on children’s later academic achievements, a reductionist approach may provide the best fit, because it is desirable that the variables of interest are studied in isolation to determine the effect. Indeed, when this question addresses the population level it can, for instance, be used in campaigns and medical advice. In a typical study, *Batstra* and colleagues [30.71] adjusted for confounders such as socioeconomic status and pre- and perinatal complications, and across 1186 children they found that maternal smoking during pregnancy was independently related to the children’s arithmetic and spelling skills between the ages of 5.5 and 11 years. Note that this study was not focused on explaining cognition, but rather on one potential risk factor, that is, the distribution of maternal smoking across the population and its statistical association with a population-defined effect (distribution of arithmetic and spelling skills).

However, as discussed earlier, the reductionist approach is also widely used to provide an understanding of the (complicated) mechanism that drives cognitive performance in real time, as well as cognitive development across the life span. The extent to which the reductionist approach on the one hand, or the CDS approach on the other hand, is most applicable to cognition that depends on whether it is a *complicated* or

*complex* model that is best able to explain cognition. In the next section we will discuss some studies, the outcomes of which render one of the two approaches more or less convincing. We will first discuss studies focused on cognitive performance in real time, after which we will discuss cognitive development across the life span.

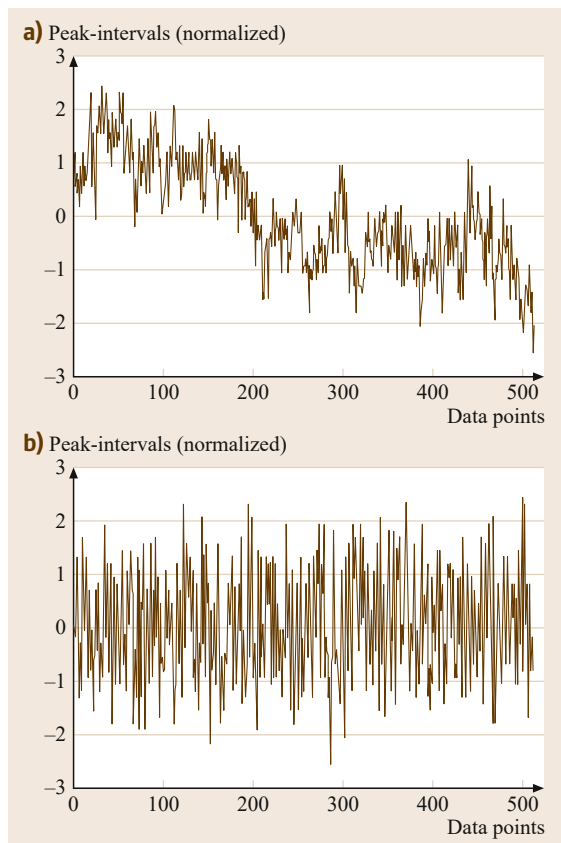
#### 30.3.1 Explaining Real-Time Cognitive Performance

In a recent study, *Den Hartigh* and colleagues [30.72] were interested in the mechanism underlying the (cognitive) control of a motor (rowing) task. The authors let rowers perform a practice session on rowing ergometers, consisting of 550 strokes at the rowers’ preferred rhythm. A force sensor was attached to the handle of the ergometer, which measured the exerted force of the rowers at 100 Hz. Subsequently, the authors analyzed the time series of the durations from force-peak-to-force peak (the force peak intervals). With the reductionist approach in mind, one would expect that each new stroke is controlled by specific modules or component processes (e.g., central pattern generators [30.73]). This entails that each new stroke would be independently controlled from the previous stroke, and that the results should reveal interval series characterized by some average interval value with random variation around it (recall that variation is typically treated as random noise in the reductionist approach).

On the other hand, a CDS is characterized by an iterative process involving interactions between various component processes at different levels (e.g., in this case cell activity, muscle contractions, limb move-

ments) and across multiple time scales (e.g., from a few seconds to several minutes of performance [30.74]). Such ongoing component interactions would cooperatively generate the rower's performance, and are assumed to generate time series characterized by a structured pattern of variation, called pink (or  $1/f$ ) noise. More specifically, the coordination among interacting component processes across multiple time scales within the system, and between the system and its (task) environment, would result in small fluctuations on a short time scale (a few rowing strokes) that are nested in larger fluctuations across longer time scales (tens or hundreds of strokes). The temporal structure of variation can be quantified in terms of the fractal dimension (FD): A FD close to 1.5 corresponds to random (white) noise, and a FD close to 1.2 corresponds to pink noise [30.38].

Figure 30.3a provides a representative example of a time series of one of the rowers. Based on only visual inspection, one can observe that minor fluctuations are embedded in waves of larger fluctuations. In line with this (seemingly) structured pat-



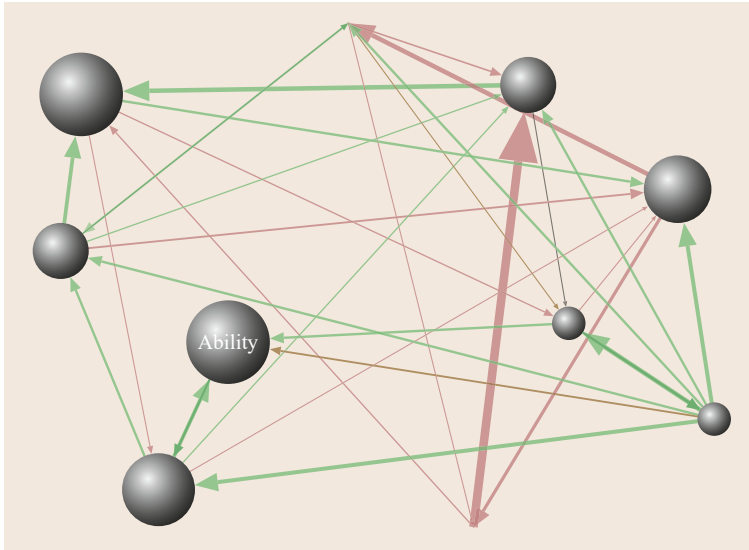
**Fig. 30.3a,b** A rower's actual peak-interval series (a) and shuffled interval series (b)

tern of variation, we found a FD of 1.22, which is close to pink noise, and a comparable pattern was found for the other rowers in the sample. These results strongly suggest that the performance of the rowers emerged from complexity, that is the *interaction* between many components on various scales, as opposed to (just) the *contribution* of many components. Figure 30.3b displays the performance data of the same rower as the one in Figure 30.3a, but in this case the force-peak interval data are randomized. Hence, the average interval and the *size* of the variation (standard deviation) are exactly the same in the two graphs, only the temporal order is different. In line with the fact that this randomization made each next rowing stroke independent of the previous one(s), we found a FD close to 1.5, which reflects random noise.

In line with the study of *Den Hartigh* and colleagues [30.72], the occurrence of pink noise in cognitive performance seems a universal phenomenon [30.39, 75]. Virtually any cognitive or motor performance in which time series of healthy individuals are analyzed reveal pink noise patterns, ranging from reaction times in psychological experiments and reading fluency, to stride intervals of human gait and rhythmical aiming tasks [30.14, 16, 34, 38, 74, 76–80]. These studies provide robust and converging evidence to the claims of the CDS approach, which makes it likely that real-time cognitive performance emerges from complexity, and cannot be reduced to separate, rather independently operating components that perform specific functions to control human behavior [30.72].

### 30.3.2 Explaining Long-Term Cognitive Development

In order to discover the model underlying cognitive development on the long term (e.g., the life span), computer simulations provide a useful tool [30.23, 81, 82]. Computer simulations can be used to (a) generate predictions in terms of which developmental patterns are generated by which kinds of model principles, and (b) compare the simulation results with actual data collected in longitudinal studies on cognitive development. As an illustration, take the development of cognitive abilities in terms of scientific talent development. According to the literature, in order to develop one's scientific abilities various factors play a role, including genetic endowment, the individual's interest and commitment, as well as environmental variables such as the support of the teacher, family support, and so forth [30.83, 84]. As noted earlier, the reductionist approach attempts to fit a complicated model to explain how a state of cognitive development can be



**Fig. 30.4** Graphical representation of a CDS model of scientific ability development. Note that this is a snapshot, and that the network constantly develops through changes in the levels of the nodes, among others as a consequence of the interactions with other nodes

predicted by particular determining factors, often including the age of the child, in a linear fashion (see also Sect. 30.1.2).

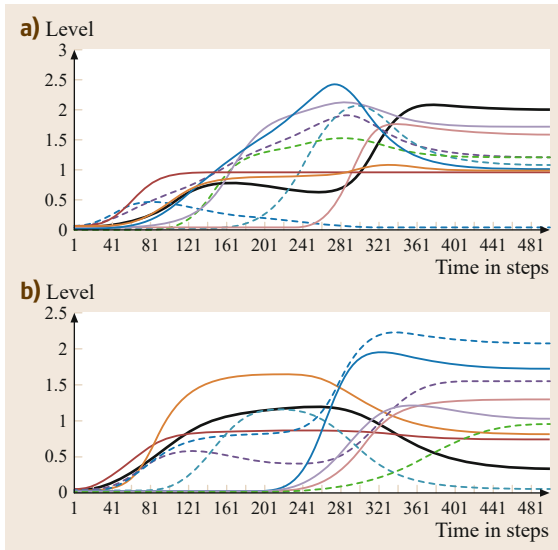
Interestingly, the literature on human development, and more specifically cognitive development, hardly reveals linear patterns [30.81]. In the specific case of scientific talent development, some defining properties have been summarized by *Simonton* in one of his articles on talent [30.85]. An example of these properties is that a similar form of (scientific) talent may emerge at different ages. Another is that the level of talent is not necessarily monotonically raising or stable. It can change or even disappear during a person's life span.

We will briefly show how to apply computer modeling to test whether a particular model would be able to generate valid predictions of scientific talent development, such as the properties mentioned above. In line with the CDS approach, we will demonstrate a model in which development is shaped by the ongoing interactions with other components, which also undergo change. The key mathematical principles of such a (relatively simple) dynamic systems model consist of the scientific ability ( $L$ ) that changes over time ( $t$ ) as a function of two kinds of resources. One remains relatively stable across time ( $K$ ), for instance, the individual's genetic endowment. The second type of resource ( $V$ ) may change on the same time scale as the change of the scientific ability, and comprises components such as commitment and teacher support. These components may dynamically interact with the scientific ability component and with each other. The interaction between the components is governed by a number of parameters, including the degree in which

an ability profits from the constant resources ( $r$ ), the weight of the connection ( $s$ ) with other components ( $i, j$ , etc.) and a general limiting factor ( $C$ ) that keeps the growth within realistic maximum values

$$\left. \begin{aligned} \frac{\Delta L_A}{\Delta t} &= \left( r_{L_A} L_A \left( 1 - \frac{L_A}{K_{L_A}} \right) + \sum_{v=1}^{v=i} s_v L_A V_v \right) \left( 1 - \frac{L_A}{C_A} \right) \\ \frac{\Delta L_B}{\Delta t} &= \left( r_{L_B} L_B \left( 1 - \frac{L_B}{K_{L_B}} \right) + \sum_{v=1}^{v=j} s_v L_B V_v \right) \left( 1 - \frac{L_B}{C_B} \right) \\ &\dots \\ &\dots \\ &\dots \end{aligned} \right\} \quad (30.3)$$

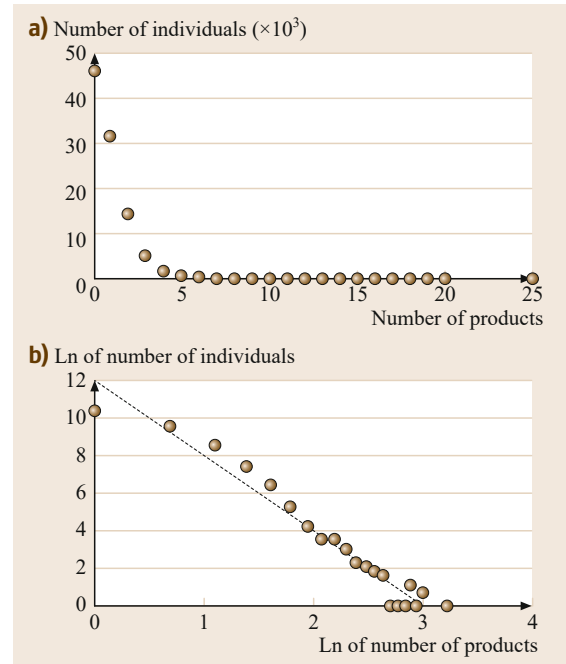
For simplicity, we simulated a system consisting of 10 components. Each simulation represents a particular individual trajectory, which is based on initial parameter values that were randomly drawn from symmetric distributions. Furthermore, the average degree of connectivity between the nodes is 25%, and the connections are randomly distributed over the nodes [30.86]. Figure 30.4 provides a graphical representation of a typical network of relationships specified by this kind of model. The nodes correspond to different variables that interact with the ability growth and with each other (think of the individual's commitment and family support). The sizes of the nodes reflect the magnitudes of the variables. Furthermore, each directed arrow between two nodes represents a supportive (green) or competitive effect (brown) of one variable on another. The strength of the relationships between the variables is reflected in the thickness of the edges.



**Fig. 30.5a,b** Simulations of the scientific ability development of two individuals (**a,b**). The *black solid lines* in the graphs represent the abilities, the *other lines* reflect the dynamic variables that have supportive, competitive, or neutral relationships with the ability and with each other

When simulating various individuals based on the dynamic systems principles explained earlier, a first observation is that the model reveals very different patterns of scientific ability development. Figure 30.5 provides two representative simulation examples. The black lines in the graphs correspond to the scientific ability, whereas the other lines correspond to the other variables. The figure shows that the ability development of one individual develops in a step-wise fashion, and reaches a plateau during the second half of the life cycle (Fig. 30.5a). On the other hand, the ability of the second individual (Fig. 30.5b) starts with a relatively rapid increase, which levels off, and in the second half of this individual's life span the ability development declines. Together, Fig. 30.5a and b correspond to the typical properties of scientific talent development, namely that it can take different forms, that it is not a linear (monotonic) process, and that talent may diminish or disappear over time [30.85] (for more extensive demonstrations of dynamic systems modeling of cognitive development, see the work of *Van Geert* [30.23, 81, 82]).

A final observation in the literature is that exceptional abilities are rare [30.86–89]. Specific abilities, such as the scientific ability required to write papers for high-ranked journals, are in most cases only measurable by referring to the typical performances or products (i. e., the number of published articles in high-



**Fig. 30.6a,b** Simulated productivity distribution according to the complexity model. The raw (simulated) data are displayed in graph (**a**), and the natural log–log representation, which approaches a straight line, is displayed in graph (**b**)

ranked journals). The actual productivity of scientists is extremely right skewed with very few researchers having many high-impact publications and relatively many with one high-impact publication. In fact, the distribution is so right skewed that the log–log representation corresponds with a straight line [30.87, 90–92]. In combination with various product models discussed in the literature [30.87, 93], simulations of the CDS model reveal an extremely skewed distribution that is in accordance with the distribution of scientific productivity of scientists in various scientific domains (Fig. 30.6). The typical reductionist model would try to explain the product distributions on the basis of linear combinations of underlying predictor variables. However, such a model is unable to predict the typical and ubiquitous heavy-skewed distribution of the products, in this case the publications [30.86].

Taken together, based on data on real-time cognitive performance and computer modeling of long-term cognitive development, researchers can choose the model that most likely underlies the empirically observed patterns. We have presented some examples of data that can be better explained by model predictions stemming from the CDS approach (a complex model) than from a reductionist approach (a complicated model).

## 30.4 Conclusion

Every human behaves and develops in a different way, and is embedded in a rich, constantly changing environment. This has made it challenging for scientists to explain cognitive development and the control of human behavior. In the past decades, human cognition has, on the one hand, been approached as localized in the brain and controlled by separate components, and, on the other hand, as a dynamic process consisting of nonlocalized interacting component processes. The first approach – the reductionist approach – assumes that research practice should be focused on finding the explanation of cognition in the specific functions of the components, whereas the second approach – the complex dynamic systems approach – assumes that we should focus on the underlying (complex) dynamic principles to understand cognition. In our belief, researchers often apply the approach that they and their close colleagues are most familiar and comfortable with. Often, this is the reductionist approach, which has been widely applied in social and behavioral sciences since the cognitive revolution in the 1950s, whereas the CDS approach has relatively slowly gained ground since the 1990s [30.7, 34, 46, 65, 81].

In this chapter, we started with an overview of some key differences between the approaches without taking a position in which of the two is the *better* one (Sects. 30.1 and 30.2). Subsequently, we discussed findings on real-time processes and long-term cognitive development (Sect. 30.3). First, we showed that cognitive performance measured in real-time reveals a structured pattern of variation (pink noise), which is difficult to reconcile with the reductionist view according to which a pattern of random variation would be expected. On the other side, it fits with the CDS

approach that cognitive performance emerges from ongoing component interactions, resulting in a time series in which short-term adaptations are embedded in slower but larger changes (Sect. 30.3.1).

Second, we demonstrated predictions that were focused on long-term cognitive development (i. e., scientific ability development). The reductionist approach assumes that cognitive development is shaped by the addition of relevant explanatory components or variables (e.g., genetic endowment, commitment, and teacher support), whereas the CDS approach proceeds from the idea that cognitive development is shaped by the ongoing *dynamic interaction* between the relevant variables. We showed that some typical properties of cognitive development, scientific talent development in particular, are generated by a model that is based on CDS principles (Sect. 30.3.2).

The plausible predictions that followed from the CDS approach suggest that cognition can best be explained by a *complex* model. Therefore, in light of future model building, we hope that researchers who apply the reductionist approach will keep an open mind regarding the potential of the CDS approach to capture the full richness of cognition and behavior. At the same time, CDS theorists should continue exploring whether a reductionist explanation may also fit with obtained results on (time-serial) cognitive processes. By doing so, researchers will be in a better position to provide a model to unlock the *mystery of the three pounds of matter between our ears*, and, importantly, how this is situated in our bodies and the environment we interact with. Given the current state of knowledge, we should keep in mind that the answer to this mystery, and the model we need, may not be *complicated*, but *complex*.

## References

- |      |  |       |   |
|------|--|-------|---|
| 30.1 | M.J. Richardson, K.L. Marsh, R.C. Schmidt: Challenging the egocentric view of coordinated perceiving, acting, and knowing. In: <i>The Mind in Context</i> , ed. by L.F. Barrett, B. Mesquita, E. Smith (Guilford, New York 2010), Chap. 15 | 30.6  | A. Clark: An embodied cognitive science?, <i>Trends Cogn. Sci.</i> <b>3</b> , 345–351 (1999)  |
| 30.2 | K.A. Ericsson, W. Kintsch: Long-term working memory, <i>Psychol. Rev.</i> <b>102</b> , 211–245 (1995)  | 30.7  | T. Van Gelder: What might cognition be, if not computation?, <i>J. Philos.</i> <b>92</b> , 345–381 (1995)   |
| 30.3 | A.B. Markman, E. Dietrich: In defense of representation, <i>Cogn. Psychol.</i> <b>40</b> , 138–171 (2000)  | 30.8  | H.A. Simon: A behavioral model of rational choice, <i>Q. J. Econ.</i> <b>69</b> , 99–118 (1955)   |
| 30.4 | K. Yarrow, P. Brown, J.W. Krakauer: Inside the brain of an elite athlete: The neural processes that support high achievement in sports, <i>Nat. Rev. Neurosci.</i> <b>10</b> , 585–596 (2009)  | 30.9  | R.F.A. Cox, W. Smitsman: Action planning in young children's tool use, <i>Dev. Sci.</i> <b>9</b> , 628–641 (2006)   |
| 30.5 | A. Chemero: Anti-representationalism and the dynamical stance, <i>Philos. Sci.</i> <b>67</b> , 625–647 (2000)  | 30.10 | P. Fitzpatrick, R. Diorio, M.J. Richardson, R.C. Schmidt: Dynamical methods for evaluating the time-dependent unfolding of social coordination in children with autism, <i>Front. Integr. Neurosci.</i> <b>7</b> (2013), doi: <a href="https://doi.org/10.3389/fnint.2013.00021">10.3389/fnint.2013.00021</a> |
|      |  | 30.11 | K.L. Marsh, R.W. Isenhower, M.J. Richardson, M. Helt, A.D. Verbalis, R.C. Schmidt, D. Fein: Autism  |

- and social disconnection in interpersonal rocking, *Front. Integr. Neurosci.* **7** (2013), doi:[10.3389/fnint.2013.00004](https://doi.org/10.3389/fnint.2013.00004)
- 30.12 E. Thelen, G. Schöner, C. Scheier, L.B. Smith: The dynamics of embodiment: A field theory of infant perseverative reaching, *Behav. Brain Sci.* **24**, 1–34 (2001)
- 30.13 M. Varlet, L. Marin, S. Raffard, R.C. Schmidt, D. Capdevielle: J.P., Boulenger, J. Del-Monte, B.G. Bardy: Impairments of social motor coordination in schizophrenia, *PLoS ONE* (2012), doi:[10.1371/journal.pone.0029772](https://doi.org/10.1371/journal.pone.0029772)
- 30.14 M.L. Wijnants, F. Hasselman, R.F.A. Cox, A.M.T. Bosman, G. Van Orden: An interaction-dominant perspective on reading fluency and dyslexia, *Ann. Dyslexia* **62**, 100–119 (2012)
- 30.15 R.F. Port, T. van Gelder: *Mind as Motion: Explorations in the Dynamics of Cognition* (MIT Press, Cambridge 1995)
- 30.16 G.C. Van Orden, J.G. Holden, M.T. Turvey: Human cognition and  $1/f$  scaling, *J. Exp. Psychol. Gen.* **134**, 117–123 (2005)
- 30.17 R.F.A. Cox, A.W. Smitsman: Special section: Towards an embodiment of goals, *Theor. Psychol.* **18**, 317–339 (2008)
- 30.18 D. Araujo, K. Davids, R. Hristovski: The ecological dynamics of decision making in sport, *Psychol. Sport Exerc.* **7**, 653–676 (2006)
- 30.19 R.J.R. Den Hartigh, S. Van Der Steen, M. De Meij, N. Van Yperen, C. Gernigon, P.L.C. Van Geert: Characterising expert representations during real-time action: A Skill Theory application to soccer, *J. Cogn. Psychol.* **26**, 754–767 (2014)
- 30.20 A.M. Williams: Perceptual skill in soccer: Implications for talent identification and development, *J. Sports Sci.* **18**, 737–750 (2000)
- 30.21 D. Bassano, P. Van Geert: Modeling continuity and discontinuity in utterance length: A quantitative approach to changes, transitions and intra-individual variability in early grammatical development, *Dev. Sci.* **10**, 588–612 (2007)
- 30.22 M. Van Dijk, P. Van Geert: Wobbles, humps and sudden jumps: A case study of continuity, discontinuity and variability in early language development, *Infant Child Dev.* **16**, 7–33 (2007)
- 30.23 P. Van Geert: A dynamic systems model of cognitive and language growth, *Psychol. Rev.* **98**, 3–53 (1991)
- 30.24 G. Park, D. Lubinski, C.P. Benbow: Contrasting intellectual patterns predict creativity in the arts and sciences tracking intellectually precocious youth over 25 years, *Psychol. Sci.* **18**, 948–952 (2007)
- 30.25 S. Van Der Steen, H. Steenbeek, M.W.G. Van Dijk, P. Van Geert: A process approach to children's understanding of scientific concepts: A longitudinal case study, *Learn. Indiv. Diff.* **30**, 84–91 (2014)
- 30.26 W.H. Dittrich: Seeing biological motion – Is there a role for cognitive strategies? In: *Gesture-Based Communication in Human-Computer Interaction*, Lecture Notes in Computer Science, Vol. 1739, ed. by A. Braffort, R. Gherbi, S. Gibet, D. Teil, J. Richardson (Springer, Berlin 1999) pp. 3–22
- 30.27 B. Hommel: The cognitive representation of action: Automatic integration of perceived action effects, *Psychol. Res.* **59**, 176–186 (1996)
- 30.28 T. Schack, H. Ritter: The cognitive nature of action – Functional links between cognitive psychology, movement science, and robotics, *Prog. Brain Res.* **174**, 231–250 (2009)
- 30.29 P. Thagard: *Mind: Introduction to Cognitive Science* (MIT Press, Cambridge 2005)
- 30.30 A.M. Williams, N.J. Hodges, J.S. North, G. Barton: Perceiving patterns of play in dynamic sport tasks: Investigating the essential information underlying skilled performance, *Perception* **35**, 317–332 (2006)
- 30.31 J.J. Gibson: *The Senses Considered as Perceptual Systems* (Houghton-Mifflin, Boston 1966)
- 30.32 J.J. Gibson: The theory of affordances. In: *Perceiving, Acting and Knowing: Toward an Ecological Psychology*, ed. by R. Shaw, J. Bransford (Lawrence Erlbaum Associates, New York 1977) pp. 67–82
- 30.33 J.J. Gibson: *The Ecological Approach to Visual Perception* (Houghton-Mifflin, Boston 1979)
- 30.34 C.T. Kello, B.C. Beltz, J.G. Holden, G.C. Van Orden: The emergent coordination of cognitive function, *J. Exp. Psychol. Gen.* **136**, 551–568 (2007)
- 30.35 P.N. Kugler, M.T. Turvey: *Information, Natural Law, and The Self-Assembly of Rhythmic Movement* (Lawrence Erlbaum Associates, Hillsdale 1987)
- 30.36 E. Thelen, L.B. Smith: *A Dynamic Systems Approach to the Development of Cognition and Action* (MIT Press, Cambridge 1994)
- 30.37 P.L.C. Van Geert, K.W. Fischer: Dynamic systems and the quest for individual-based models of change and development. In: *Toward a New Grand Theory of Development? Connectionism and Dynamic Systems Theory Reconsidered*, ed. by J.P. Spencer, M.S.C. Thomas, J. McClelland (Oxford Univ. Press, Oxford 2009), Chap. 16
- 30.38 G.C. Van Orden, J.G. Holden, M.T. Turvey: Self-organization of cognitive performance, *J. Exp. Psychol. Gen.* **132**, 331–335 (2003)
- 30.39 M.L. Wijnants: A review of theoretical perspectives in cognitive science on the presence of scaling in coordinated physiological and cognitive processes, *J. Nonlinear Dyn.* (2014), doi:[10.1155/2014/962043](https://doi.org/10.1155/2014/962043)
- 30.40 K. Davids, D. Araújo: The concept of Organismic Asymmetry in sport science, *J. Sci. Med. Sport* **13**, 633–640 (2010)
- 30.41 J.F. Grehaigne, D. Bouthier, B. David: Dynamic-system analysis of opponent relationships in collective actions in soccer, *J. Sports Sci.* **15**, 137–149 (1997)
- 30.42 N. Hurtado, V.A. Marchman, A. Fernald: Does input influence uptake? Links between maternal talk, processing speed and vocabulary size in Spanish-learning children, *Dev. Sci.* **11**, F31–F39 (2008)
- 30.43 A. Weisleder, A. Fernald: Talking to children matters early language experience strengthens processing and builds vocabulary, *Psychol. Sci.* **24**, 2143–2152 (2013)
- 30.44 P.L.C. Van Geert: Nonlinear complex dynamic systems in developmental psychology. In: *Chaos and Complexity in Psychology: The Theory of Nonlinear*

- Dynamical Systems*, ed. by S.J. Guastello, M. Koopmans, D. Pincus (Cambridge Univ. Press, New York 2009) pp. 242–281
- 30.45 M. Van Dijk, P. Van Geert, K. Korecky-Kröll, I. Maillochon, S. Laaha, W.U. Dressler, D. Bassano: Dynamic adaptation in child–adult language interaction, *Lang. Learn.* **63**, 243–270 (2013)
- 30.46 J.M. Ottino: Engineering complex systems, *Nature* **427**, 399 (2004)
- 30.47 C.T. Kello, G.D. Brown, R. Ferrer-i-Cancho, J.G. Holden, K. Linkenkaer-Hansen, T. Rhodes, G.C. Van Orden: Scaling laws in cognitive sciences, *Trends Cogn. Sci.* **14**, 223–232 (2010)
- 30.48 J.A.S. Kelso: *Dynamic Patterns: The Self-Organization of Brain and Behavior* (MIT Press, Cambridge 1995)
- 30.49 R. Duarte, D. Araújo, L. Freire, H. Folgado, O. Fernandes, K. Davids: Intra- and inter-group coordination patterns reveal collective behaviors of football players near the scoring zone, *Hum. Mov. Sci.* **31**, 1639–1651 (2012)
- 30.50 R. Duarte, D. Araújo, K. Davids, B. Travassos, V. Gazimba, J. Sampaio: Interpersonal coordination tendencies shape 1-vs-1 sub-phase performance outcomes in youth soccer, *J. Sports Sci.* **30**, 871–877 (2012)
- 30.51 J. Headrick, K. Davids, I. Renshaw, D. Araújo, P. Passos, O. Fernandes: Proximity-to-goal as a constraint on patterns of behaviour in attacker-defender dyads in team games, *J. Sports Sci.* **30**, 247–253 (2012)
- 30.52 W. Frencken, J. Van Der Plaats, C. Visscher, K. Lemmink: Size matters: Pitch dimensions constrain interactive team behaviour in soccer, *J. Syst. Sci. Complex.* **26**, 85–93 (2013)
- 30.53 P. Van Geert: The contribution of complex dynamic systems to development, *Child Dev. Perspect.* **5**, 273–278 (2011)
- 30.54 F. Hasselman, M.P. Seevinck, R.F.A. Cox: *Caught in the Undertow: There is structure beneath the ontic stream*, SSRN 2010)
- 30.55 Z.O. Weizman, C.E. Snow: Lexical output as related to children's vocabulary acquisition: Effects of sophisticated exposure and support for meaning, *Dev. Psychol.* **37**, 265–279 (2001)
- 30.56 P. Van Geert: Group versus individual data in a dynamic systems approach to development, *Enfance* **2014**, 283–312 (2014)
- 30.57 P.C. Molenaar: A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever, *Measurement* **2**, 201–218 (2004)
- 30.58 P.C.M. Molenaar, C.G. Campbell: The new person-specific paradigm in psychology, *Curr. Dir. Psychol. Sci.* **18**, 112–117 (2009)
- 30.59 R.J.R. Den Hartigh, C. Gernigon, N.W. Van Yperen, L. Marin, P.L.C. Van Geert: How psychological and behavioral team states change during positive and negative momentum, *PLoS ONE* (2014), doi:[10.1371/journal.pone.0097887](https://doi.org/10.1371/journal.pone.0097887)
- 30.60 P. van Geert, H. Steenbeek: Explaining after by before: Basic aspects of a dynamic systems approach to the study of development, *Dev. Rev.* **25**, 408–442 (2005)
- 30.61 R. Ruhland, P. Van Geert: Jumping into syntax: Transitions in the development of closed class words, *Brit. J. Dev. Psychol.* **16**, 65–95 (1998)
- 30.62 M. Van Dijk, P. Van Geert: Disentangling behavior in early child development: Interpretability of early child language and its effect on utterance length measures, *Infant Behav. Dev.* **28**, 99–117 (2005)
- 30.63 P. Van Geert, M. Van Dijk: Focus on variability: New tools to study intra-individual variability in developmental data, *Infant. Behav. Dev.* **25**, 340–374 (2002)
- 30.64 W. Briki, R.J.R. Den Hartigh, K.D. Markman, C. Gernigon: How do supporters perceive positive and negative psychological momentum changes during a simulated cycling competition?, *Psychol. Sport Exerc.* **15**, 216–221 (2014)
- 30.65 A. Nowak, R.R. Vallacher: *Dynamical Social Psychology* (Guilford, New York 1998)
- 30.66 P. Van Geert: The dynamic systems approach in the study of L1 and L2 acquisition: An introduction, *Mod. Lang. J.* **92**, 179–199 (2008)
- 30.67 B.R. Jansen, H.L. Van der Maas: Evidence for the phase transition from Rule I to Rule II on the balance scale task, *Dev. Rev.* **21**, 450–494 (2001)
- 30.68 H.L. Van der Maas, P.C. Molenaar: Stagewise cognitive development: An application of catastrophe theory, *Psychol. Rev.* **99**, 395–417 (1992)
- 30.69 H.S. Hock, J.A.S. Kelso, G. Schöner: Bistability and hysteresis in the organization of apparent motion patterns, *J. Exp. Psychol. Hum. Percept. Perform.* **19**, 63–80 (1993)
- 30.70 B. Bardy, O. Oullier, R.J. Bootsma, T.A. Stoffregen: Dynamics of human postural transitions, *J. Exp. Psychol. Hum. Percept. Perform.* **28**, 499–514 (2002)
- 30.71 L. Batstra, M. Hadders-Algra, J. Neeleman: Effect of antenatal exposure to maternal smoking on behavioural problems and academic achievement in childhood: Prospective evidence from a Dutch birth cohort, *Early Hum. Dev.* **75**, 21–33 (2003)
- 30.72 R.J.R. Den Hartigh, R.F.A. Cox, C. Gernigon, N.W. Van Yperen, P.L.C. Van Geert: Pink noise in rowing ergometer performance and the role of skill level, *Motor Control* **19**, 355–369 (2015)
- 30.73 M.R. Dimitrijevic, Y. Gerasimenko, M.M. Pinter: Evidence for a spinal central pattern generator in humans, *Ann. N. Y. Acad. Sci.* **860**, 360–376 (1998)
- 30.74 M.L. Wijnants, R.F.A. Cox, F. Hasselman, A.M.T. Bosman, G. Van Orden: A trade-off study revealing nested timescales of constraint, *Front. Physiol.* **3**, 116 (2012), doi:[10.3389/fphys.2012.00116](https://doi.org/10.3389/fphys.2012.00116)
- 30.75 C.T. Kello: Critical branching neural networks, *Psychol. Rev.* **120**, 230–254 (2013)
- 30.76 N.M. De Ruiter, R.J.R. Den Hartigh, R.F.A. Cox, P.L.C. Van Geert, E.S. Kunnen: The temporal structure of state self-esteem variability during parent-adolescent interactions: More than random fluctuations, *Self Identity* **14**, 314–333 (2015)
- 30.77 D.L. Gilden: Cognitive emissions of  $1/f$  noise, *Psychol. Rev.* **108**, 33–56 (2001)



- 30.78 A.L. Goldberger, L.A. Amaral, J.M. Hausdorff, P.C. Ivanov, C.K. Peng, H.E. Stanley: Fractal dynamics in physiology: Alterations with disease and aging, *Proc. Natl. Acad. Sci.* **99**, 2466–2472 (2002)
- 30.79 J.M. Hausdorff, Y. Ashkenazy, C.K. Peng, P.C. Ivanov, H.E. Stanley, A.L. Goldberger: When human walking becomes random walking: Fractal analysis and modeling of gait rhythm fluctuations, *Phys. Stat. Mech. Appl.* **302**, 138–147 (2001)
- 30.80 M.L. Wijnants, A.M. Bosman, F. Hasselman, R.F.A. Cox, G. Van Orden:  $1/f$  scaling in movement time changes with practice in precision aiming, *Nonlinear Dyn. Psychol. Life Sci.* **13**, 75–94 (2009)
- 30.81 P. Van Geert: *Dynamic Systems of Development: Change Between Complexity and Chaos* (Harvester, New York 1994)
- 30.82 P. Van Geert: Dynamic modeling for development and education: From concepts to numbers, *Mind Brain Educ.* **8**, 57–73 (2014)
- 30.83 D. Lubinski: Exceptional cognitive ability: The phenotype, *Behav. Genet.* **39**, 350–358 (2009)
- 30.84 D. Lubinski, C.P. Benbow: Study of mathematically precocious youth after 35 years: Uncovering antecedents for the development of math-science expertise, *Perspect. Psychol. Sci.* **1**, 316–345 (2006)
- 30.85 D.K. Simonton: Talent development as a multidimensional, multiplicative, and dynamic process, *Curr. Dir. Psychol. Sci.* **10**, 39–43 (2001)
- 30.86 R.J.R. Den Hartigh, M.W.G. Van Dijk, H.W. Steenbeek, P.L.C. Van Geert: A dynamic network model to explain the development of excellent human performance, *Front. Psychol.* **7** (2016), doi:[10.3389/fpsyg.2016.00532](https://doi.org/10.3389/fpsyg.2016.00532)
- 30.87 J.C. Huber, R. Wagner-Dobler: Scientific production: A statistical analysis of authors in mathematical logic, *Scientometrics* **50**, 323–337 (2001)
- 30.88 E. O'Boyle Jr, H. Aguinis: The best and the rest: Revisiting the norm of normality of individual performance, *Pers. Psychol.* **65**, 79–119 (2012)
- 30.89 D.K. Simonton: Talent and its development: An emergenic and epigenetic model, *Psychol. Rev.* **106**, 435–457 (1999)
- 30.90 J. Laherrere, D. Sornette: Stretched exponential distributions in nature and economy: "Fat tails" with characteristic scales, *Eur. Phys. J.* **2**, 525–539 (1998)
- 30.91 S. Redner: How popular is your paper? An empirical study of the citation distribution, *Eur. Phys. J. B.* **4**, 131–134 (1998)
- 30.92 M. Sutter, M.G. Kocher: Power laws of research output, *Scientometrics* **51**, 405–414 (2001)
- 30.93 J.C. Huber: A statistical analysis of special cases of creativity, *J. Creat. Behav.* **34**, 203–225 (2000)

# 31. From Neural Circuitry to Mechanistic Model-Based Reasoning

Jonathan Waskan

Model-based reasoning in science is often carried out in an attempt to understand the kinds of mechanical interactions that might give rise to particular occurrences. One hypothesis regarding in-the-head reasoning about mechanisms is that scientists rely upon mental models that are like scale models in crucial respects. Behavioral evidence points to the existence of these mental models, but questions remain about the neural plausibility of this hypothesis.

This chapter will provide an overview of the psychological literature on mental models of mechanisms with a specific focus on the question of how representations that share the distinctive features of scale models might be realized by neural machinations. It is shown how lessons gleaned from the computational simulation of mechanisms and from neurological research on mental maps in rats can be applied to make sense of how neurophysiological processes might realize mental models.

The goal of this chapter is to provide readers with a general introduction to the central challenge facing those who would maintain that in-the-head model-based reasoning about mechanisms in science is achieved through the use of scale-model-like mental representations.

A central form of model-based reasoning in science, particularly in the special sciences, is model-based reasoning about mechanisms. This form of reasoning can be affected with the aid of external representational aids (e.g., formalisms, diagrams, and computer simulations) and through the in-the-head manipulation of representations. Philosophers of science have devoted most of their attention to the former, but the latter is arguably at the heart of most of what passes for explanatory understanding in science (Sect. 31.1). Psychologists have long theorized that humans and other creatures (e.g., rats) reason about spatial, kinematic, and dynamic relationships through the use of mental representations, often termed *mental models*, that are structurally sim-

31.1	<b>Mechanistic Reasoning in Science</b> .....	672
31.2	<b>The Psychology of Model-Based Reasoning</b> .....	673
31.3	<b>Mental Models in the Brain: Attempts at Psycho-Neural Reduction</b> .....	675
31.3.1	From Structural to Functional Isomorphism.....	676
31.3.2	Distinctive Features of Scale Models ....	678
31.3.3	Does Computational Realization Entail Sentential Representation?.....	681
31.3.4	What About POPI? .....	682
31.3.5	Bridging the Divide .....	684
31.3.6	Bottom-Up Approaches.....	685
31.4	<b>Realization Story Applied</b> .....	686
31.4.1	AI and Psychology: Towards an Intuitive Physics Engine.....	686
31.4.2	Exduction .....	687
31.5	<b>Mechanistic Explanation Revisited</b> .....	687
31.5.1	The Prediction and Ceteris Paribus Problems .....	688
31.5.2	Beyond Mental Models.....	689
31.6	<b>Conclusion</b> .....	690
	<b>References</b> .....	690

ilar to scale models, though clearly the brain does not instantiate the very properties of a modeled system in the way that scale models do (Sect. 31.2). A key challenge facing this view is thus to show that brains are capable of realizing representations that are like scale models in crucial respects. There have been several failed attempts to show precisely this, but a look at how computers are utilized to model mechanical interactions offers a useful way of understanding how brains might realize mental representations of the relevant sort (Sect. 31.3). This approach meshes well with current research on mental maps in rats. In addition, it has useful ramifications for research in artificial intelligence (AI) and logic (Sect. 31.4), and it offers a promising account

of the generative knowledge that scientists bring to bear when testing mechanistic theories while also shedding

light on the role that external representations of mechanisms play in scientific reasoning (Sect. 31.5).

## 31.1 Mechanistic Reasoning in Science

A common reason that scientists engage in model-based reasoning is to derive information that will enable them to explain or predict the behavior of some target system. Model-based explanations provide scientists with a way of understanding how or why one or more *explanandum* occurrences came about. A good model-based explanation will typically provide the means for determining what else one ought to expect if that explanation is accurate – that is, it will enable one to formulate predictions so that the explanation may (within widely known limits) be tested. (One must bear in mind, however, that models are often accurate only in certain respects and to certain degrees [31.1].)

Model-based reasoning can, corresponding to the diversity of representational structures that count as models – including external scale models, biological models, mathematical formalisms and computer simulations – take many forms in science. As for what models represent, it is now widely accepted that mechanisms are one of the principal targets of model-based reasoning. This is most obviously true in the nonbasic sciences (e.g., biology, medicine, cognitive science, economics, and geology).

In philosophy of science, much of the focus on mechanisms has thus far been on the role they play in scientific explanation. The idea that *all* genuine scientific explanation is mechanistic began to gain traction in contemporary philosophy of science with the work of *Peter Railton*, who claimed that [31.2]

“if the world is a machine – a vast arrangement of nomic connections – then our theory ought to give us some insight into the structure and workings of the mechanism, above and beyond the capability of predicting and controlling its outcomes [...].”

Inspired by Railton, *Wesley Salmon* abandoned his statistical-relevance model of explanation in favor of the view that “the underlying causal mechanisms hold the key to our understanding of the world” [31.3]. In this view, an “explanation of an event involves exhibiting that event as it is embedded in its causal network and/or displaying its internal causal structure” [31.4]. Salmon was working in the shadow of Carl Hempel’s covering law model of explanation, according to which explanations involve inferences from statements describing laws and, in some cases, particular conditions. Salmon tended, in contrast, to favor an ontic account,

according to which explanations are out in the world. He thought that progress in understanding those explanations requires *exhibiting* the relevant mechanisms. However, even though he rejected representational and inferential accounts of explanation, he naturally recognized that reasoning about mechanisms, which requires representations (models), plays a big part in the process of exhibiting those mechanisms.

A more recent formulation of the mechanistic account of explanation is supplied by *Machamer et al.*, who claim that “Mechanisms are entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions” [31.5]. A central goal of science, in their view, is to formulate models, which take the form of descriptions of mechanisms that render target occurrences intelligible [31.5]:

“Mechanism descriptions show *how possibly*, *how plausibly*, or *how actually* things work. Intelligibility arises [...] from an elucidative relation between the *explanans* (the set-up conditions and intermediate entities and activities) and the *explanandum* (the termination condition or the phenomenon to be explained) [...].”

As with *exhibiting* for Salmon, the process of *elucidating* how set-up conditions lead to termination conditions requires a significant contribution from model-based reasoning.

*Bechtel* offers a related account of mechanisms. He claims [31.6]:

“A mechanism is a structure performing a function in virtue of its component parts. The orchestrated functioning of the mechanism is responsible for one or more phenomena.”

As compared with other mechanists, Bechtel is much more explicit about the role that model-based reasoning plays in science and about the diverse forms of representation that may be involved (e.g., descriptions, diagrams, scale models and animal models). He is, moreover, among the few to acknowledge the importance of *in-the-head* model-based reasoning. He suggests that its central form may involve a kind of mental animation. As *Bechtel* and *Wright* put it, “One strategy is to use imagination to put one’s representation of the mechanism into motion so as to visualize how

that phenomenon is generated” [31.7]. Bechtel claims that the representations underlying this mental animation process may have a structure similar to that of the diagrams scientist use in their thinking and to the animated renderings of computer simulations scientists construct to represent proposed mechanisms in action. As for prediction, he notes [31.6]:

“what the scientist advances is a representation of a mechanism [...] She or he then evaluates the representation by using it to reason about how such a mechanism would be expected to behave under a variety of circumstances and testing these expectations against the behavior of the actual mechanism.”

In other words, once the scientist possesses a model of the mechanisms that may be responsible for an occurrence, which may take the form of a *mental* model,

## 31.2 The Psychology of Model-Based Reasoning

Given the potentially crucial role that mental models play in the process of mechanistic explanation and prediction, it may be that we cannot hope to attain a truly adequate, deep understanding of science without first understanding how the mental modeling process works. An obvious way of going about making sense of the role mental models play in science is to inquire into the nature of those models themselves. A good question to ask here is: What form must our mental models take if they are to play the role that they do in science? One increasingly popular answer has its origins in *Craik’s* landmark monograph, *The Nature of Explanation*. Regarding everyday reasoning, *Craik* suggests [31.9]:

“If the organism carries a *small-scale model* of external reality and of its own possible actions within its head, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise [...] and in every way to react in a much fuller, safer, and more competent manner to the emergencies which face it.”

In *Craik’s* view, scientific explanation is just an extension of this everyday reasoning process – that is, it involves the construction of internal world models that are akin to scale models. (*Bechtel* is explicit in crediting *Craik*, when he maintains that the use of mental models in scientific reasoning about mechanisms is to be understood by analogy with the use of external images and scale models [31.6]. Fellow mechanists *Nancy Ners-*

he or she may then use it to formulate predictions in order to test that model.

While external representational artifacts may sometimes be required in order to achieve explanatory understanding of how a mechanism could produce a given phenomenon, plausibly those artifacts are not themselves sufficient for explanatory understanding. (For evidence that there is a crucial psychological component to explanatory understanding, see [31.8].) Instead, representational artifacts may have the important function of facilitating understanding by enhancing the scientist’s ability to mentally simulate the process by which the proposed mechanism would produce the target phenomenon. (As shown in Sect. 31.4.2, external representational aids may also enable forms of reasoning that would otherwise (e.g., due to the complexity of the mechanism) be impossible.) Through manipulation of those mental simulations, scientists may also discover novel predictions of a given model.

*essian* [31.10] and *Paul Thagard* [31.11] also credit *Craik*.)

What may be considered the first attempt to put this view to experimental scrutiny came in the prelude to the cognitive revolution with *Edward Tolman’s* seminal studies of spatial navigation in rats [31.12]. In his most famous experiment, *Tolman’s* team placed rats in a simple alley maze, similar to the one depicted in Fig. 31.1a, and rewarded the animals with food when they reached the end. After learning to perform the task without hesitation, the maze was replaced with a radial maze similar to the one in Fig. 31.1b, where the alley that the rats had previously learned to traverse was blocked. Upon discovering this, the vast preponderance of rats then chose the alley that led most directly to where the food source had been in previous trials. On the basis of such experiments, *Tolman* concluded that rats navigate with the

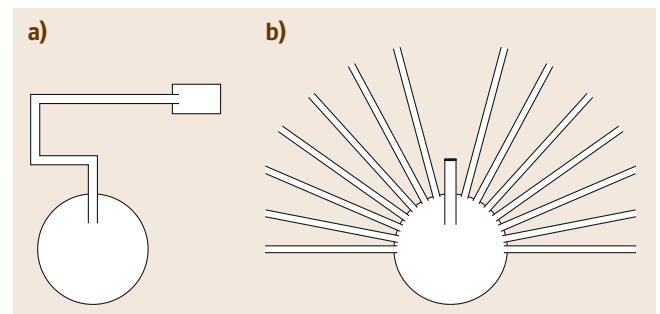


Fig. 31.1a,b Alley maze (a) and radial maze (b) (after [31.12])

aid of cognitive maps of the relative spatial locations of objects in their environment.

Later, *Shepard* and *Metzler* would show that the time it takes for people to determine if two three-dimensional (3-D) structures have the same shape is proportional to the relative degree of rotational displacement of those structures [31.13]. One neat explanation for this finding is that people engage in the mental rotation of 3-D models of the two structures until they are aligned in such a fashion as to enable easier comparison. In another landmark study of mental imagery, *Kosslyn* showed that reaction times for scanning across mental images of a map was proportional to distance, but not to the number of intervening objects, suggesting that spatial reasoning is better explained by a process akin to scanning across a real map than to a process of sentence-based reasoning (e.g., working through a list structure) [31.14].

All of this research points to the existence of mental models of two-dimensional (2-D) and 3-D spatial relationships, but to support the full range of inferences implicated in mechanistic model-based scientific reasoning, mental models would need to capture kinematic and dynamic relations as well. There is some support for the existence of these models as well. For instance, *Schwartz* and *Black* observed similar, proportional reaction times when subjects were asked to determine whether or not a knob on one gear would, when that gear is rotated, fit into a groove on a connecting gear (Fig. 31.2a) [31.15]. *Schwartz* and *Black* found, moreover, that subjects were able to “induce patterns of behavior from the results depicted in their imaginations” [31.16]. Subjects might, for instance, infer and remember that the second in a series of gears will, along with every other even-numbered gear, turn in the opposite direction of the drive gear (Fig. 31.2b). Having inferred this through simulation, the information becomes stored as explicit knowledge, thereby eliminating the need to generate the knowledge anew for each new application.

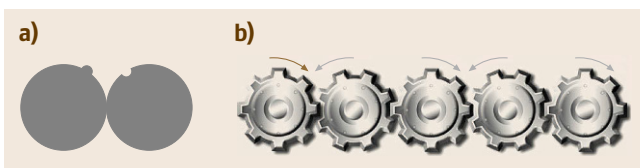
In addition, [31.17, 18] have shown that mental modeling of dynamic relationships is often affected in piecemeal fashion, a process that is much better suited for tracing a sequence of interactions through a system than for simulating collections of dynamic effects all at

once. All of this research fits well with *Norman*'s early assessment of mental models. He notes [31.19]:

- “1. Mental models are incomplete.
2. People’s abilities to *run* their models are severely limited.
3. Mental models are unstable: People forget the details of the system they are using [ . . . ]
4. Mental models do not have firm boundaries: similar devices and operations get confused with one another.”

These limitations on the human ability to construct and manipulate mental models surely have a great deal to do with more general limitations on the capacity of human working memory and with the high cognitive load associated with creating, maintaining, and manipulating mental models.

In everyday reasoning with mental models, the behaviors of the component structures in our models will not typically be tied in any direct way to fundamental physical laws (e.g., Newtonian, quantum mechanical, or relativistic). Rather, many of the kinematic and dynamic principles governing object behavior in our mental simulations will be rooted in early experiences of collisions, impenetrability, balance and support, projectiles, blocking, and so forth [31.20–22]. In addition, in everyday reasoning, and even more so in scientific reasoning about mechanisms, many of the behaviors of the components of our models will not be the result of early learning. Some of these will be one-off *brute* events – such as a meteor striking the earth, a gene mutating, or a latch coming undone – for which one does not have or require (in order to formulate a satisfactory answer to the question of why the *explanandum* occurred) any deeper explanation. Such occurrences might be imposed upon a mental model in much the same way that one would impose them – that is, through direct intervention – on a scale model. In the same way, one could also impose newly learned or hypothesized regularities on a model. Some of these might be discovered through simple induction (one might notice that one’s car engine becomes louder in cold weather) or through prior model-based reasoning (as in *Schwartz*’ study with gears). However, when formulating mechanical explanations, particularly in science, one sometimes simply hypothesizes, as a way of making sense of the available data, that a particular regularity obtains. A good example of this is the way that the hypothesis of periodic geomagnetic pole flipping was used to make sense of the patterns of magnetization in rocks found lateral to mid-ocean rifts [31.1]. Such ideas accord well with recent work regarding mechanistic explanation in the philosophy of science, where it is generally recognized that our models of mechanisms



**Fig. 31.2a,b** Knob and groove on connecting gears (a), (after [31.15]). Gears in series (b), (after [31.16])

typically bottom out at brute *activities* [31.5] or *functions* [31.6].

The above empirical research sheds light on the properties of the models we use to reason about mechanisms in everyday life and in science. There is, in addition, a great deal of research that simply hypothesizes that we do utilize such models to understand other cognitive processes such as language comprehension, concepts [31.23], or learning [31.24–30].

The hypothesis of mental models has also been invoked by *Johnson-Laird* to explain deductive reasoning, though here the term *mental model* is used somewhat differently than it is in the research cited above [31.31]. (Below, I explain in greater depth, the contrast between deductive reasoning more generally and the mental models approach to mechanistic reasoning espoused here.) Like many proponents of mental models, *Johnson-Laird* and *Byrne* do claim to be directly inspired by Craik, an inspiration that shows up in their suggestion that mental models have “a structure that is remote from verbal assertions, but close to the structure of the world as humans conceive it” [31.32]. However, if we look more closely at the way in which *Johnson-Laird* employs the mental models hypothesis in accounting for reasoning processes (deductive, inductive, and abductive), it begins to look as though he has something very different in mind. For instance, with regard to deductive reasoning – that is, reasoning

that mainly involves the semantic properties of top-neutral logical operators such as *if...then...*, *and*, *all*, and *some* – *Johnson-Laird* proposes that we reason internally through a process not unlike the formal method of truth table analysis. For instance, on *Johnson-Laird*’s view, the conditional, *If the door is pushed, then the bucket will fall*, would be mentally represented as something like the following spatial array, which lists those scenarios (models) that would be consistent with the truth of the statement ( $\neg$  here signals negation)

door pushed	bucket falls
$\neg$ door pushed	bucket falls
$\neg$ door pushed	$\neg$ bucket falls

If presented with the additional premise, *The bucket did not fall*, one could then eliminate all but the last of these models, enabling a valid deduction to *The door was not pushed*. The formal, topic-neutral nature of this strategy means that it works in exactly the same way regardless of what items (e.g., balloons, satellites, or mice) we are reasoning about. To say nothing of the viability of the approach, *Johnson-Laird*’s proposals regarding deductive (as well as inductive and abductive) reasoning thus seem, except insofar as they appeal to such structures as *spatial* arrays, at odds with his avowed view that mental models have a structure closer to the world than to our descriptions of it.

### 31.3 Mental Models in the Brain: Attempts at Psycho-Neural Reduction

While there has been considerable research on mental models in recent years, what has been very slow to materialize is a demonstration that brains do or, what is even more worrisome, *that they could* harbor mental models that are like scale models in crucial respects. One can see how this might raise concerns about the mental models hypothesis. After all, if brains cannot realize such models then the above explanatory appeals to mental models come out looking misguided from the outset. At the same time, there is a competing hypothesis which faces no such difficulties. In its most audacious form, it is the proposal that all of cognition is affected through formal computational operations – that is, operations that involve the application of syntax-sensitive inference rules to syntactically structured (sentential) representations.

Proponents of the computational theory of cognition know that they have nothing to fear, at least with regards to the matter of whether or not brains are capable of realizing the relevant kinds of syntax-crunching operations. *McCulloch* and *Pitts* showed,

decades ago, that collections of neuron-like processing units can implement logic gates and, in principle, a universal Turing machine [31.33]. Indeed, it was in no small part because von Neumann recognized the functional similarities between *McCulloch–Pitts* neurons and electronic switches (e.g., transistors) that he was inspired to create the first fully programmable computers, ENIAC and EDVAC. More recently, it has been shown that recurrent neural networks are, memory limitations notwithstanding, capable of implementing computers that are Turing complete [31.34]. There is, then, no longer any doubt that it is possible to bridge the divide between neural machinations and syntax-crunching operations.

In contrast, a satisfactory demonstration that neural machinations might realize mental models – that is, nonsentential mental representations that are like scale models in crucial respects – has proven far more elusive. Indeed, difficulties arise the moment one tries to specify what the *crucial respects* might be, as is evidenced by the fact that each past attempt at doing this

has been argued, not without justification, to run afoul of one or the other of the following two desiderata:

1. An adequate account of mental models must be compatible with basic facts about the brain.
2. An adequate account of mental models must be specific enough to distinguish mental models from other kinds of representation (sentential representations).

Again, this is no small matter, for given that brains are known to be capable of formal computational operations, if it cannot be shown that they are also capable of realizing mental models, this will cast doubt on all those psychological theories mentioned above that advert to mental models. This is a concern made all the more pressing by the fact that proponents of the computational theory of cognition have no shortage of alternative explanations for the behavioral data cited in support of mental models. For instance, to the extent that people report having model-like phenomenology, this might be dismissed as a mere *epiphenomenon* of the actual, underlying computational operations. Similarly, to the extent that behavioral data, such as reaction times, suggests reliance upon model-like mental representations that undergo continuous transformations, this might be chalked up to demand characteristics (subjects may feel compelled to pretend that they are scanning a map). Some of these specific objections could be vulnerable in that they give rise to their own testable predictions [31.35], but, as explained below, proponents of the computational theory have an ace up their sleeve, for computational accounts are flexible enough to handle virtually any behavioral data. All of this is quite general, so let us turn to some of the specific attempts to spell out the distinctive features of mental models.

### 31.3.1 From Structural to Functional Isomorphism

As we have seen, from the outset, the claim made on behalf of putative mental models is that they are *like* scale models in one or more crucial respects. Of course, scale models are themselves *like* the actual systems they represent in a very obvious respect: They instantiate the very same properties as what they represent. It is thus no surprise that the dominant theme in attempts to specify what makes mental models *models* is the invocation of one form or another of isomorphism between mental models, scale models and the modeled world.

#### Mere Isomorphism

The most straightforward form of isomorphism invoked in this literature is what might be termed *bare* isomorphism, or isomorphism *simpliciter*, which is

a purported relationship between mental models and what they represent. Despite initial appearances, this is the form of isomorphism that *Craik* seems to have had in mind. He claims, for instance: “By a model we thus mean any physical or chemical system which has a similar relation-structure to that of the process it imitates” [31.9]. Latter-day proponents of this proposal include *Cummins* [31.36] and *Hegarty*, who, in an attempt to summarize the dominant view of mental models in psychology, notes [31.18]:

“a mental model (or situation model) is a representation that is isomorphic to the physical situation that it represents and the inference processes simulate the physical processes being reasoned about.”

One serious concern about this approach is that it is too liberal, which is to say that it leads one to classify too wide a range of representations as *models*. Consider, for instance, that one of *Craik*’s favored examples of a representation with a *similar relation structure* to what it represents is Kelvin’s Tide Predictor, a device that consists of an ingenious system of gears and pulleys arranged so as to support truth-preserving inferences regarding the tides (Fig. 31.3). Says *Craik* [31.9],

“My hypothesis then is that thought models, or parallels, reality—that its essential feature is [...] symbolism, and that this symbolism is largely of the



**Fig. 31.3** Kelvin’s first tide predicting device (photo by William M. Connoley)

same kind as that which is familiar to us in mechanical devices which aid thought and calculation.”

This, of course, is no different from what proponents of the computational theory of cognition currently maintain. After all, any syntax-crunching system capable of supporting truth-preserving inferences with respect to a given physical system will have to be isomorphic with it – that is, there will have to be correspondences between the parts and relations in the system and the components of the representation – in ways that get preserved over the course of computation. To that extent, one might even say that the inference process *simulates*, or even *pictures* [31.37], the process being reasoned about. In short, then, the proposal that mental models are merely isomorphic with what they represent is thus far too vague to satisfy desideratum (2.) above. Indeed, it is for this very reason that researchers have tried to find a more restrictive notion of isomorphism, one that can distinguish models from sentential representations.

#### Physical Isomorphism

Perhaps the most restrictive such notion is that of *structural* [31.38] or *physical* [31.39] isomorphism, which involves instantiating the very same properties, and arrangements thereof, as the represented system. This appears to be the kind of isomorphism that *Thagard* has in mind when he claims [31.11] (also see [31.40]):

“Demonstrating that neural representation can constitute mental models requires showing how they can have the same relational structure as what they represent, both statically and dynamically.”

Thagard cites Kosslyn’s research as indicative of how this demand might be met, and in Kosslyn too, we do find frequent appeals to structural isomorphisms. For instance, noting the retinotopic organization of areas of visual cortex that are implicated in mental imagery, *Kosslyn* claims, “these areas represent depictively in the most literal sense [...]” [31.41].

Unfortunately, the postulation of physically isomorphic mental representations is highly suspect for several reasons. To start with, the kind of retinotopy that one finds in areas such as V1 is highly distorted relative to the world due to the disproportionate amount of cortex devoted to the central portion of the retina (the fovea). A square in the visual field is thus not represented in the cortex by sets of neurons that lie in straight, let alone in parallel, lines. Moreover, visual representation seems not to be carried out through the activity of any single retinotopically organized neural ensemble. Rather, vision involves the combined activity of a variety of systems that are, to a considerable extent, anatomically

and functionally distinct [31.42–44]. Lastly, the kind of retinotopy pointed out by Kosslyn is restricted to two spatial dimensions, and a 2-D representational medium cannot realize representations that are physically isomorphic with what they represent in three dimensions. Nor, a fortiori, can such a medium realize representations that are physically isomorphic in both 3-D and causal respects. Crudely put, there are no literal buckets, balls, or doors in the brain. (Perhaps it is worth noting, as well, how inessential structural isomorphism is to information processing in neural networks, even in the case of 2-D retinotopic maps. The relative physical locations of neural cell bodies seems irrelevant when compared to the patterns of connectivity between neurons, the strengths and valences of connections, and the schemes of temporal coding the neurons employ. One would expect then that, so long as all of this is preserved, cell bodies might be tangled up in arbitrary ways without affecting processing.)

#### Functional Isomorphism

The main problem with the appeal to physical isomorphism, one that has long been appreciated, is that it fails to satisfy desideratum (1.). As *Shepard* and *Chipman* note, “With about as much logic, one might as well argue that the neurons that signal that the square is green should themselves be green!” [31.38]. Recognizing this, and recognizing the weakness of appeals to mere isomorphism, *Shepard* and *Chipman* push for the following moderate notion of isomorphism [31.38, italics added for emphasis]:

“isomorphism should be sought-not in the first-order relation between (a) an individual object, and (b) its corresponding internal representation-but in the second-order relation between (a) the relations among alternative external objects, and (b) the relations among their corresponding internal representations. Thus, although the internal representation for a square need not itself be square, it should [...] at least have a *closer functional relation* to the internal representation for a rectangle than to that, say, for a green flash or the taste of persimmon.”

The appeal to second-order isomorphism would, they hoped, provide an alternative to physical isomorphism that is both consistent with basic brain facts (desideratum (1.)) and distinct from sentential accounts (desideratum (2.)).

Another moderate account of isomorphism was put forward at the same time by *Huttenlocher* et al. [31.45]. They had a particular interest in how subjects make ordering inferences (viz., those involving the ordering of three items along such dimensions as size, weight and



height) like this one

Linus is taller than Prior.  
 Prior is taller than Mabel.  
 ∴ Linus is taller than Mabel.

*Huttenlocher* et al. suggested that subjects might use representations that “are isomorphic with the physically realized representations they use in solving analogous problems (graphs, maps, etc.) [...]” [31.45]. The essence of their proposal was that the mental representations that subjects form in order to solve such problems might function like spatial arrays rather than like sentences. For instance, what seems distinctive about *external* sentential representations of three-term ordering syllogisms like the one above is that, because each premise is represented in terms of a distinct expression, terms that denote particular individuals must be repeated. On the other hand, when such inferences are made with the aid of external spatial arrays, the terms need not be repeated. For instance, one can make inferences about the taller-than relation on the basis of the left-of relation with the help of marks on a paper like these

L P M

In fact, the introspective reports obtained by *Huttenlocher* et al. did support the idea that subjects were constructing the functional equivalents of spatial arrays – for instance, subjects reported that symbols representing individuals were not repeated – and on this basis they claimed that subjects might be carrying out three-term ordering inferences using mental representations that *function* like actual spatial arrays and unlike lists of sentences (see also [31.40]). This kind of isomorphism is thus sometimes termed *functional* isomorphism [31.39].

*Shepard* and *Chipman* [31.38] and *Huttenlocher* et al. [31.45] were clearly after a notion of isomorphism that satisfies desideratum (1.). Unfortunately, the solutions they offer appears, at least at first glance, to run afoul of desideratum (2.) – that is, appeals to functional isomorphism, of either the first or second-order variety, seem not to distinguish between computational representations and model-like representations. *Huttenlocher* et al. were among the first to suspect this. They note [31.45]:

“It is not obvious at present whether any theory which postulates imagery as a mechanism for solving problems can or cannot, in general, be reformulated in an abstract logical fashion that, nevertheless makes the same behavioral predictions.”

*Anderson* is generally credited with confirming this suspicion by pointing out the possible tradeoffs that can be made between assumptions about representational structure and those concerning the processes that operate over the representations [31.46]. He showed that the possible structure-process tradeoffs render computational accounts flexible enough to handle virtually any behavioral finding. Most have since endorsed his thesis that it is, at least after the fact, always possible to “generate a propositional (i. e., sentential) model to mimic an imaginal model” [31.46]. Alternatively, as *Palmer* puts it, if you create the right sentential model it will be functionally isomorphic to what it represents in just the sense that a nonsentential model is supposed to be [31.39].

### Imagery and Perception

One last way in which one might try to satisfy the above desiderata, at least with regard to spatial models, is to point out that visual mental imagery involves the utilization of visual processing resources. *Brooks* [31.47] and *Segal* and *Fusella* [31.48], for instance, discovered that performance on visual imagery tasks is diminished when subjects must perform a concurrent visual processing task but not when they perform an auditory task – that is, they found that there is interference between mental imagery and auditory perception but not between mental imagery and visual perception (see also [31.36]). However, if these findings are meant to provide a model-based alternative to computational theories, the attempt would appear to have the same fundamental flaw as the appeal to functional isomorphism. As *Block* notes, because perceptual processing can, in principle, also be explained in terms of computational processes [31.49] (see also [31.46, 50]):

“the claim that the representations of imagery and perception are of the same kind is irrelevant to the controversy over pictorialist versus descriptionalist interpretation of experiments like the image scanning and rotation experiments [...]”

That is, the claim that imagery utilizes visual processing resources fails to satisfy desideratum (2.).

### 31.3.2 Distinctive Features of Scale Models

The overall realization problem facing putative mental models, then, is just that it has proven exceedingly difficult to specify what sorts of representational structures mental models are in a way that is consistent with basic brain facts but that also distinguishes models from sentential representations. In order to finally see our way past these concerns, it will be helpful if we first take stock of a handful of features that are widely taken, even

by proponents of the computational theory of mind, to distinguish *external* images and scale models from sentential representations. Three such features concern the sorts of entities, properties, and processes that each form of representation is naturally suited for representing:

1. Images and scale models are not naturally suited for representing abstract entities, properties, and processes (e.g., war criminal, ownership, or economic inflation). They are much better suited for representing concrete entities, properties, and processes (e.g., a bucket, one object being over another, or compression).
2. Images and scale models are not naturally suited for representing general categories (e.g., triangles or automobiles). They are better suited for representing specific instances of categories (Note: Genera differ from abstracta in that the former can be concrete (e.g., rocks) and the latter can be specific (e.g., the enlightenment)).
3. Images and scale models are not naturally suited for singling out specific properties of specific objects [31.37, 50]. For instance, it would be difficult, using a scale model, to represent just the fact that Fred's car is green, for any such model will simultaneously represent many other properties, such as the number of doors and wheels, the body type, and so on.

In contrast, sentential representations (those constructed using natural and artificial languages) have little trouble representing abstracta (e.g., *war criminal*), genera (*triangle*), and specific properties of specific objects (e.g., *Fred's car is green*).

While images and scale models are relatively disadvantaged in the above respects, they are much better suited for supporting inferences regarding the consequences of alterations to specific, concrete systems. The fact that syntax-crunching systems are quite limited in this regard first came to light as a consequence of early work in formal-logic-inspired, sentence-and-rule-based AI. The general problem confronting syntax-crunching approaches came to be known as the *frame problem* [31.51].

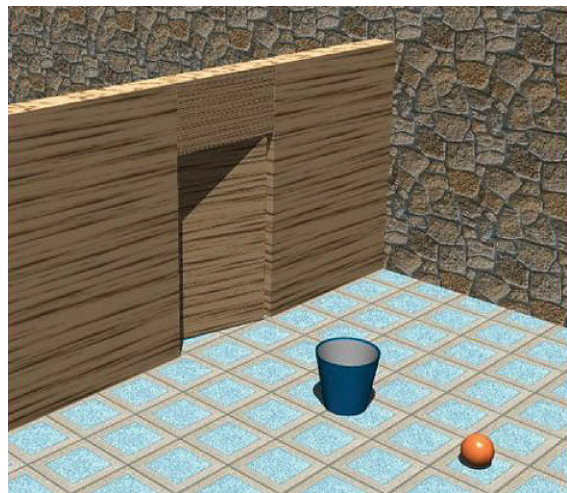
In its original formulation, the frame problem had much to do with the challenge of endowing a sentence-and-rule-based representational system with the ability to anticipate what will *not* change following an alteration to the world (e.g., tipping over a bottle changes its orientation but *not* its color). Today, however, the frame problem is regarded as something more general – namely, the problem of endowing computational systems (and other artifacts) with the kind of com-

monsense knowledge that the average human possesses about what will change and what will stay the same following alterations to the objects in the world. As Hayes puts it [31.52]:

“The frame problem arises in attempts to formalise problem-solving processes involving interactions with a complex world. It concerns the difficulty of keeping track of the consequences of the performance of an action in, or more generally of the making of some alteration to, a representation of the world.”

The frame problem can actually be broken down into at least two component problems, the prediction problem [31.53] and the qualification problem [31.54].

As it confronts computational devices, the prediction problem can be summed up as follows: In order to support inferences about the consequences of alterations to even simple physical systems, a sentence-and-rule system would have to contain innumerable rules that explicitly specify how objects will behave relative to one another following each of innumerable possible alterations. For a simple illustration, consider what we all know about the consequences of different ways of altering the items in Fig. 31.4. We know, for example, what would happen were we to use the bucket to throw the ball through the open doorway, were we to place the bucket over the ball and slide the bucket through the doorway, were we to set the bucket containing the ball atop the slightly ajar door and then shove the door open, and so on indefinitely. To endow a sentence-and-rule system with the ability to predict the consequences of these various alterations, one would have to build in, corresponding to each one, a separate data structure



**Fig. 31.4** A toy world: A doorway, a bucket, and a ball (after [31.55])

specifying the starting conditions, the alteration, and the consequences of that alteration. If these take the form of conditional statements, the system could then make inferences utilizing domain-general (e.g., topic-neutral, deductive) machinery. Alternatively, the information could be encoded directly as domain-specific inference rules (e.g., production-system operators). Either way, from an engineering standpoint, the problem that quickly arises is that no matter how many of these statements or rules one builds into the knowledge base of the system, there will generally be countless other bits of commonsense knowledge that one has overlooked. Notice, moreover, that scaling the scenario up even slightly (e.g., such that it now includes a board) has an *exponential* effect on the number of potential alterations and, as such, on the number of new data structures that one would have to incorporate into one's model [31.53]. As Hayes says [31.52]:

“One does not want to be obliged to give a law of motion for every aspect of the new situation [...] especially as the number of frame axioms increases rapidly with the complexity of the problem.”

Moreover, as explained in the manual for a past incarnation of the production system Soar [31.56]:

“when working on large (realistic) problems, the number of operators (i.e., domain-specific rules) that may be used in problem solving and the number of possible state descriptions will be very large and probably infinite.”

As if the prediction problem were not problem enough, it is actually compounded by the other facet of the frame problem, the qualification problem [31.54]. This is because in order to capture what the average human knows about the consequences of alterations to a physical system, not only would innumerable distinct conditionals or inference rules be required, but each would have to be qualified in an indefinite number of ways. Notice, for instance, that placing the bucket over the ball and sliding it through the doorway will result in the ball being transferred to the other side of the wall, but *only if it is not the case that* there is a hole in the floor into which the ball might fall, there is a hole in the bucket through which it might escape, the ball is fastened securely to the floor, and so on indefinitely. To once again quote Hayes, “Almost any general belief about the result of his own actions may be contradicted by the robot's observations [...] there are no end to the different things that can go wrong, and he cannot be expected to hedge his conclusions round with thousands of qualifications” [31.52]. Thus, to capture what the average human knows, if only implicitly, about the consequences of this one alteration, all of the relevant

qualifications would have to be added to the relevant sentence or rule. Once again, in realistic situations, the challenge of specifying all of the qualifications is magnified exponentially.

The general failing of sentence-and-rule-based representations that the frame problem brings to light is that they only support predictions concerning the consequences of alterations and the defeaters of those consequences if those alterations, consequences, and defeaters have been spelled out, antecedently and explicitly, as distinct data structures. Representations of this sort – that is, representations that require distinct structures to support predictions regarding the consequences of each type of alteration to the represented system – are sometimes termed *extrinsic* representations. (The intrinsic-extrinsic distinction discussed here was introduced by Palmer [31.39] but modified by Waskan [31.57, 58].)

It is worth a quick digression to note that, while the terminology has changed, these general concerns about the limitations of extrinsic representations antedate work in contemporary AI by over three hundred years. They show up, for instance, in Descartes' *best-explanation* arguments for dualism in his *Discourse on the Method*. Descartes there despairs of there ever being a mechanical explanation for, or an artifact that can duplicate, the average human's boundless knowledge of the consequences of interventions on the world [31.59]:

“If there were machines which bore a resemblance to our bodies and imitated our actions [...] we should still have two very certain means of recognizing that they were not real men [...] (Firstly, humans have the ability to converse.) Secondly, even though some machines might do some things as well as we do them [...] they would inevitably fail in others, which would reveal that they are acting not from understanding, but only from the disposition of their organs. For whereas reason is a universal instrument, which can be used in all kinds of situations, these organs need some particular action; hence it is for all practical purposes impossible for a machine to have enough different organs to make it act in all the contingencies of life in the way in which our reason makes us act.”

Descartes thought that to match wits with even a *dull-witted* human, any natural or artificial device would need, *per impossibile*, to rely upon an infinite number of specific sensory-motor routines – which bear a striking resemblance to production-system operators – for each new situation the device might confront. What Descartes could not imagine, because he thought that all such knowledge had to be represented explic-

itly, was the possibility of (to use Chomsky's term) a *generative* inference mechanism – that is, one that embodies boundless knowledge of implications through finite means.

What Descartes failed to notice was that there were already artifacts (i. e., scale models) that exhibited the requisite generativity. Indeed, in contemporary AI, the benefits of an appeal to scale-model-like representations are now well known. Starting with the prediction problem, one can use a reasonably faithful scale model of the setup depicted in Fig. 31.4 in order to predict what would happen were one to use the bucket to throw the ball through the open doorway, were one to place the bucket over the ball and slide the bucket through the doorway, were one to set the bucket containing the ball atop the slightly ajar door and then shove the door open, and so on indefinitely. To use *Haugeland's* terms, the side effects of alterations to such representations mirror the side effects of alterations to the represented system *automatically* [31.60] – which is to say, without requiring their explicit specification. (This only holds, of course, to the extent that the model is a faithful reproduction. Unless the model is a perfect replica, which includes being *to* scale, there will be some limits on inferential fidelity, though this does not undermine the claim that scale models are generative.) Notice also that incremental additions to the represented system will only have an incremental effect on what needs to be built into the representation. The addition of a board to the system above, for instance, can be handled by the simple addition of a scale model of the board to the representation.

Nor do scale models suffer from the qualification problem. To see why, notice that much of what is true of a modeled domain will be true of a scale model of that domain. For instance, with regards to a scale model of the setup in Fig. 31.4, it is true that the scale model of the ball will fall out of the scale model of the bucket when it is tipped over, but only if the ball is not wedged into the bucket, there is no glue in the bucket, and so on indefinitely. Just like our own predictions, the predictions generated using scale models are implicitly qualified in an open-ended number of ways. With scale models, all of the relevant information is *implicit* in the models and so there is no need to represent it all explicitly using innumerable distinct data structures. Representations of this sort are termed *intrinsic* representations. Summing up, scale models are immune to the frame problem, for one can use them to determine, on an as-needed basis, both the consequences of countless alterations to the modeled system and the countless possible defeaters of those consequences – that is, one simply manipulates the model in the relevant ways and reads off the consequences.

### 31.3.3 Does Computational Realization Entail Sentential Representation?

The above distinguishing features can help us to know better whether we are dealing with model-like or sentence-like representations and, ultimately, to appreciate how one might bridge the gap from neurophysiology to mental models. As noted above, a similar bridge was constructed from neurophysiology to computational processes by showing that artifacts (e.g., collections of McCulloch–Pitts neurons or wires and transistors) characterized by a complex circuitry not unlike that of real brains can be configured so as to implement, at a higher level of abstraction, processes that exhibit the hallmarks of traditional syntax-crunching. Because neurons have similar information-processing capabilities as these artifacts, implementing a set of formal operations on an electronic computer is already very nearly an existence proof that brain-like systems can realize the same set of operations.

Might this strategy offer a template for constructing a similar bridge to high-level models? There is surely no shortage of computer simulations of mechanical systems, and at least as they are depicted on a computer's display, these simulations look for all the world like images and scale models. Many would argue, however, that this approach to bridging the neuron-model divide is a nonstarter. The worry, in short, is that it fails to satisfy desideratum (2.) above. To see why, it will be helpful to look at the kinds of computational models of mental imagery offered up by researchers such as *Kosslyn* [31.14] and *Glasgow* and *Papadias* [31.61].

*Kosslyn's* model of mental imagery has several components [31.14]. One is a long-term store that contains sentential representations of the shape and orientation of objects. These descriptive representations are utilized for the construction of representations in another component, the visual buffer, which encodes the same information in terms of the filled and empty cells of a computation matrix. The cells of the matrix are indexed by  $x$ ,  $y$  coordinates, and the descriptions in long-term memory take the form of polar coordinate specifications (i. e., specifications of the angle and distance from a point of origin) of the locations of filled cells. Control processes operate over the coordinate specifications in order to perform such functions as panning in and out, scanning across, and mental rotation.

One distinctive feature of actual (e.g., paper-and-ink) spatial matrix representations is that they embody some of the very same properties and relationships (spatial ones) as – which is just to say that they are physically isomorphic with – the things they represent. But *Kosslyn's* computational matrix representa-

tions (CMRs) are clearly not physically isomorphic with what they represent. After all, Kosslyn's visual buffer representations are not *real* matrix representations that utilize cells arranged in Euclidean space; they are *computational* matrix representations. To be sure, modelers may sometimes see literal pictures on the output displays of their computers, but the representations of interest are located in the central processing unit (CPU) (viz., in random-access memory (RAM)) of the computer running the model. Accordingly, the control operations responsible for executing representational transformations like rotation do not make use of inherent spatial constraints, but rather they operate over the coordinate specifications that are stored in the computer's memory. Details aside, at a certain level of description, there can be no doubt that the computer is implementing a set of syntax-sensitive rules for manipulating syntactically structured representations; this is what computers *do*. As Block puts it, "Once we see what the computer does, we realize that the representation of the line is *descriptive*" [31.49]. The received view, then, a view that has gone nearly unchallenged, is that if a representation of spatial, kinematic, or dynamic properties is implemented using a high-level computer program, then the resulting representations must be *sentential* in character [31.49, 62, 63]. (That Fodor shares this sentiment is suggested by his claim that "if [...] you propose to co-opt Turing's account of the nature of computation for use in a cognitive psychology of thought, you will have to assume that *thoughts themselves have syntactic structure*" [31.64]).

It would thus seem that the strongest claim that can possibly be supported with regard to CMRs is that they *function* like images. Yet, as Anderson notes, it is always possible, through clever structure-process tradeoffs, to create a *sentential* system that mimics an *imagistic* one [31.46]. Indeed, rather than supporting the mental models framework, one might well take computer simulations of mental modeling as concrete evidence for Anderson's claim. Likewise, there is a case to be made that CMRs and their brethren are, unlike scale models, *extrinsic* representations [31.62]. After all, the computers that run them implement syntax-sensitive rules that provide explicit specifications of the consequences of alterations. This is no small matter. From the standpoint of cognitive science, one of the most important virtues of the hypothesis that we utilize mental representations akin to scale models was that scale models constitute *intrinsic* representations of interacting worldly constraints and are thus immune to the frame problem. One could, then, be forgiven for thinking that any attempt to build a bridge from neurons to models by following the template set by computational theories – that is, by noting that certain computational

artifacts instantiate the relevant kind of processing – will be doomed to fail from the outset.

### 31.3.4 What About POPI?

Consider, however, that upon gazing directly at a vast collection of electrical or electrochemical circuits, one will see no evidence of the harboring or manipulation of *sentential* representations. In *Monadology*, Leibniz turned an analogous observation about perceptual experience into an objection to materialism [31.65]:

"It must be confessed, moreover, that perception, and that which depends on it, are inexplicable by mechanical causes, that is, by figures and motions. And, supposing that there were a mechanism so constructed as to think, feel and have perception, we might enter it as into a mill. And this granted, we should only find on visiting it, pieces which push one against another, but never anything by which to explain a perception."

A similar objection might be leveled regarding computational processes. Again, one sees no evidence of this kind of processing when one looks at electronic or electrochemical circuitry. Clearly something has gone wrong.

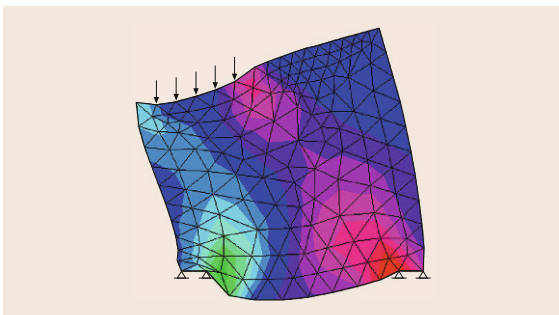
What Leibniz overlooked – and this may be because he lacked the conceptual tools made available by the information age – was a grasp of the principle of *property independence* (POPI). The basic idea of POPI is that properties characterizing a system when it is studied at a relatively low level of abstraction are often absent when it is studied at a higher level, and vice versa. It is POPI that allows computer scientists to say that a system which is characterized by electronic switches and relays at level  $n$  may nevertheless be best described in terms of the storing of bits of information in numerically addressable memory registers at level  $n + 1$  and in terms of the application of syntax-sensitive rules to syntactically structured representations at level  $n + 2$ . It is also the very thing that enables proponents of computational theories of cognition to say that brains and computational artifacts are, despite superficial appearances, capable of implementing the application of syntax-sensitive rules to syntactically structured representations.

However, when proponents of computational theories of cognition insist that computational implementation (e.g., of CMRs) entails *sentential* representation, they are turning their backs on the very principle that enabled them to bridge divide between low-level circuitry and high-level computational operations; they are turning their back on POPI. Indeed, nothing about POPI entails that all syntax-crunching systems must be char-

acterized in terms of sentences and inference rules at the *highest* level of abstraction. POPI thus opens up at least logical space for systems that engage in syntax-crunching operations at one level but that harbor and manipulate nonsentential models at a higher level.

In point of fact, in this logical space reside actual systems, including finite element models (FEMs). These were first developed in the physical (e.g., civil and mechanical) engineering disciplines for testing designs, but they have since become a staple tool in the sciences for exploring the ramifications of theories, generating novel predictions, and facilitating understanding. For our current purposes, what matters most about FEMs is that they provide an existence proof that computational processes can realize nonsentential representations that are like scale models and unlike sentential representations in all of the crucial respects listed above.

To see why, notice first that there are (among others) two important levels of abstraction at which a given FEM may be understood. As with scale models, one may understand FEMs at the relatively low level of the principles that govern their implementing medium. What one finds at this level are sentential specifications of coordinates (e.g., for polygon vertices) along with rules, akin to the fundamental laws of nature, which constrain how those coordinates may change (e.g., due to collisions and loads) (Fig. 31.5). (For a close analogy, think the basic rules of Conway's Game of Life.) When a given model is *run*, at this low level one finds a massive number of iterative number crunching operations. Not unlike Leibniz, enemies of the idea of computationally realized nonsentential models have seized upon this low level with their suggestion that computational systems harbor only sentential representations. At this level, however, it is not even obvious that we are dealing with representations (worldly objects and properties) at all, any more than we are, for instance, when we fixate upon the constraints governing the behaviors of individual Lego blocks.



**Fig. 31.5** Polymesh representation of a blunt impact to a semirigid sheet of material (after [31.57])

One only finds representations of objects when one turns to the higher level of the models that are realized, *and multiply realizable*, by the aforementioned modeling media. And when we take a close look at the properties of these high-level FEMs, we find that they share several characteristics that have long been taken, including by those who suggest that computational implementation entails sentential representation, to distinguish sentential representations from scale models.

To start with, like scale models and unlike sentential representations, FEMs are not (by themselves) naturally suited to representing abstract entities, properties, and processes (e.g., war criminal, ownership, economic inflation). They are much better suited for representing concrete entities, properties, and processes (e.g., a bucket, one object being over another, and compression). Nor are FEMs naturally suited to representing general categories (e.g., triangles or automobiles). They are far better suited for representing specific instances of those categories. Lastly, FEMs are not naturally suited to singling out specific properties of specific objects. For instance, using an FEM, it would be difficult to represent just the fact that Fred's car is green, for any such model will simultaneously represent many other properties, such as the number of doors and wheels, the body type, and so on. In short, just like scale models, FEMs are always representations of specific, concrete systems. By these reasonable standards, FEMs ought to be considered computationally-*realized* nonsentential models that are the close kin of scale models.

The case for this claim looks even stronger once we consider whether or not FEMs constitute intrinsic representations. As we have seen, the received view is that FEMs and their brethren (e.g., CMRs) are extrinsic representations, for the constraints governing how the coordinates of primitive modeling elements may change must be encoded antecedently and explicitly. Indeed, at the level of coordinates and transformation rules, one gets nothing *for free*. However, once a modeling medium has been used to construct a suitable FEM of a collection of objects, the model can then be altered in any of countless ways in order to determine the possible consequences of the corresponding alterations to the represented objects. One can, for instance, use a high-fidelity FEM of the door, bucket, ball system to infer, among other things, what would happen were we to place the bucket over the ball and slide the bucket through the doorway, what would happen were the bucket used to throw the ball at the open doorway, what would happen were the air pressure dramatically decreased, and so on indefinitely [31.57]. The consequences of *these* alterations need not be anticipated or explicitly incorporated into the system. Indeed, as with scale models, much of the point of constructing FEMs

is to *find out* how a system will behave in light of whichever alterations an engineer or scientist can dream up.

It bears repeating that it is not at the level of the primitive operations of an implementation base that we find intrinsic representations, but at the level of the representations *realized by* a given, primitively constrained implementation base. Part of what justifies this claim is the fact that certain constraints will be inviolable *at the level of the model*, and thus a great deal of information will be implicit in the model, because it has been implemented using a particular kind of medium. As *Pylyshyn* notes [31.63]:

“the greater number of formal properties built into a notation in advance, the weaker the notational system’s expressive power (though the system may be more efficient for cases to which it is applicable). This follows from the possibility that the system may no longer be capable of expressing certain states of affairs that violate assumptions built into the notation. For example, if Euclidean assumptions are built into a notation, the notation cannot be used to describe non-Euclidean properties [ . . . ]”

This, in fact, is very close to an apt characterization of what is going on in the case of FEMs. *Given* that a particular model has been realized through the use of a primitively constrained medium, certain constraints will be inviolable at the representational level and a great deal of information will be implicit [31.57]. As Mark Bickhard (in correspondence) summarizes the point:

“Properties and regularities are only going to be *intrinsic* at one level of description if they are built-in in the realizing level – or else they are ontologically *built-in* as in the case of strictly spatial relationships in physical scale models.”

While scale models are intrinsic for the latter reason, FEMs are intrinsic for the former. This shows up in the fact that FEMs exhibit a comparable degree of generativity to scale models and a correlative immunity to the frame problem. Like scale models, FEMs provide a finite embodiment of boundless tacit knowledge, which can be made explicit at any time, of the consequences of innumerable alterations to the systems they represent.

So how does all of this square with *Ander-son’s* [31.46] contention that it is always possible to construct a sentential system to mimic an imagistic or model-based system or *Palmer’s* [31.39] claim that if you create the right sentential model it will be functionally isomorphic to what it represents in just the sense that a nonsentential model is supposed to be? Ander-

son and Palmer are surely right that, post hoc, one can always constrain a sentential representational system so that it mimics the output of a model-based system, but the post hoc character of the strategy is precisely what gets sentential approaches into trouble vis-à-vis the frame problem. Consider, for instance, that the traditional AI approach is to take any physical implication of which humans express knowledge and, after the fact, to build it into the knowledge base of one’s system as a sentence or inference rule. (Despite its shortcomings, this strategy is alive and well, as is evidenced by Lenat’s massive ongoing Cyc project.) But to solve, or rather to avoid, the frame problem, one must rely upon representations that embody all of this boundless information as tacit knowledge – that is, the information cannot be explicitly encoded at the outset, but it can later be generated, and thereby become explicit knowledge, on an as-needed basis. Put simply, to exhibit anything approaching *true* functional isomorphism with scale models, what is needed are high-level, intrinsic, nonsentential models.

To sum up, those who would contend that FEMs (or even CMRs of 2-D spatial properties) are, *qua* computational, necessarily extrinsic and sentential have overlooked the fact that there are multiple levels of abstraction at which a given computational model can be understood. At the relatively low level of the modeling medium, there are unquestionably extrinsic representations of the principles governing the permissible transformation of primitive modeling elements. At a higher level, one finds models that share many distinguishing features, including immunity to the frame problem, with the scale models they were in large part invented to replace. Thus, we find once again that FEMs are like scale models and unlike paradigmatic sentential representations.

### 31.3.5 Bridging the Divide

All of this bears directly on the longstanding concern that there is no way to bridge the divide between neural machinations and the nonsentential models hypothesized by proponents of mental models. What the foregoing makes clear is that computational processes can realize nonsentential models that share with scale models the main characteristics that distinguish nonsentential models from sentential representations. Given that the brain is capable, at least in principle, of realizing any such computational processes, then one must also agree that brains can realize nonsentential models. Thus, by appealing to the above distinguishing features of scale models, we see that there is an account of mental models that (i) distinguishes them (on multiple grounds) from sentential representations and (ii) is

compatible with basic facts about how brains operate. All of this provides a much-needed foundation for all of that psychological work cited above that adverts to mental models.

One advantage of showing that low-level *computations* can realize higher-level mental models is that it renders the mental models hypothesis robust enough to withstand the discovery that the brain is a computational system at some level of description. Even if the brain is not a computational system (i. e., in the syntax-crunching sense), the manner in which computational systems realize intrinsic, nonsentential models will nevertheless remain quite instructive. It suggests a general recipe for the creation of intrinsic models that can be followed even without the computational intermediary: Start by creating a representational medium such that a large number of primitive elements are constrained to obey a handful of simple behavioral principles. Next construct models from this highly productive medium. (*Productive* is here used in *Fodor's* sense – that is, to denote a medium capable of representing an open-ended number of distinct states of affairs [31.66].) What emerges are generative structures capable of supporting an open-ended number of mechanical inferences. At the level of the medium, *running* such a model involves the recursive application of the basic constraints on the modeling-element behaviors. This will typically be a massive, parallel, constraint-satisfaction process. Given that this form of processing is the forte of neural networks, there should be little doubt that neural machinations are up to the task. (At the same time, one should not overestimate inherent immunity to the frame problem of neural networks [31.58]. It is only by implementing a primitively constrained modeling medium that neural networks can be expected to realize intrinsic representations of complex, interacting worldly constraints.)

### 31.3.6 Bottom-Up Approaches

Thus far, we have largely approached the question of the neural realizability of mental models in the abstract, and from the top down. This is partly because there has been a relative dearth of work that moves in the opposite direction, from the bottom up. One exception is *Thagard's* [31.11] recent work on the topic, which appeals to such biologically plausible simulations of neural networks as those of *Eliasmith* and *Anderson* [31.67]. Unfortunately, Thagard has yet to offer evidence that the neural encoding strategies he discusses exhibit any of the central features, discussed here, that distinguish modeling from syntax-crunching. Most notably, the neural representations he cites have not yet been shown to exhibit a significant degree of

spatial, kinematic, or causal generativity. The proof of the pudding here is in the eating.

To the extent that there have been significant advances in the bottom-up endeavor, they mostly issue from research – such as that of Nobel laureates John O'Keefe, May-Britt Moser, and Edward Moser – on the biological neural networks that underwrite spatial reasoning abilities in rats. As you will recall, Tolman's pioneering work on maze navigation suggested that rats have an onboard medium for the construction of generative spatial maps of their location relative to barriers and important items such as food and drink. *O'Keefe* and *Nadel* are famous for showing that the rat's hippocampus contains *place* cells which fire preferentially when an animal reaches a particular location in its environment, cells that fire in sequence as a rat moves from one location to another [31.68]. *Moser* and *Moser* subsequently showed that the rat's uncanny spatial navigation abilities also depend upon *grid* cells in the nearby entorhinal cortex [31.69]. Individual grid cells fire when an animal is in any of several, roughly evenly spaced locations. When lines are drawn to connect these points, they collectively form what (purely by coincidence) looks a great deal like the kind of 2-D polymesh shown in Fig. 31.5. While each grid cell is tuned to a collection of locations, different grid cells have sparser or denser coverage of the same region of space. Collectively they provide effective coverage of the entire region of space in which the animal finds itself.

Importantly, *O'Keefe* et al. note regarding place cells that [31.70]

“there does not appear to be any obvious topographical relation between the field locations (i. e., the places to which cells become temporarily tuned) and the anatomical locations of the cells relative to each other within the hippocampus.”

Nor do grid cells in the entorhinal cortex exploit any obvious structural isomorphisms between their respective anatomical locations and the spatial layout of the environment. However, acting in concert, the two types of cells enable effective navigation, as if the organism had an internal map that preserves relative locations (place cells) and distances (grid cells). In other words, the two systems encode maps that are *functionally* isomorphic with real maps of the environment. Moreover, they provide a productive modeling medium, one which, not unlike a collection of Lego blocks, can be used and reused, through a process called *remapping*, to encode information about an open-ended number of new environments [31.71]. The maps constructed in this medium are generative with regards to 2-D spatial properties in the aforementioned sense, as is shown by their role in enabling rats to find efficient new



ways to a destination when familiar routes are blocked. More recent research suggests that the rat's place cells are also somewhat sensitive to vertical displacement from a reference plane, perhaps enabling 3-D mapping capabilities [31.72]. Nor are the lessons learned here applicable only to rats, for a large body of research suggests that the same anatomical systems may be implicated in human spatial navigation [31.73].

Our deepest understanding of how real neural networks create spatial mental models thus suggests that brains implement a reusable modeling medium and, by exploiting the kinds functional, rather than physical, isomorphisms that make neural realizability feasible, nothing is lost in the way of generativity. It also bears mentioning that this modeling medium is well suited for producing models of the organism's location rela-

tive to its specific, concrete environment. As such, it may (taken in isolation) be ill suited for representing abstracta or genera. As for the singling out of specific properties of specific objects, it may be that models that are realized by neurophysiological processes have a natural advantage over scale models in that populations representing specific properties may *cry out* for attention (by oscillating at the appropriate frequency). There is, moreover, no reason why these lessons could not scale up, so to speak, to account for the human ability to run off-line models of spatial, kinematic, and dynamic relationships. Of course, in humans, the neocortex is likely to play a much more prominent role. As of yet, however, there is little understanding of the precise manner in which the neocortex does, or might, realize mental models.

## 31.4 Realization Story Applied

Though we clearly have a long way to go, the above hypothesis about what mental models are such that neural systems might realize them looks to have important ramifications for work in several fields, ranging from AI to the philosophy of science.

### 31.4.1 AI and Psychology: Towards an Intuitive Physics Engine

One obvious ramification of the above is what it suggests about how one might go about endowing computational artifacts with the kind of boundless commonsense knowledge of the consequences of alterations to the world that humans seem to possess. FEMs prove that there is a determinate computational solution to the prediction and qualification problems. FEMs are generative in that they can be manipulated in any of countless ways in order to make inferences about how alterations to the environment might play out and, by the same token, about the ways in which those consequences might be defeated. It would thus behoove AI researchers to incorporate media for the construction of intrinsic models within the core inference machinery of their devices. Indeed, there has been some movement in this direction in recent years. For instance, though past manuals for the Soar production-system architecture evidence a certain degree of exasperation when it comes to the frame problem, more recent manuals indicate that Soar's designers have begun to offload mechanical reasoning to nonsentential models. *Laird* notes, for instance [31.74]:

“With the addition of visual imagery, we have demonstrated that it is possible to solve spatial rea-

soning problems orders of magnitude faster than without it, and using significantly less procedural knowledge. Visual imagery also enables processing that is not possible with only symbolic reasoning, such as determining which letters in the alphabet are symmetric along the vertical axis (A, H, I, M, O, T, U, V, W, X, Y).”

While Soar's imagery module still only supports simple spatial reasoning, it is clearly a step in the direction of richer, intrinsic models of 3-D kinematics and dynamics.

There has also been some movement in the direction of using computationally realized intrinsic models as a way of making sense of behavioral findings regarding how humans engage in commonsense reasoning about the world. For instance, after paying homage to Craik, MIT researchers *Battaglia* et al. describe their innovative approach to commonsense reasoning as follows [31.75]:

“We posit that human judgments are driven by an *intuitive physics engine* (IPE), akin to the computer physics engines used for quantitative but approximate simulation of rigid body dynamics and collisions, soft body and fluid dynamics in computer graphics, and interactive video games.”

They simulate the IPE with FEMs of full-blown 3-D kinematic and dynamic relationships. They note that a similar IPE in humans might allow us to read off from our simulations the answers to questions of *What will happen?* regarding innumerable novel scenarios. Their pioneering work also breaks new ground in that it be-

gins to account for probabilistic reasoning by building a bit of uncertainty into models and treating multiple runs of a model as a statistical sample.

All of this work is very much in the spirit of Schwartz' claim that "inferences can emerge through imagined actions *even though people may not know the answer explicitly*" [31.76, italics mine]. It also fits with the following suggestion of Moulton and Kosslyn [31.35, italics mine]:

"the primary function of mental imagery is to allow us to generate specific predictions based upon past experience. Imagery allows us to answer *what if* questions by making explicit and accessible the likely consequences of being in a specific situation or performing a specific action."

### 31.4.2 Exduction

Another important lesson to be learned from computationally realized intrinsic models is that they support a form of mechanistic reasoning that has found its way into few, if any, standard reasoning taxonomies. As Glasgow and Papadias claim [31.61]:

"The spatial structure of images has properties not possessed by deductive sentential representations [...] spatial image representations [...] support nondeductive inference using built-in constraints on the processes that construct and access them."

Of course, there is more to be said about the process of model-based mechanistic reasoning than that it is *not* deductive. In fact, the process shares with (valid) deductive reasoning the property of being monotonic. What makes deduction a monotonic (indefeasible) reasoning process is that the conclusion of a valid argument cannot be overturned simply by adding premises; it can only be overturned by rejecting one or more of the premises from which the conclusion was deduced. Other forms of reasoning (inductive generalization, analogical reasoning, abduction) are defeasible in that one *can* overturn their conclusions simply by adding relevant premises. For instance, if I hear a meowing noise emanating from my daughter's closet door, I may

infer that the cat is trapped inside. But if I then see the cat walking through the kitchen and am told that my daughter was given a new electronic cat toy, my conclusion would be undermined while at the same time leaving the original premise (that there is meowing coming from the closet) intact.

One diagnosis for why deduction is monotonic is that, in a certain sense, the premises of a valid deduction already *contain* the information stated in the conclusion, so adding information takes nothing away from the support that those premises lend to the conclusion. That means that insofar as the original premises are true, the conclusion must be as well, and insofar as the conclusion is false, there must be something wrong with the premises used to derive it. But deduction is formal, in that topic-neutral logical particles are what bear the entirety of the inferential load – that is, the specific contents (consistently) connected and quantified over drop out as irrelevant.

The use of scale models and FEMs makes evident that there is another form of monotonic reasoning in addition to deduction. As explained above, information derived regarding the consequences of interventions on a modeled system are to a significant extent already *contained* (i. e., they are implicit) in the models themselves. The only way to overturn a model-based inference is to call into question some aspect or other of the model from which it was derived. By the same token, if the conclusion is incorrect, there must be something wrong with the model. But unlike deduction, model-based reasoning is not affected by abstracting away from specific contents and allowing logical particles to bear the inferential load. Instead, it is the specific, concrete contents of the models that do all of the work. As yet, this form of monotonic reasoning lacks a name. Let us thus call it *exduction* (*ex*-out and *duce*-lead). Like deduction, exduction may be implemented externally through the use of representational artifacts, but the hypothesis being explored here is just that we also sometimes engage in internal exductive reasoning through the use of mental models. If this hypothesis is correct, then exduction must be added to our standard taxonomy of internal reasoning processes and placed alongside deduction under the broader heading of monotonic reasoning.

## 31.5 Mechanistic Explanation Revisited

It was noted in the beginning that mental models may well play a crucial role in the process of mechanistic explanation and prediction. If that is so, then we can only hope to attain a deep understanding of science if

we first account for how the mental modeling process works. Now that we have a clearer conception of the distinctive features of mental models and of the way in which they might be realized by neurophysiological

processes, we can begin to see what the payoff might be in terms of our understanding of model-based reasoning about mechanisms in science.

### 31.5.1 The Prediction and *Ceteris Paribus* Problems

To give some of the flavor of where this might lead, consider that one benefit of the foregoing realization story regarding mental models for the philosophy of science is that it offers a solution to two longstanding problems: the surplus-meaning problem and the *ceteris paribus* problem (a.k.a., the problem of *provisos*). Both problems arose as a consequence of attempts to apply the methods of formal, mostly deductive methods in an attempt to provide a logical reconstruction of scientific reasoning.

The surplus-meaning problem has to do with the fact that explanatory hypotheses have, and are *known* to have, countless implications beyond the happenings they explain. To keep things simple, consider the non-scientific case of what a mechanic knows about the operation of an automobile engine. Imagine, in particular, that someone has brought an automobile into the mechanic's shop complaining that the engine has suffered a drop-off in power. Listening to the engine, the mechanic might decide that the engine has blown a ring. Setting aside the question of creativity, one might provide the following formal, deductive reconstruction of his explanatory model

If one of the cylinders has lost a ring,  
 then the engine will have lost power.  
 One of the cylinders has lost a ring.  
 ∴ The engine has lost power.

Consider, however, that the mechanic knows not only that the faulty ring will result in a loss of power (the *explanandum* occurrence); he also knows many other implications of this explanatory hypothesis, such as that oil will leak into the combustion chamber, the exhaust will look smoky, the end of the tailpipe will become oily, and the sparkplugs will turn dark. He also knows what interventions will and will not alleviate the problem – for instance, he knows that replacing the rings will restore power but replacing the air filter will not, and so on indefinitely. Any suitable deductive reconstruction of the mechanic's mental model of the source of the problem must thus account not just for the fact that it implies the *explanandum*; it must account for its countless further implications. The general problem with deductive reconstructions of explanations – a problem that will surely beset any attempt

to reconstruct mechanistic explanations using extrinsic representations – is that they fail to capture the full complexity of what anyone who possesses that explanation must know, if only implicitly. The problem is that there is too much *surplus-meaning* to express it all explicitly [31.77], and these added implications are essential to how we assess the adequacy of explanations, whether in everyday life or in science. As Greenwood explains [31.78]:

“Where this *surplus meaning* comes from [...] is a matter of some dispute, but that genuine theories poses [sic.] such surplus meaning is not—for this is precisely what accounts for their explanatory power and creative predictive potential.”

Notice also that the mechanic's model of why the car lost power not only contains information about the various other things he should expect to find if that model is correct; it also contains information about the countless ways in which each of these expectations might, consistent with the truth of the explanation, be defeated. The mechanic knows, for instance, that replacing the rings will restore power, but *only if it is not the case that one of the spark plug wires was damaged in the process, the air filter has become clogged with dust from a nearby construction project, and so on indefinitely.*

Whether we are dealing with commonsense or scientific reasoning about mechanisms, the problem with attempts at formalizing our knowledge of the ways in which a given implication is qualified is that what we know far outstrips what can be expressed explicitly in the form of, say, a conditional generalization. In philosophy of science, the qualifications that would have to be added are termed *ceteris paribus* clauses and *provisos*. As Fodor claims, the general problem is that “as soon as you try to make these generalizations explicit, you see that they have to be hedged about with *ceteris paribus* clauses” [31.79]. And as Giere claims, “the number of *provisos implicit* in any law is *indefinitely large*” [31.80, italics mine]. Deductive models, and others that rely upon extrinsic representations, are thus unable to capture the full breadth of what we know when we possess a mechanistic explanation for an occurrence. And all of this *dark information* plays a crucial role in the testing and retention of mechanistic hypotheses [31.81]. Accounting for it must thus be viewed as a central goal for the philosophy of science.

If we abandon deductive (or other extrinsic) schemes for making sense of scientific reasoning about mechanisms and instead adopt an exductive (and intrinsic) model-based account of reasoning, the surplus-meaning and *ceteris paribus* problems dissolve, and the

source of the dark information comes into focus. After all, these two problems are just variants on the prediction and qualifications problems of AI. This is not surprising given that both sets of problems were discovered through early attempts to deductively reconstruct everyday and scientific reasoning about mechanisms. Naturally, the same solution applies in both cases: Eschew the appeal to extrinsic representations and formal inferences in favor of an appeal to intrinsic models and exductive inferences. The promising idea that emerges is that scientists may be utilizing intrinsic mental models to understand the mechanisms that are (or might be) responsible for particular phenomena. Such models would endow the scientist with boundless tacit knowledge of the further implications of a given mechanistic hypothesis and of the countless ways in which those implications are qualified.

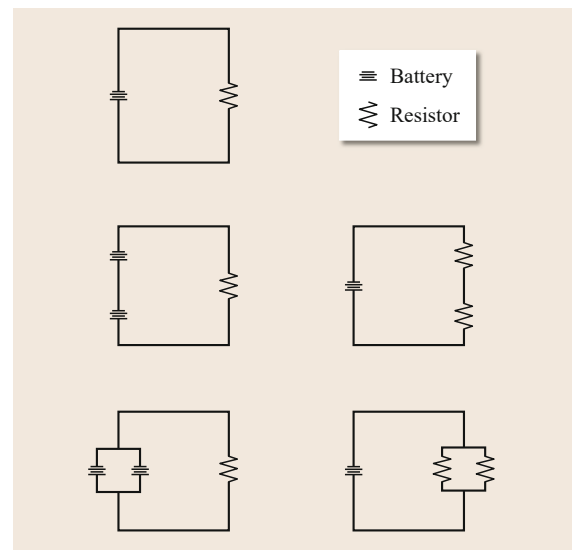
### 31.5.2 Beyond Mental Models

We must not forget, however, that our mental models are limited by working memory capacity and by the heavy cognitive load associated with mental modeling. Even so, scientists are somehow able to formulate and comprehend some remarkably complex mechanical explanations. Seen in this light, it is no surprise that, in their reasoning about mechanisms, humans rely heavily upon external representational artifacts such as diagrams. These can act as aids to memory, both short- and long-term, enabling us to off-load some of the cognitive burden to the world and thereby compensating for our otherwise limited ability to keep track of the simultaneous influences of many mechanical components (see [31.82]). Indeed, when aided by external diagrams, *Hegarty* et al. found that high- and low-imagery subjects performed about *equally well* on a task that required model-based reasoning about mechanisms [31.18]. The compensatory influence of external representations is strong indeed (see also *Bechtel*, Chap. 27).

Of course we do not just utilize static pictures to make sense of natural phenomena; we also sometimes use scale models [31.83]. On the present view, the reason so many have come to view these models as an apt metaphor for in-the-head reasoning may be that scale models recapitulate, albeit in a way that overcomes many of our cognitive frailties, the structure of our internal models of mechanisms. However, with the advent of sophisticated computer models, we now have an even better tool for investigating the implications of mechanical hypotheses. As we have seen, certain computer models (e.g., FEMs) are like scale models in that they constitute intrinsic nonsentential representations of actual or hypothetical mechanisms. However,

these models have the added virtue that one can easily freeze the action, zoom in or out, slow things down, and even watch things play out in reverse. These models thus constitute what *Churchland* and *Sejnowski* term a “fortunate preparation” [31.84]. Such models provide an even more apt analogy for understanding our own native mental models, for both sorts of models are realized at a low level by complicated circuitry, and both tend to bottom-out well above the level of nature’s fundamental laws. (One way of describing the interplay between external models (in their various forms) and internal mental models would be to say that the latter are part and parcel of scientific cognition, whereas the former are representational artifacts created to aid cognition. An alternative, somewhat speculative, proposal is that our external models are no less part of the fabric of scientific cognition than are our internal mental models [31.85]).

As noted in Sect. 31.2, over and above the quantitative limits imposed by working memory capacity, scale models and FEMs are, and our own mental models may well be, limited in certain qualitative respects, such as their ability to represent abstracta. But surely thoughts about abstracta play a big role in scientific reasoning. One way of accounting for this is to say that our deficiencies with regards to modeling abstracta are the precise problem that analogy and metaphor were created to solve. This would make sense of why the language we use to represent abstracta (e.g., *economic inflation*) is so shot through with analogies and metaphors rooted in concrete domains [31.86–88].



**Fig. 31.6** Various configurations of circuitry, batteries, and resistors

Gentner and Gentner's study of human reasoning about electricity lends some credence to this view [31.89]. Gentner and Gentner found that in conversation, nonexpert subjects commonly likened the flow of electricity to water moving through pipes or crowds moving through corridors. Each analogy happens to yield its own unique set of inaccurate predictions about how electrical current moves through particular configurations of electrical components (Fig. 31.6).

## 31.6 Conclusion

It was noted in the beginning that mental models might play a crucial role in the process of mechanistic explanation and prediction in science. As such, we can only attain a deep understanding of science itself if we first understand that nature of this mental, model-based, reasoning process. We then saw that experimental psychologists have long maintained that mental models are distinct from sentential representations in much the way that scale models are. Insofar as this hypothesis is viable, we can expect that experimental psychology will provide crucial insight into both the nature and limits of our onboard mental models. At the same time, it is important to recognize that the many appeals to distinctively model-like mental representations in psychology will be considered suspect so long as we lack a reasonable way of spelling out what sorts of representational structures mental models are supposed to be in a way that (i) shows these models to be distinct from sentential representations while (ii) allowing for their realization by neurophysiological processes. We can see the way forward, however, if we first pay attention to some

Gentner and Gentner found that subjects' errors in reasoning about these components tracked the analogies they invoked when discussing electricity. This suggests that these analogies run deeper than the surface features of language and penetrate right into subjects' mental models of the flow of electricity. Analogies and metaphors are perhaps not the whole story of how we think about abstracta, but they may well be an important part of it.

of the distinctive features of scale models that distinguish them from sentential representations. If we then turn to the computational realm, we see that these very features (including immunity to the notorious frame problem) are exhibited by certain computational models of mechanisms such as FEMs. An appeal to the principle that sustains the computational theory of cognition (POPI) enables us to understand how this could be so and how high-level, nonsentential, intrinsic models of mechanisms could in principle be realized by neurophysiological processes. The broader viability of this realization story for mental models is suggested by recent work in both AI and experimental psychology and by the elegant solution it offers to the surplus-meaning and *ceteris paribus* problems in the philosophy of science. Going forward, the idea that our scientific reasoning about mechanisms might, to a large extent, involve the manipulation of representations that are like scale models in crucial respects can be regarded as at least one, sturdy pillar of a promising hypothesis regarding the nature of model-based reasoning in science.

## References

- 31.1 R. Giere: *Explaining Science: A Cognitive Approach* (Univ. Chicago Press, Chicago 1988)
- 31.2 P. Railton: A deductive-nomological model of probabilistic explanation, *Philos. Sci.* **45**, 206–226 (1978)
- 31.3 W. Salmon: *Scientific Explanation and the Causal Structure of the World* (Princeton Univ. Press, Princeton 1984)
- 31.4 W. Salmon: *Causality and Explanation* (Oxford Univ. Press, New York 1998)
- 31.5 P. Machamer, L. Darden, C. Craver: Thinking about mechanisms, *Philos. Sci.* **67**, 1–25 (2000)
- 31.6 W. Bechtel: *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience* (Psychology Press, New York 2009)
- 31.7 C. Wright, W. Bechtel: Mechanisms and psychological explanation. In: *Philosophy of Psychology and Cognitive Science*, ed. by P. Thagard (Elsevier, New York 2007) pp. 31–79
- 31.8 J. Waskan, I. Harmon, A. Higgins, J. Spino: Investigating lay and scientific norms for using explanation. In: *Modes of Explanation: Affordances for Action and Prediction*, ed. by M. Lissack, A. Graber (Palgrave Macmillan, New York 2014) pp. 198–205
- 31.9 K. Craik: *The Nature of Explanation* (Cambridge Univ. Press, Cambridge 1943)
- 31.10 N. Nersessian: Mental modeling in conceptual change. In: *International Handbook of Conceptual Change*, ed. by S. Vosniadou (Routledge, London 2007) pp. 391–416
- 31.11 P. Thagard: *The Cognitive Science of Science* (MIT Press, Cambridge 2012)
- 31.12 E. Tolman: Cognitive maps in rats and men, *Psychol. Rev.* **55**, 189–208 (1948)
- 31.13 R. Shepard, J. Metzler: Mental rotation of three-dimensional objects, *Science* **171**, 701–703 (1971)
- 31.14 S. Kosslyn: *Image and Mind* (Harvard Univ. Press, Cambridge 1980)

- 31.15 D. Schwartz, J. Black: Analog imagery in mental model reasoning: Depictive models, *Cogn. Psychol.* **30**, 154–219 (1996)
- 31.16 D. Schwartz, J. Black: Shuttling between depictive models and abstract rules: Induction and fall-back, *Cogn. Sci.* **20**, 457–497 (1996)
- 31.17 M. Hegarty: Mechanical reasoning by mental simulation, *Trends Cogn. Sci.* **8**, 280–285 (2004)
- 31.18 M. Hegarty, S. Kriz, C. Cate: The roles of mental animations and external animations in understanding mechanical systems, *Cogn. Instruct.* **21**, 325–360 (2003)
- 31.19 D. Norman: Some observations on mental models. In: *Mental Models*, ed. by D. Gentner, A. Stevens (Lawrence Erlbaum Associates, Hillsdale 1983) pp. 7–14
- 31.20 A. Leslie, S. Keeble: Do six-month old infants perceive causality?, *Cognition* **25**, 265–288 (1987)
- 31.21 A. Schlottman: Seeing it happen and knowing how it works: How children understand the relation between perceptual causality and underlying mechanism, *Dev. Psychol.* **35**, 303–317 (1999)
- 31.22 R. Baillargeon, J. Li, Y. Gertner, D. Wu: How do infants reason about physical events? In: *The Wiley-Blackwell Handbook of Childhood Cognitive Development*, ed. by U. Goswami (Blackwell, Oxford 2011) pp. 11–48
- 31.23 L. Barsalou, K. Solomon, L. Wu: Perceptual simulation in conceptual tasks, *PROCBEGINProc. 4th Conf. Int. Cogn. Linguist. Assoc. Cult. Typol. Psychol. Perspect. Cogn. Linguist. PROCEND*, Vol. 2., ed. by M. Hiraga, C. Sinha, S. Wilcox (John Benjamins, Amsterdam 1999) pp. 209–228
- 31.24 W. Brewer: Scientific theories and naive theories as forms of mental representation: Psychologism revived, *Sci. Educ.* **8**, 489–505 (1999)
- 31.25 R. Langacker: *Concept, Image, and Symbol: The Cognitive Basis of Grammar* (Mouton de Gruyter, New York 1991)
- 31.26 A. Goldberg: *Constructions: A Construction Grammar Approach to Argument Structure* (Univ. Chicago Press, Chicago 1995)
- 31.27 G. Fauconnier: *Mental Spaces* (MIT Press, Cambridge 1985)
- 31.28 A. Garnham: *Mental Models as Representations of Discourse and Text* (John Wiley, New York 1987)
- 31.29 L. Talmy: Force dynamics in language and cognition, *Cogn. Sci.* **12**, 49–100 (1988)
- 31.30 P. Johnson-Laird: How is meaning mentally represented? In: *Meaning and Mental Representation*, ed. by E. Eco, M. Santambrogio, P. Violi (Indiana Univ. Press, Bloomington 1988) pp. 99–118
- 31.31 P. Johnson-Laird: *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness* (Harvard Univ. Press, Cambridge 1983)
- 31.32 P. Johnson-Laird, R. Byrne: *Deduction* (Lawrence Erlbaum Associates, Hillsdale, New Jersey 1991)
- 31.33 W. McCulloch, W. Pitts: A logical calculus of the ideas immanent in nervous activity, *Bull. Math. Biophys.* **5**, 115–113 (1943)
- 31.34 S. Franklin, M. Garzon: Computation by discrete neural nets. In: *Mathematical Perspectives on Neural Networks*, ed. by P. Smolensky, M. Mozer, D. Rumelhart (Lawrence Erlbaum Associates, Mahwah, New Jersey 1996) pp. 41–84
- 31.35 S. Moulton, S. Kosslyn: Imagining predictions: Mental imagery as mental emulation, *Philos. Trans. R. Soc. B* **364**, 1273–1280 (2009)
- 31.36 R. Cummins: *Representations, Targets, and Attitudes* (MIT Press, Cambridge 1996)
- 31.37 L. Wittgenstein: *Philosophical Investigations* (Macmillan, New York 1953)
- 31.38 R. Shepard, S. Chipman: Second-order isomorphism of internal representations: Shapes of states, *Cogn. Psychol.* **1**, 1–17 (1970)
- 31.39 S. Palmer: Fundamental aspects of cognitive representation. In: *Cognition and Categorization*, ed. by E. Rosch, B. Lloyd (Lawrence Erlbaum Associates, Hillsdale, New Jersey 1978) pp. 259–303
- 31.40 N. Nersessian: The cognitive basis of model-based reasoning in science. In: *The Cognitive Basis of Science*, ed. by P. Carruthers, S. Stich, M. Siegal (Cambridge Univ. Press, Cambridge 2002) pp. 133–153
- 31.41 S. Kosslyn: *Image and Brain: The Resolution of the Imagery Debate* (MIT Press, Cambridge 1994)
- 31.42 S. Zeki: The functional organization of projections from striate to prestriate visual cortex in the rhesus monkey, *Cold Spring Harbor Symp. Quant. Biol.* **40**, 591–600 (1976)
- 31.43 M. Mishkin, L. Ungerleider, K. Macko: Object vision and spatial vision: Two cortical pathways, *Trends Neurosci.* **6**, 414–417 (1983)
- 31.44 E. DeYoe, D. Van Essen: Concurrent processing streams in monkey visual cortex, *Trends Neurosci.* **11**, 219–226 (1988)
- 31.45 J. Huttenlocher, E. Higgins, H. Clark: Adjectives, comparatives, and syllogisms, *Psychol. Rev.* **78**, 487–514 (1971)
- 31.46 J.R. Anderson: Arguments concerning representations for mental imagery, *Psychol. Rev.* **85**, 249–277 (1978)
- 31.47 L. Brooks: Spatial and verbal components in the act of recall, *Can. J. Psychol.* **22**, 349–368 (1968)
- 31.48 S. Segal, V. Fusella: Influence of imaged pictures and sounds on detection of visual and auditory signals, *J. Exp. Psychol.* **83**, 458–464 (1970)
- 31.49 N. Block: Mental pictures and cognitive science. In: *Mind and Cognition*, ed. by W.G. Lycan (Basil Blackwell, Cambridge, Massachusetts 1990) pp. 577–606
- 31.50 J.A. Fodor: Imagistic representation. In: *Imagery*, ed. by N. Block (MIT Press, Cambridge 1981) pp. 63–86
- 31.51 J. McCarthy, P. Hayes: Some philosophical problems from the standpoint of artificial intelligence. In: *Machine Intelligence*, ed. by B. Meltzer, D. Michie (Edinburgh Univ. Press, Edinburgh 1969) pp. 463–502
- 31.52 P. Hayes: The frame problem and related problems in artificial intelligence. In: *Readings in Artificial Intelligence*, ed. by B. Webber, N. Nilsson (Morgan Kaufman, Los Altos 1981) pp. 223–230
- 31.53 L. Janlert: The frame problem: Freedom or stability? With pictures we can have both. In: *The Robot's*

- Dilemma Revisited: The Frame Problem in Artificial Intelligence*, ed. by K.M. Ford, Z. Pylyshyn (Ablex Publishing, Norwood, New Jersey 1996) pp. 35–48
- 31.54 J. McCarthy: Applications of circumscription to formalizing common-sense knowledge, *Artif. Intell.* **28**, 86–116 (1986)
- 31.55 J. Waskan: Applications of an implementation story for non-sentential models. In: *Model-Based Reasoning in Science and Technology*, ed. by L. Magnani, W. Carnielli, C. Pizzi (Springer, Berlin 2010) pp. 463–476
- 31.56 C. Congdon, J. Laird: *The Soar User's Manual: Version 7.0.4* (Univ. Michigan, Ann Arbor 1997)
- 31.57 J. Waskan: Intrinsic cognitive models, *Cogn. Sci.* **27**, 259–283 (2003)
- 31.58 J. Waskan: *Models and Cognition* (MIT Press, Cambridge 2006)
- 31.59 R. Descartes: Discourse on the method. In: *The Philosophical Writings of Descartes*, (Cambridge Univ. Press, Cambridge 1988) pp. 20–56, trans. by J. Cottingham, R. Stoothoff, D. Murdoch
- 31.60 J. Haugeland: An overview of the frame problem. In: *Robot's Dilemma*, ed. by Z. Pylyshyn (Ablex Publishing Corp, Norwood 1987) pp. 77–93
- 31.61 J. Glasgow, D. Papadias: Computational imagery, *Cogn. Sci.* **16**(3), 355–394 (1992)
- 31.62 K. Sterelny: The imagery debate. In: *Mind and Cognition*, ed. by W. Lycan (Blackwell, Cambridge 1990) pp. 607–626
- 31.63 Z. Pylyshyn: *Computation and Cognition: Toward a Foundation for Cognitive Science* (MIT Press, Cambridge, Massachusetts 1984)
- 31.64 J.A. Fodor: *The Mind Doesn't Work That Way* (MIT Press, Cambridge 2000)
- 31.65 G. Leibniz: The Monadology. In: *The Monadology and Other Philosophical Writings* (Clarendon Press, Oxford 1898) pp. 215–271, trans. by R. Lotta
- 31.66 J.A. Fodor: *The Language of Thought* (Thomas Y. Crowell, New York 1975)
- 31.67 C. Eliasmith, C.H. Anderson: *Neural Engineering: Computation, Representation and Dynamics in Neurobiological Systems* (MIT Press, Cambridge 2003)
- 31.68 J. O'Keefe, L. Nadel: *The hippocampus as a cognitive map* (Oxford Univ. Press, New York 1978)
- 31.69 V. Brun, M. Otnass, S. Molden, H. Steffenach, M. Witter, M. Moser, E. Moser: Place cells and place recognition maintained by direct entorhinal-hippocampal circuitry, *Science* **296**, 2243–2246 (2002)
- 31.70 J. O'Keefe, N. Burgess, J. Donnett, K. Jeffery, E. Maguire: Place cells, navigational accuracy, and the human hippocampus, *Philos. Trans. R. Soc. B* **353**, 1333–1340 (1998)
- 31.71 M. Fyhn, T. Hafting, A. Treves, M. Moser, E. Moser: Hippocampal remapping and grid realignment in entorhinal cortex, *Nature* **446**, 190–194 (2007)
- 31.72 F.S.J. Knierim: Coming up: In search of the vertical dimension in the brain, *Nat. Neurosci.* **14**, 1102–1103 (2011)
- 31.73 K. Woollett, E. Maguire: Acquiring 'the knowledge' of London's layout drives structural brain changes, *Curr. Biol.* **21**, 2109–2114 (2011)
- 31.74 J. Laird: Extending the Soar cognitive architecture, Proc. 1st AGI Conf., ed. by P. Wang, B. Goertzel, S. Franklin (IOS Press, Amsterdam 2008) pp. 224–235
- 31.75 P. Battaglia, J. Hamrick, J. Tenenbaum: Simulation as an engine of physical scene understanding, *Proc. Natl. Acad. Sci. USA* **110**, 18327–18332 (2013)
- 31.76 D. Schwartz: Physical imagery: Kinematic versus dynamic models, *Cogn. Psychol.* **38**, 433–464 (1999)
- 31.77 H. Reichenbach: *Experience and Prediction: An Analysis of the Foundations and the Structure of Knowledge* (Univ. Chicago Press, Chicago 1938)
- 31.78 J. Greenwood: Folk psychology and scientific psychology. In: *The Future of Folk Psychology*, ed. by J. Greenwood (Cambridge Univ. Press, Cambridge 1991) pp. 1–21
- 31.79 J.A. Fodor: *Psychosemantics: The Problem of Meaning in the Philosophy of Mind* (MIT Press, Cambridge 1984)
- 31.80 R. Giere: Laws, theories, and generalizations. In: *The Limits of Deductivism*, ed. by A. Grünbaum, W. Salmon (Univ. California Press, Berkeley 1988) pp. 37–46
- 31.81 J. Waskan: Knowledge of counterfactual interventions through cognitive models of mechanisms, *Int. Stud. Philos. Sci.* **22**, 259–275 (2008)
- 31.82 A. Clark: *Mindware: An Introduction to the Philosophy of Cognitive Science* (Oxford Univ. Press, New York 2013)
- 31.83 M. Weisberg: Who is a modeler?, *Br. J. Philos. Sci.* **58**, 207–233 (2007)
- 31.84 P. Churchland, T. Sejnowski: Neural representation and neural computation, *Philos. Perspect. Action Theory Philos. Mind* **4**, 343–382 (1988)
- 31.85 R. Giere: Models as parts of distributed cognitive systems. In: *Model Based Reasoning: Science, Technology, Values*, ed. by L. Magnani, N. Nersessian (Kluwer, Amsterdam 2002) pp. 227–241
- 31.86 G. Lakoff, M. Johnson: *Metaphors We Live By* (Univ. Chicago Press, Chicago 1980)
- 31.87 G. Lakoff: *Women, Fire, and Dangerous Things* (Univ. Chicago Press, Chicago 1987)
- 31.88 D. Fernandez-Duque, M. Johnson: Attention metaphors: How metaphors guide the cognitive psychology of attention, *Cogn. Sci.* **23**, 83–116 (1999)
- 31.89 D. Gentner, D.R. Gentner: Flowing waters or teeming crowds: Mental models of electricity. In: *Mental Models*, ed. by D. Gentner, A. Stevens (Lawrence Erlbaum Associates, Hillsdale 1983) pp. 99–129

---

# Modelling **Part G**

## Part G Modelling and Computational Issues

Ed. by Francesco Amigoni, Viola Schiaffonati

### 32 Computational Aspects

#### of Model-Based Reasoning

Gordana Dodig-Crnkovic, Göteborg, Sweden

Antonio Cichetti, Västerås, Sweden

### 33 Computational Scientific Discovery

Peter D. Sozou, London, UK

Peter C.R. Lane, Hatfield, UK

Mark Addis, Twickenham, UK

Fernand Gobet, Liverpool, UK

### 34 Computer Simulations

#### and Computational Models in Science

Cyrille Imbert, Nancy, France

### 35 Simulation of Complex Systems

Paul Davidsson, Malmö, Sweden

Franziska Klügl, Örebro, Sweden

Harko Verhagen, Kista, Sweden

### 36 Models and Experiments in Robotics

Francesco Amigoni, Milano, Italy

Viola Schiaffonati, Milano, Italy

### 37 Biorobotics

Edoardo Datteri, Milano, Italy



Since the 1950s computational progress has been contributing to increasing the relevance of the discourse about models, making it not only relevant to scientists and philosophers, but also to computer scientists, programmers, and logicians. Emphasis has been put both on the application of computational tools to a range of disciplines and on the computational issues themselves. Computing is playing an increasing role in several scientific endeavors, in modeling and simulating different entities, and in practically enhancing the performance of various scientific activities. At the same time, computational tools are becoming more and more complex and present several open issues that require consideration from technical, methodological, and epistemological points of view.

This part of the *Handbook of Model-Based Science* discusses the modeling and computational issues arising in this context, with the aim of giving back an articulated, although necessarily incomplete, picture of the field. It is composed of six different chapters that alternate with general perspectives and specific fields of application.

This part opens with **Chap. 32** on *Computational Aspects of Model-Based Reasoning* by *Gordana Dodig-Crnkovic* and *Antonio Cicchetti* offering an introductory overview on the use of computational models and tools for the study of cognition and model-based reasoning. From simple agents, like bacteria, to the complex human cognitive systems, computation is meant as physical, natural, embodied, and distributed, and it is discussed in relation to the view of symbol manipulation of classical computationalism.

**Chapter 33** by *Peter Sozou*, *Peter Lane*, *Mark Addis*, and *Fernand Gobet* discusses *computational scientific discovery* as a particularly interesting and successful field of application of computer-supported model-based science. The chapter reviews the application of computational methods in the formulation of scientific ideas and acknowledges the importance of this field not only for historical reasons, with the first systems having played a disruptive role in the philosophical debate on scientific discovery, but also for testifying its increasing importance in many areas of science.

**Chapter 34**, *Computer Simulations and Computational Models in Science* by *Cyrill Imbert* presents a very rich examination of computational science and computer simulations by giving reason to the constant attempts of extending human computational capacities. The chapter covers a wide variety of topics and themes

with the awareness that epistemological analyses of simulations are, to a large degree, contextual and that these analyses require developing insights about the evolving relation between human capacities and computational science.

Simulation is also the topic of **Chap. 35**, *Simulation of Complex Systems* by *Paul Davidsson*, *Franziska Klügl*, and *Harko Verhagen*, which opens with a discussion on what characterizes complex systems, such as huge ecosystems and traffic systems. Different approaches to model complex systems are presented, but particular attention is devoted to agent-based simulations, to their intuitiveness and flexibility, to some solutions proposed in the last years, and to the still open problems that are discussed in a critical perspective.

**Chapter 36**, by *Francesco Amigoni* and *Viola Schiaffonati*, *Models and Experiments in Robotics*, surveys the practices being employed in experimentally assessing the special class of computational models embedded in robots. This assessment is particularly challenging due to the difficulty of satisfactorily estimating the interactions between robots and their environments. Moreover, by considering also related topics such as simulations, benchmarks, standards, and competitions, this chapter shows how the recent debate on the implementation of the experimental method in this field is still very open.

**Chapter 37**, *Biorobotics* by *Edoardo Datteri*, provides an overview of the biorobotic strategy for testing mechanistic explanations of animal behavior starting from a reflection on the various roles played by robotic simulations in scientific research. Besides the history and state of the art of biorobotics, the chapter also addresses some key epistemological and methodological issues mainly concerning the relationships between biorobots and the theoretical models under investigation.

It is not by chance that this part of the volume ends with this chapter on biorobotics: if one of the main common traits of the other chapters has been the foundational role of the philosophical tools in discussing computational models in model-based science, this last chapter also shows how computational models and tools can offer new insights to traditional philosophical problems, and thus represents an ideal and critical conclusion offering further reflections on the articulation between computation and philosophy.

## 32. Computational Aspects of Model-Based Reasoning

Gordana Dodig-Crnkovic, Antonio Cichetti

Computational models and tools provide increasingly solid foundations for the study of cognition and model-based reasoning, with knowledge generation in different types of cognizing agents, from the simplest ones like bacteria to the complex human distributed cognition. After the introduction of the computational turn, we proceed to models of computation and the relationship between information and computation. A distinction is made between mathematical and computational (executable) models, which are central for biology and cognition. Computation as it appears in cognitive systems is physical, natural, embodied, and distributed computation, and we explain how it relates to the symbol manipulation view of classical computationalism. As present day models of distributed, asynchronous, heterogeneous, and concurrent networks are becoming increasingly well suited for modeling of cognitive systems with their dynamic properties, they can be used to study mechanisms of abduction and scientific discovery. We conclude the chapter with the presentation of software modeling with computationally automated reasoning and the discussion of model transformations and separation between semantics and ontology.

32.1	<b>Computational Turn Seen from Different Perspectives</b> .....	695
32.2	<b>Models of Computation</b> .....	697
32.2.1	Turing Model of Computation and Its Scope .....	698
32.2.2	Computation as Information Processing .....	698
32.3	<b>Computation Versus Information</b> .....	700
32.4	<b>The Difference Between Mathematical and Computational (Executable) Models</b> .....	702
32.5	<b>Computation in the Wild</b> .....	703
32.5.1	Physical Computation – Computing Nature as Info-Computation .....	703
32.5.2	New Computationalism. Nonsymbolic versus Symbolic Computation .....	704
32.6	<b>Cognition: Knowledge Generation by Computation of New Information</b> ..	706
32.6.1	Distributed Cognition and Model-Based Reasoning .....	707
32.6.2	Computational Aspects of Model-Based Reasoning in Science ..	708
32.7	<b>Model-Based Reasoning and Computational Automation of Reasoning</b> .....	709
32.8	<b>Model Transformations and Semantics: Separation Between Semantics and Ontology</b> .....	712
	<b>References</b> .....	715

### 32.1 Computational Turn Seen from Different Perspectives

Computation is central for our entire contemporary civilization – research and sciences, communications, government, manufacturing, control, transports, financial sector, education, entertainment, arts, humanities, technology in general, and especially engineering. It also affects in profound ways our outlook of the world and determines what can be conceptualized and how. In the following, we will describe current understanding of the processes and structures of the computational turn from different perspectives.

From the point of view of conceptual, cognitive aspects, *Brian Cantwell Smith* recognizes the computational turn as present constantly increasing importance of computation (cited in [32.1, p. 3]):

“Everyone knows that computational and information technology has spread like wildfire throughout academic and intellectual life. But the spread of computational ideas has been just as impressive.

Biologists not only model life forms on computers; they treat the gene, and even whole organisms, as information systems. Philosophy, artificial intelligence, and cognitive science don't just construct computational models of mind; they take cognition to be computation, at the deepest levels.

Physicists don't just talk about the information carried by a subatomic particle; they propose to unify the foundations of quantum mechanics with notions of information. Similarly for linguists, artists, anthropologists, critics, etc."

With the emphasis on philosophical facets of the computational turn, *Charles Ess* and *Ruth Hagengruber* depict the same process of steady increase in the role of computation as follows [32.2]:

"In the West, philosophical attention to computation and computational devices is at least as old as Leibniz. But since the early 1940s, electronic computers have evolved from a few machines filling several rooms to widely diffused – indeed, ubiquitous – devices, ranging from networked desktops, laptops, smartphones and *the internet of things*. Along the way, initial philosophical attention – in particular, to the ethical and social implications of these devices (Norbert Wiener, 1950) – became sufficiently broad and influential as to justify the phrase *the computational turn* by the 1980s.

In part, the computational turn referred to the multiple ways in which the increasing availability and usability of computers allowed philosophers to explore a range of traditional philosophical interests – e.g., in logic, artificial intelligence, philosophical mathematics, ethics, political philosophy, epistemology, ontology, to name a few – in new ways, often shedding significant new light on traditional issues and arguments. Simultaneously, computer scientists, mathematicians, and others whose work focused on computation and computational devices often found their work to evoke (if not force) reflection and debate precisely on the philosophical assumptions and potential implications of their research."

Looking from the perspective of arts, humanities, and social sciences *David M. Berry* [32.3] observes:

"The importance of understanding computational approaches is increasingly reflected across a number of disciplines, including the arts, humanities and social sciences, which use technologies to shift the critical ground of their concepts and theories – something that can be termed a computational turn.

This is shown in the increasing interest in the digital humanities (Schreibman et al., 2008) and

computational social science (Lazer et al., 2009), as evidenced, for example, by the growth in journals, conferences, books and research funding. In the digital humanities 'critical inquiry involves the application of algorithmically facilitated search, retrieval, and critical process that [...] originat[es] in humanities-based work'; therefore 'exemplary tasks traditionally associated with humanities computing hold the digital representation of archival materials on a par with analysis or critical inquiry, as well as theories of analysis or critical inquiry originating in the study of those materials' (Schreibman et al., 2008: xxv). In social sciences, *Lazer et al.* argue that 'computational social science is emerging that leverages the capacity to collect and analyze data with an unprecedented breadth and depth and scale' (2009).

*Latour* speculates that there is a trend in these informational cascades, which is certainly reflected in the ongoing digitalisation of arts, humanities and social science projects that tends towards 'the direction of the greater merging of figures, numbers and letters, merging greatly facilitated by their homogeneous treatment as binary units in and by computers (Latour, 1986: 16)."

Finally, from the perspective of computing itself as a field, *Peter J. Denning* ascertains [32.4, p. 1-1; p. 1-4]:

"Computing is integral to science – not just as a tool for analyzing data but also as an agent of thought and discovery. It has not always been this way. Computing is a relatively young discipline. It started as an academic field of study in the 1930s with a cluster of remarkable papers by Kurt Gödel, Alonzo Church, Emil Post, and Alan Turing. The papers laid the mathematical foundations that would answer the question, *what is computation?* and discussed schemes for its implementation. These men saw the importance of automatic computation and sought its precise mathematical foundation."

"All this suggests that computing has developed a paradigm all its own (Denning and Freeman, 2009). Computing is no longer just about algorithms, data structures, numerical methods, programming languages, operating systems, networks, databases, graphics, artificial intelligence, and software engineering, as it was prior to 1990. It now also includes exciting new subjects including Internet, web science, mobile computing, cyberspace protection, user-interface design, and information visualization. The resulting commercial applications have spawned new research challenges in social networking, endlessly evolving computation, music, video, digital photography, vision, massive

multiplayer online games, user-generated content, and much more. [...] The central focus of the computing paradigm can be summarized as information processes—natural or constructed processes that transform information. They can be discrete or continuous.”

Denning reminds that in the beginning of the field “Computation was taken to be the mechanical steps followed to evaluate mathematical functions. Computers were people who did computations.” while, today computing plays a much more diverse role: “Computing is not only a tool for science but also a new method of thought and discovery in science.” [32.4, p. 1-3]

Computing is the fourth great domain of science, together with traditional domains of physical-, life-, and social sciences [32.5].

In more detail, here is how Computer Science is relevant for other sciences in view of *Samson Abramsky* and *Bob Coecke* [32.6]:

“Computer Science has something more to offer to the other sciences than the computer. In particular, in the mathematical and logical understanding of fundamental transdisciplinary scientific concepts such as interaction, concurrency and causality, synchrony and asynchrony, compositional modeling and reasoning, open versus closed systems, qual-

itative versus quantitative reasoning, operational methodologies, continuous versus discrete, hybrid systems and more. Computer Science is far ahead of many other sciences, due in part to the challenges arising from the amazing rapidity of the technology change and development it is constantly being confronted with. One could claim that computer scientists (maybe without realizing it) constitute an avant-garde for the sciences in terms of providing fresh paradigms and methods.”

All the above observations made by researchers with different perspectives on computing provide evidence supporting the commonly observed emergence of computational turn in its different manifestations – from technological to cultural, artistic, cognitive, conceptual, and modeling aspects. As Edsger Dijkstra famously said, “Computing is no more about computers than astronomy is about telescopes” (as quoted in [32.7]) – that is to say that computers (as we know them and develop them) are only the tools for computing. Understanding of computing requires understanding of computational processes and their mechanisms.

Generative computational methods, according to *Stephen Wolfram*, enable the development of a *new kind of science* [32.8], and a new way of thinking that *Jeanette M. Wing* termed *computational thinking* [32.9].

## 32.2 Models of Computation

Interestingly, in spite of the broadly shared perception of the current computational turn [32.10], it is still an open question what exactly is computing.

English word *computing* [32.4] in German, French, and Italian languages translates into the respective terms *informatik*, *informatique*, and *informatica* (that would correspond to *informatics* in English). However, there is a slight difference in the emphasis. While the English term *computing* has an *empirical* character, the analogous German, French, and Italian term *informatics* has an *abstract* nature [32.11]. The question of nomenclature, *informatics* versus *computing*, can also be seen in the light of the dichotomy *information* versus *computation* or *structure* versus *process* [32.1].

Information as the central idea of informatics/computing is both scientifically and sociologically indicative. Scientifically, it suggests a view of informatics as a generalization of information theory that is concerned not only with the transmission/communication of information but also with its transformation and interpretation. Sociologically, similar to the industrial revolution, which was concerned with the utilizing of matter/energy, we have the information revolution,

which is concerned with the utilizing of information/computation [32.11].

Information is a field of intense theory building, with diverse perspectives and levels of description, with various goals addressed. In the work of *Mark Burgin*, the focus of theory of information is on its fundamentality, diversity, and unification with specific chapters dedicated to general, statistical, semantic, algorithmic, dynamics, and pragmatic information theory [32.12]. *The Handbook of Philosophy of Information* [32.13] addresses major technical approaches, transformation and use of information in the sciences and humanities. Information logic and dynamics is a separate topic addressed in the work of *Johan van Benthem* and *Patrick Allo* [32.14–17]. No wonder that no single definition can embrace the complexity of the knowledge about information.

Equivalent situation can be found in the studies of computation and the subject of computing [32.4, 5, 18, 19]. Among variety of definitions of computing, the following example is indicative of instrumental approach as one describes *the processes* performed by computing machinery: Computing is [32.20]

“[t]he process of utilizing computer technology to complete a task. Computing may involve computer hardware and/or software, but must involve some form of a computer system.”

It is general enough, as a description of *computation* process but leaves open the question of what actually is *computer*. It implicitly seems to assume computer to be a technological device. However in recent years a new field of computing is being developed labeled as *Natural Computing* [32.21] or *Computing Nature* [32.22, 23], where processes in nature are understood as being some kind of (intrinsic, physical) computation.

### 32.2.1 Turing Model of Computation and Its Scope

The first classical model of computation that also served as a definition of computation is the Turing machine model, which takes computation to be *computation of mathematical function*. Logical Computing Machine (Turing’s own expression for Turing machines) was an attempt to give a mathematically precise definition of an algorithm that is a mechanical procedure (followed by a human using pencil and paper, and given unlimited resources). The Church–Turing thesis says that *a function on the natural numbers is computable* [32.24, 25] if and only if it is describable by a Turing machine model.

Besides the Turing machine model, several other models of computation were defined such as lambda calculus, cellular automata, register machines, and substitution systems, which have been shown to be equivalent to using general recursive functions. The Church–Turing thesis has long served as a *definition for computation*. There has never been a proof, but the evidence for its validity comes from the evident practical equivalence of mentioned computational models.

*Georg Kampis* claims that the Church–Turing thesis applies only to *simple systems* [32.26]. According to Kampis, complex systems such as found in biology must be modeled as self-referential, self-organizing structures called *component systems* whose behavior goes far beyond the simple Turing machine model as a more general model of computation [32.26, p. 223]:

“a component system is a computer which, when executing its operations (software) builds a new hardware [...] [W]e have a computer that re-wires itself in a hardware-software interplay: the hardware defines the software and the software defines new hardware. Then the circle starts again.”

I would add an obvious remark. The Turing machine is supposed to be given from the outset – its logic, its (unlimited) physical resources, and the meanings ascribed to its actions. *The Turing Machine essentially*

*presupposes a human as a part of a system* – the human is the one who poses the questions, provides resources, sets up the rules and interprets the answers.

However, today the dramatically increased interactivity and connectivity of computational devices have changed our understanding of the nature of computing [32.27]. Computing models have been successively extended from the initial abstract symbol manipulating models of stand-alone, discrete sequential machines, to the models of physical computing in the natural world, which are in general concurrent, asynchronous processes. For the first time it is possible to model living systems, their informational structures, and dynamics on both symbolic and subsymbolic information processing levels [32.28]. Currently the computation models are being developed to describe embedded, interactive, and networked computing systems [32.29] with an ambition to encompass present-day distributed computational architectures with their concurrent time behavior.

Ever since the time of Turing, the definition of computation is the subject of a debate. The special issue of the journal *Minds and Machines* (1994, 4, 4) was devoted to the question *What is Computation?* The most general is the view of computation as information processing, found in number of mathematical accounts of computing; see [32.30] for exposition. Understanding of computation as information processing is also widespread in biology, neuroscience, cognitive science, and number of other fields. An illuminating case is presented by *David Baltimore* in *How biology became an information science* [32.31]. *Barry Cooper* and *Jan van Leeuwen* Turing centenary volume [32.32] illustrates the current state of the art regarding Turing model and its scope.

In general, for a process to qualify as computation, a mechanism that ensures definability of its behavior must exist, such as algorithm, network topology, physical process, or similar [32.11].

The characterization of computing can be made in several dimensions with orthogonal types: digital/analog, symbolic/subsymbolic, interactive/batch, and sequential/parallel. Nowadays digital computers are used to simulate all sorts of natural processes, including those that in physics are understood as continuous. However, it is important to distinguish between the *mechanism* and *model* of computation [32.33].

### 32.2.2 Computation as Information Processing

*Luciano Floridi* [32.34] presents the list of the five most interesting areas of research for the field of information (and computation) philosophy, containing 18 fundamental questions. Information dynamics is of special interest, as information processing (computation).

Information and computation are two complementary ideas in a similar way to continuum and a discrete set. In its turn continuum – discrete set dichotomy may be seen in a variety of disguises such as time – space; wave – particle; geometry – arithmetic; interaction – algorithm; computation – information. Two elements in each pair presuppose each other, and are inseparably related to each other so that *Dodig-Crnkovic* introduces the concept of info-computation which emphasizes this dual character of information and computation as its dynamics [32.1, 35, 36]. The field of Philosophy of Information is so closely interconnected with the Philosophy of Computation that it would be appropriate to call it Philosophy of Information and Computation, having in mind the dual character of information-computation. *Burgin* puts it in the following way [32.30]:

“It is necessary to remark that there is an ongoing synthesis of computation and communication into a unified process of information processing. Practical and theoretical advances are aimed at this synthesis and also use it as a tool for further development. Thus, we use the word computation in the sense of information processing as a whole. Better theoretical understanding of computers, networks, and other information processing systems will allow us to develop such systems to a higher level.”

The traditional mathematical theory of computation is the theory of algorithms. Ideal, theoretical computers are mathematical objects and they are equivalent to algorithms, or abstract automata (Turing machines), or effective procedures, or recursive functions, or formal languages. New envisaged future computers are information-processing devices. That is what makes the difference. Syntactic mechanical symbol manipulation is replaced by information (both syntactic and semantic) processing. Compared to new computing paradigms, Turing machines form the proper subset of the set of information-processing devices, in much the same way as Newton’s theory of gravitation is a special case of Einstein’s theory, or the Euclidean geometry is a limit case of non-Euclidean geometries.

According to [32.30] there are three distinct components of information-processing systems: hardware (physical devices), software (programs that regulate its functioning), and infoware that represents information processed by the system. Infoware is a shell built around the software–hardware core, which was the traditional domain of automata and algorithm theory. Communication of information and knowledge takes place on the level of infoware.

*Peter J. Denning* comments on the relationship between computation and information with respect to the

classical Turing idea that computation is equivalent to *algorithm execution* [32.4]:

“First, some information processes are natural. Second, we do not know whether all natural information processes are produced by algorithms. The second statement challenges the traditional view that algorithms (and programming) are at the heart of computing. Information processes may be more fundamental than algorithms.”

*Floridi’s* philosophy of information, developed as [32.37]

“a new philosophical discipline, concerned with a) the critical investigation of the conceptual nature and basic principles of information, including its dynamics (especially computation and flow), utilization and sciences and b) the elaboration and application of information-theoretic and computational methodologies to philosophical problems.”

can be seen as parallel to the philosophy of computation as developed by *Cantwell Smith* in his *Origins of Objects* [32.18].

To better represent information processing in biological systems, computational modeling is applied with computation taken to be *distributed, massively concurrent, heterogeneous, and asynchronous*. *Dodig-Crnkovic* proposes to adopt *Hewitt et al.’s* actor model of computation [32.38, 39]. In this model, computation is the process of message passing between actors (agents) in an interacting network. *Hewitt* provides the following description [32.40, p. 161]:

“In the Actor Model, computation is conceived as distributed in space, where computational devices communicate asynchronously and the entire computation is not in any well-defined state. (An Actor can have information about other Actors that it has received in a message about what it was like when the message was sent.) Turing’s Model is a special case of the Actor Model.”

*Hewitt’s* computational systems consist of computational agents – informational structures capable of acting on their own behalf. Unlike logical model of Turing machine, *Hewitt’s* model of computation is inspired by physics [32.23].

When defining computation as information processing in a network of agents, those networks can consist of molecules or cells like bacteria or neurons, thus constituting networks of networks on the hierarchy of scales of informational structures with computational dynamics.

As we base the model of computation on the concept of information, it is in place to analyze the relationships between the two in more detail.

### 32.3 Computation Versus Information

In this section, we detail the concept of computation as information processing to later on make explicit its connections to cognition in the following sections.

Sometimes the current lack of the consensus about the definition of information is termed *scandal of information*. At the same time, we can talk even about the corresponding *scandal of computation*, that is, the current lack of consensus about the concept of computation. *Denning* describes historical development of the concept of computation [32.4] as it appears in different contexts and various communities of practice.

However, this situation is not as unique as it may seem. There is no simple commonly accepted definition of life, and yet biologists study it and make constant

progress in understanding its different aspects, forms, and processes. Actually no such unique definition of many among central concepts of sciences exists, and yet this situation is not experienced as a scandal. A fresh example is *dark energy*, which together with *dark matter* constitutes 95.1% of the content of the universe. We do not know what they actually are.

Instead of understanding a concept as equivalent to a (linear) string of symbols that constitute its definition, we can rather represent it as a network of relationships such as *is kind of*, *is a member of*, *is a part of*, *is a substance of*, *is similar to*, *pertains to*, *causes*, etc., which connect the concept with other concepts that are connected with still further concepts.

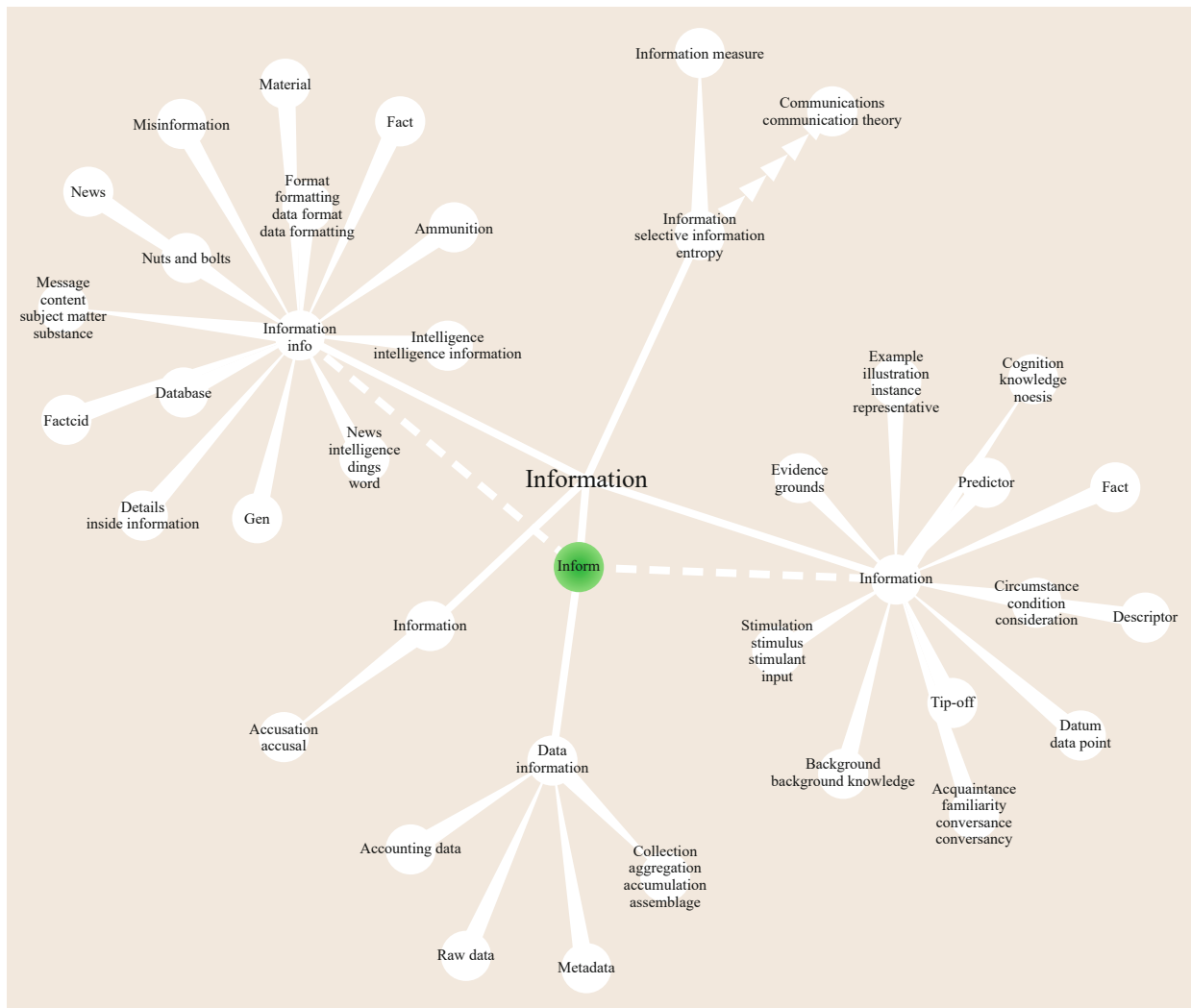


Fig. 32.1 Information concept (after [32.41])

This constitutes *family resemblance* (Familienähnlichkeit) of Ludwig Wittgenstein [32.42, §66] where

“There is no reason to look, as we have done traditionally – and dogmatically – for one, essential core in which the meaning of a word is located and which is, therefore, common to all uses of that word. We should, instead, travel with the word’s uses through ‘a complicated network of similarities overlapping and criss-crossing (PI §66)’”

Wittgenstein explains in [32.42, §67]:

“Why do we call something a *number*? Well, perhaps because it has a direct relationship with several things that have hitherto been called number; and this can be said to give it an indirect relationship to other things we call the same name. And we extend our concept of number as in spinning a thread we twist fibre on fibre. And the strength of the thread does not reside in the fact that someone fibre runs through its whole length, but in the overlapping of many fibres.”

As a result, boundaries dissolve [32.42, §68]:

“What still counts as a game and what no longer does? Can you give the boundary? No.”

Figures 32.1 and 32.2 illustrate how visualizations of the concepts *information* and *computation* may look like and where Wittgenstein’s idea of *family resemblance* becomes apparent.

Given that information/computation are common ideas spanning fields from physics, chemistry, biology, and theories of mind, to ecologies and social systems, the bridging over such a large range is achieved by a network of inter-related concepts (*family of concepts*) that enable us to traverse from field to field, the main point being keeping a common thread. That means that the concept is not a reductionist one, but networked rhizomatic idea in the sense of [32.43].

World today is seen as governed and controlled by natural laws that in-materio execute *programs* unlike the mechanistic world governed by fixed and unchangeable laws expressed by equations of Newtonian physics. Wolfram is arguing that “information processes underlie every natural process in the universe” [32.8].

The important difference is that while equations operate on *data*, programming languages can operate even on higher level *data structures* and even on *physical objects* such as in case of cyber-physical systems.

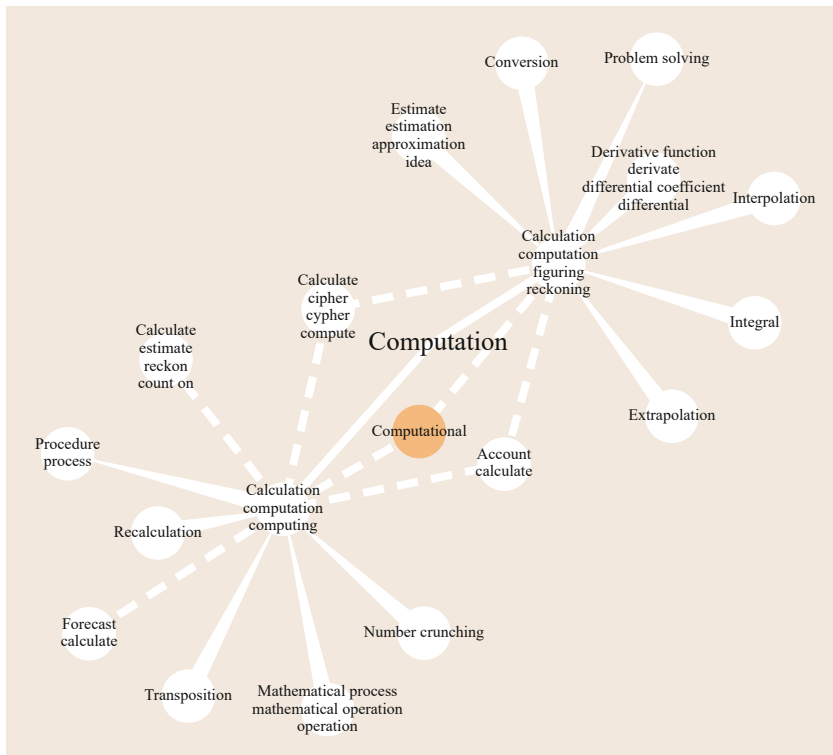


Fig. 32.2 Computation concept network (after [32.41])



## 32.4 The Difference Between Mathematical and Computational (Executable) Models

The belief in mathematical/geometrical essence of the world can be traced back to Plato and the Pythagoreans, which later on reappears with Galileo, in his 1632 *Dialogue Concerning the Two Chief World Systems*, where he argues that the book of nature is written in language of mathematics. Plato's ideal of eternal, unchangeable forms can be found in mathematics to this day. Even though mathematical formulas can be used to compute time-dependent processes, equations themselves are symbolic structures, persistent and immovable. Time dependency comes from human performing computation, actively using static structures of mathematical algorithms to trace time behaviors of real-world systems. Platonic ideal forms, however remote from the physical realizations and questions of finite material resources, were long considered to represent the true nature of the world, while changes were supposed to be something ephemeral, uninteresting, and too earthly for a mathematician or a scientist to bother about. Up to quite recently this detachment from the *real-time* aspects of the world was commonly taken for granted and justified. The change happened with computing machinery getting integrated with dynamically changing physical objects, such as in embedded systems technologies or process control, where real-time computation processes must match real-time behaviors of the physical environment. This situation radicalized even more with the mobile distributed information and communication technologies for which the *system dynamics* is the most prominent characteristics. Rapidly, eternal forms are becoming something remote and less noted. Everything is in the process of change, communication, timely response, and resource optimization, as this new world of embodied and embedded computation is physical in nature and thus substrate-dependent. The whole field of cyber-physical systems is emerging, at different levels of scale, from nano to macroscopic. In this decisive step from idealized abstract forms toward concrete material processes, computation has come close to the messy and complex world of concurrent and context-dependent processes in living beings [32.27]. One important shift is also in the role of an *observer* [32.4]:

“Computational expressions are not constrained to be outside the systems they represent. The possibility of self-reference makes for very powerful computational schemes based on recursive designs and executions and also for very powerful limitations on computing, such as the noncomputability

of halting problems. Self-reference is common in natural information processes; the cell, for example, contains its own blueprint.”

One important aspect of modeling is the direction of their generation – bottom up or top down. Mathematical models are typically top-down while computational models are frequently bottom-up or generative, described by *Wolfram* as a *new kind of science* [32.8]. Fields modeling living organisms like synthetic biology present challenge of bridging the gap between the two, enabling the circular motion from bottom up to top down and back.

*Barry S. Cooper* addresses this difference between mathematical and computational approaches [32.44] in his article detecting *the mathematician's bias* and the current return to embodied (physical, natural) computation.

Unlike pure mathematics, computing can provide modeling tools for biomolecular systems such the abstraction of molecule as computational agent in which a system of interacting molecules is modeled as a system of interacting computational agents [32.45]. *Petri nets, State charts, and the Pi-calculus*, developed for systems of interacting computations, can be successfully used for modeling of biomolecular systems, such as signaling, regulatory, and metabolic pathways and even multicellular processes. Processes, the basic interacting computational entities of these languages, have an internal state and interaction capabilities. *Process behavior is governed by reaction rules specifying the response to an input message based on its content and the state of the process*. The response can include state change, a change in interaction capabilities, and/or sending messages. Complex entities are described hierarchically [32.45].

Computer science distinguishes between two levels of description of a system: *specification* (what the system does) and *implementation* (how the system is built). Biological function of a biomolecular system emerges thus from the semantic equivalence between the low-level and high-level computational descriptions [32.45].

The difference between mathematical and computational models can be summarized as distinction between denotational and operational semantics models given by *Fisher and Henzinger* in the following [32.46]:

*Denotational (mathematical) models* present set of equations showing relationships between different quantities and their time changes. They are approximated numerically.

*Operational semantics models* are based on algorithms, which execution (computation) simulates the behavior of the system in time.

The two semantics have different roles – while ex-

ecutable algorithms connect directly to the physical process and can be used in interactions with them, denotational semantics is descriptive and can be used to reason about systems.

## 32.5 Computation in the Wild

### 32.5.1 Physical Computation – Computing Nature as Info-Computation

Some processes in nature that show high regularity and symmetry have fascinated people for millennia – intricate forms of crystals, shells, flowers, patterns on water or sand, etc. Those have been studied extensively in the literature [32.8, 47–50] as they by their regularities invite algorithmic models. Wolfram, for example, emphasizes the importance of simple generative computational models that in number of iterative steps produce intricate forms that resemble ones found in nature, such as in shells or plants. Forms and processes found in nature in recent years inspired the development of a field of *natural computing*. In the Handbook of Natural Computing [32.21], the following types of natural computation are presented: amorphous computing; molecular computation; RNA-guided DNA assembly; synthetic biology; evolutionary computation; DNA nanomachines; artificial immune systems; evolutionary computation; computational swarm intelligence; quantum computing; genetic programming; membrane computing; and neural networks.

According to [32.51]:

“Natural computing is the field of research that investigates models and computational techniques inspired by nature and, dually, attempts to understand the world around us in terms of information processing. [...] There is information processing in nature, and the natural sciences are already adapting by incorporating tools and concepts from computer science at a rapid pace. Conversely, a closer look at nature from the point of view of information processing can and will change what we mean by computation.”

James Crutchfield et al. [32.52] make a distinction between *designed computation* (that is what computer machinery performs) from *intrinsic computation* (that is all processes in nature that are inherently computational and that are used in computing machinery to compute on the basic hardware level). In that way, they are able to *argue that information processing in dynamical systems as well is computational in nature* –

a claim that earlier was strongly denied. Nature thus presents an implementation substrate for computation. Similar claim is presented in [32.53] that argues that all dynamics of information present some computation, on different levels of description. *Nir Fresco* [32.54] provides an argument that the explanatory frameworks of computationalism, connectionism, and dynamicism, contrary to frequent claims, are not mutually exclusive but rather mutually complementary approaches, suggesting the way for their integration under the assumption of physical computation.

One way to characterize physical computation is via morphology – defining computational processes as the dynamic of morphological structure. *Turing* made pioneering contributions in the field of morphogenesis [32.55] even though he did not think of a process itself as computation. Later on models of robotics systems are made using morphology of a robot body to perform *computation in materio* in that way providing solutions to passive walking robots or artificial muscles [32.56].

*Helmut Hauser, Rudolf M. Füchslin and Rolf Pfeifer* in their *Opinions and Outlooks on Morphological Computation* point out [32.57]:

“Morphological Computation is based on the observation that biological systems seem to carry out relevant computations with their morphology (physical body) in order to successfully interact with their environments. This can be observed in a whole range of systems and at many different scales. It has been studied in animals – e.g., while running, the functionality of coping with impact and slight unevenness in the ground is *delivered* by the shape of the legs and the damped elasticity of the muscle-tendon system – and plants, but it has also been observed at the cellular and even at the molecular level – as seen, for example, in spontaneous self-assembly.”

As can be observed in nature, living systems from the simplest unicellular to the most complex multicellular organisms are heterogeneous, distributed, massively concurrent, and largely asynchronous (in spite of certain common regularities like circadian rhythm

with oscillation of about 24 hours, which is found in animals, plants, fungi, and bacteria). *This kind of systems are hardest to cope with in conventional computing* [32.58]. The approach to modeling such systems suggested by *Luca Cardelli* is as “collectives of interacting stochastic automata, with each automaton representing a molecule that undergoes state transitions” [32.59]. *Aviv Regev* and *Ehud Shapiro* suggest that we should use lessons learned from modeling sequence and structure in biomolecular systems, which already have good computational models [32.45]:

“Sequence and structure research use computers and computerized databases to share, compare, criticize and correct scientific knowledge, to reach a consensus quickly and effectively. Why can’t the study of biomolecular systems make a similar computational leap? Both sequence and structure research have adopted good abstractions: *DNA-as-string* (a mathematical string is a finite sequence of symbols) and *protein-as-three-dimensional-labeled-graph*, respectively. Biomolecular systems research has yet to find a similarly successful one.”

### 32.5.2 New Computationalism. Nonsymbolic versus Symbolic Computation

*Greg Chaitin* argued in [32.60], in the tradition from Leibniz to  $\Omega$  number, epistemology can be seen as information theory. Reality for an agent is defined by information and its processes [32.61] where information processes for a cognizing agent proceed from deepest levels of cell-cognition, self-organized and with emergent properties from subsymbolic signal processing up to symbolic level of human (natural and formal) languages.

As presented in [32.23, pp. 1–22], it is often argued that computationalism is the opposite of connectionism and *that connectionist networks and dynamic systems are not computational*. This would imply that human mind, as network of processes resulting from the activity of human brain, cannot be adequately modeled computationally. However, if we define computation in a sense of natural computation, instead of symbol manipulation as in the Turing machine, it is obvious that processes in the physical substrate of the human brain are natural computation and consequently models of *connectionist networks and dynamical systems do correspond to computational processes*.

One of the central and long-standing controversies when it comes to understanding of computation in biological (cognitive) systems is the relationship between symbolic and subsymbolic computation, where symbol

manipulation is the way of classical Turing computation, while subsymbolic processes such as going on in dynamic systems are frequently not even considered as computing. *Andy Clark* argues convincingly for the necessity of both kinds of processes, subsymbolic and symbolic for human-level cognition [32.62]. Information is relative to a cognizing agent and what is information, symbolic, subsymbolic, continuous, or discrete is a question of level of description or agency. From the everyday experience of a human agent, water and air are continua. However, on the molecular level (thus from the perspective of molecular agency) water and air consist of discrete objects – molecules. Atomic nucleus is seen as a continuum in Bohr’s liquid drop model, while on the level of constituent nucleons, it is a discrete system. On the level of nucleons, we can see continuum but on the deeper level of their constituent quarks as agents, there is a discrete behavior. In general, what an agent registers as continuous or discrete depends on both the system one examines and the type of agent – its structures and ways of interaction. In the dynamic systems models, [32.63]:

“[T]he general idea is that cognition should be characterized as a continual coupling among brain, body, and environment that unfolds in real time, as opposed to the discrete time steps of digital computation. The emphasis of the dynamical approach is on how the brain/body/environment system as a whole changes in real time, and dynamics is proposed as the best framework for capturing that change. This is said to contrast with computation’s focus on *internal structure* i. e., its concern with the static organization of information processing and representational structure in a cognitive system.”

Computational modeling of cognitive processes requires computing tools that contain not only Turing Machine model but also physical computation on the level of biological substrate. That is also the claim made by *Matthias Scheutz* in the Epilogue of the book *Computationalism: New Directions* [32.64, p. 176] where he notices that:

“Today it seems clear, for example, that classical notions of computation alone cannot serve as foundations for a viable theory of the mind, especially in light of the real-world, real-time, embedded, embodied, situated, and interactive nature of minds, although they may well be adequate for a limited subset of mental processes (e.g., processes that participate in solving mathematical problems). Reservations about the classical conception of computation, however, do not automatically transfer and

apply to real-world computing systems. This fact is often ignored by opponents of computationalism, who construe the underlying notion of computation as that of Turing-machine computation.”

Classical computationalism was the view that the classical theory of computation (Turing-machine model, universal, and disembodied) might be enough to explain cognitive phenomena. New computationalism (natural computationalism) emphasizes that embodiment is essential and thus physical computation, hence natural computationalism. The view of Scheutz is supported by *Gerard O'Brien* [32.65] who is arguing that

“cognitive processes, are not governed by exceptionless, representation-level rules; they are instead the work of defeasible cognitive tendencies subserved by the non-linear dynamics of the brains neural networks.”

Dynamical characterization of the brain is consistent with the analog interpretation of connectionism. But dynamical systems theory is typically not considered to be a computational framework. *O'Brien* and *Opie* [32.66] thus search for an answer to the question how connectionist networks compute, and come with the following characterization:

“Connectionism was first considered as the opposed to the classical computational theory of mind. Yet, it is still considered by many that a satisfactory account of how connectionist networks compute is lacking. In recent years networks were much in focus and agent models as well so the number of those who cannot imagine computational networks has rapidly decreased.”

As in classical computationalism only symbolic computation was taken into account, it is important to understand the connection between symbolic and sub-symbolic information processing [32.67, p. 119]:

“Symbolic simulation is thus a two-stage affair: first the mapping of inference structure of the theory onto hardware states which defines symbolic computation; second, the mapping of inference structure of the theory onto hardware states which (under appropriate conditions) qualifies the processing as

a symbolic simulation. Analog simulation, in contrast, is defined by a single mapping from causal relations among elements of the simulation to causal relations among elements of the simulated phenomenon.”

Both symbolic and subsymbolic (analog) simulations depend on causal/analog/physical and symbolic type of computation on some level but *in the case of symbolic computation it is the symbolic level where information processing is observed*. Similarly, even though in the analog model symbolic representation exists at some high level of abstraction, it is the *physical agency of the substrate and its causal structure that define computation* (simulation).

*Gianfranco Basti* [32.68] suggests how to

“integrate in one only formalism the physical (*natural*) realm, with the *logical-mathematical* (*computation*), studying their relationships. That is, the passage from the realm of the *causal necessity* (*natural*) of the physical processes, to the realm of the *logical necessity* (*computational*), and eventually representing them either in a sub-symbolic, or in a symbolic form. This foundational task can be performed, by the discipline of *theoretical formal ontology*.”

*Walter Jackson Freeman* offers an accurate characterization of the relationship between physical/subsymbolic and logical/symbolic level in the following passage [32.69]:

“The symbols are outside the brain. Inside the brains, the construction is effected by spatiotemporal patterns of neural activity that are operators, not symbols. The operations include formation of sequences of neural activity patterns that we observe by their electrical signs. [...] Neural operators implement non-symbolic communication of internal states by all mammals, including humans, through intentional actions. [...] I propose that symbol-making operators evolved from neural mechanisms of intentional action by modification of non-symbolic operators.”

Subsequently, brain uses internal subsymbolic computing to manipulate relevant external objects/symbols.

## 32.6 Cognition: Knowledge Generation by Computation of New Information

In the framework of *Humberto Maturana* and *Francisco Varela*, cognition is capacity of all living beings, no matter how small or simple [32.70]. It is characteristics of organisms that increase in complexity from the simplest ones such as bacteria, to the most complex forms of cognition found in humans. Maturana and Varela's view is gaining substantial support through the study of cognitive capacities of bacteria [32.71–79] and others. Social cognition has been studied in bacterial colonies, swarms, and films. *Lorenzo Magnani* and *Emanuele Bardone* summarize current findings in the following [32.80]:

“[A]ll organisms, including bacteria, are able to perform elementary cognitive functions because they *sense* the environment and process internal information for ‘thriving on latent information embedded in the complexity of their environment’ (Ben Jacob, Shapira, and Tauber, 2006) p. 496.”

In light of the contemporary research results, the earlier completely dominating, and to this day still widespread view of cognition as an exclusively human capacity occurs as a gross simplification. The more we learn about the ways living organisms cope with their environments, the more we understand that even the simplest organisms exhibit cognitive behaviors, that is, adaptive information processing increasing their probability of survival. According to *James Alan Shapiro* [32.78]:

“bacteria utilize sophisticated mechanisms for intercellular communication and even have the ability to commandeer the basic cell biology of *higher* plants and animals to meet their own needs.”

As an example of the level at which the subtleties of bacterial cognitive behavior is known, we refer to *Stephan Schauder* and *Bonni L. Bassler* who reveal the specifics of bacterial communication and quorum sensing both within and between species of bacteria [32.75]:

“Bacteria communicate with one another using chemical signaling molecules as words. Specifically, they release, detect, and respond to the accumulation of these molecules, which are called autoinducers. Detection of autoinducers allows bacteria to distinguish between low and high cell population density, and to control gene expression in response to changes in cell number. This process, termed quorum sensing, allows a population of bacteria to coordinately control the gene expression of the entire community.”

Insights into the cognitive processes of bacteria have very far-reaching consequences for our understanding of life and cognition [32.78]:

“[T]he recognition of sophisticated information processing capacities in prokaryotic cells represents another step away from the anthropocentric view of the universe that dominated pre-scientific thinking. Not only are we no longer at the physical center of the universe; our status as the only sentient beings on the planet is dissolving as we learn more about how smart even the smallest living cells can be.”

Regarding information processing (computation) in bacteria, [32.73] emphasizes that bacterial information processing differs from the Turing machine model of computation. Unlike the Turing machine, in a bacterial colony, in response to an input (signals or molecules from the environment), *hardware* (physical system, bacteria) changes through information processing resulting in a new configuration/form plus possibly some output in signals/molecules. Bacteria typically exchange molecules as information, and it might also be exchange of genetic material. This type of computation is example of *Kampis' component systems* [32.26] and presents *physical computation* [32.54, 81–83]. Yet another take on physical processes behind living agency and its evolution is elaborated by *Terrence Deacon* [32.84]. Even though Deacon himself is not a computationalist, models he develops can be understood as mechanistic and interpreted as computation. For [32.85], bacterial cognition is a case of interactive biological (hyper)computation, that is, computation beyond Turing machine model.

According to [32.86] dynamical systems [32.87], analog neural networks [32.88] and oracle Turing machines [32.89] have in common that they introduce elements that are not Turing computable, that is, they introduce *hyper-computation*. *Bournez* and *Cosnard* compare capabilities of discrete versus dynamical systems and conclude that “many dynamical systems have intrinsically super-Turing capabilities.” Models of hypercomputation or super-Turing computation models of biological systems are studied in [32.90].

Advancing computational models of cognition and bridging the gap between bacterial and human cognition calls for studies of cognition in other living organisms. Even though human cognition is usually superior to animal and plant cognition, it is not always the case. For example, many animals have superior senses like vision, hearing, far better motoric skills, and some

of them like chimpanzees can beat humans in certain memory tasks [32.91]

“Young chimpanzees have an extraordinary working memory capability for numerical recollection – better even than that of human adults tested in the same apparatus following the same procedure.”

Understanding cognition in other organisms has extraordinary value in understanding mechanisms of cognition and their evolution. For example study of cognition in fish can help us find ecological factors that affect the evolution of particular cognitive abilities. One can study the relationship between size of specific brain areas and cognitive abilities and the stages in the development of decision abilities [32.92]. Animal studies can be used for tracing evolution of cognitive capacities, and quantitatively testing possible correlations between certain cognitive abilities and life history, morphology, or socioecological variables, measure if phylogenetic similarity corresponds to the cognitive skills throughout species, etc. [32.93]. However, traditional and even to this day predominant view is that only humans possess cognition. As a consequence, cognitive science developed vast majority of its models and theories exclusively about human cognition.

### 32.6.1 Distributed Cognition and Model-Based Reasoning

Even though animals (including birds) use tools for different purposes, human intelligence is defined as “faculty to create artificial objects, in particular tools to make tools” [32.94]. In the embodied interaction with the environment humans are “engaged in a process of *cognitive niche construction*” as they delegate certain cognitive functions to the environment [32.80]:

“In this sense, we argue that a cognitive niche emerges from a network of continuous interplays between individuals and the environment, in which people alter and modify the environment by mimetically externalizing fleeting thoughts, private ideas, etc., into external supports. [...] Artifactual cognitive objects and devices extend, modify, or substitute *natural* affordances actively providing humans and many animals with new opportunities for action.”

Underlying computational cognitive mechanisms enable the process of model construction. If we take *knowledge* to include not only the propositional knowledge (knowledge that) but also nonpropositional knowledge (knowledge how), we can say that bacteria *know how* to find food and avoid dangers in the environment.

According to [32.95]:

“Knowledge generation can be naturalized by adopting computational model of cognition and evolutionary approach. In this framework knowledge is seen as a result of the structuring of input data (data → information → knowledge) by an interactive computational process going on in the agent during the adaptive interplay with the environment, which clearly presents developmental advantage by increasing agent’s ability to cope with the situation dynamics.”

Scientific knowledge is obviously human knowledge and it includes both propositional (typically theoretical) and nonpropositional (typically practical) knowledge. *Ronald N. Giere* suggests that models in science are best understood “as being components of distributed cognitive systems,” where the process of scientific cognition “is distributed between a person and an external representation” [32.96].

The idea of distributed cognition can be traced back to *David Rumelhart* and *James McClelland* [32.97], and it has been developed during the years in a number of prominent works such as [32.98, 99, 99–102].

The related idea termed *extended mind* has been proposed by *Andy Clark* and *David Chalmers* in [32.103] meaning that humans use tools and other suitable objects in the environment to perform cognitive tasks. *Enactivism* is a connected movement in the philosophy of mind whose proponents argue that we should understand mental abilities as essentially related to the extended body and to action [32.104].

In the study of the capabilities of networks of simple processors, *Rumelhart* and *McClelland* [32.97] found that they are good at recognizing patterns in the input. The generalization to human brain is that it recognizes patterns through the activation of changes in the states of neurons induced by sensory inputs. *Rumelhart* and *McClelland* suggest that “humans do the kind of cognitive processing required for these linear activities by creating and manipulating external representations.”

In a distributed cognitive system information processing happens through parallel distributed processing (PDP). In this view “the regular or law-like behavior of a complex system is the consequence of interactions among constituent elements” [32.105]. The main ideas of PDP models – such as that cognitive functions arise from neural mechanisms, representations are distributed, cognitive development is driven by learning, cognitive structure is quasi-regular, behavior is sensitive to multiple ordered constraints, processing is ordered and continuous – are now standard assumptions in many research domains.

As a consequence of distributed cognition language is seen as a means of socially distributed cognition that supports human communication [32.106]:

“There is no *language of thought*. Rather, thinking in language is a manifestation of a pattern matching brain trained on external linguistic structures (see also [32.107]).”

This distributed view of language implies that “cognition is not only embodied, but also embedded in a society and in a historically developed culture” [32.101].

Even here, as in the case of *language games* we see social aspect of cognitive artifacts. In the same way as a concept is a node in a network of related concepts, connected with several types of relationships, distributed cognition in general is a network that in principle can extend indefinitely. What we consider to be relevant depends on the agent and the context.

The computational model of cognition is closely related to the idea of agency. An agent in this context is defined as an entity capable of acting on its own behalf. Agent models are especially suitable for modeling of distributed cognitive systems and they are used for study of adaptive behavior and learning. Among new trends in modeling there are generative agent-based models, where complexity of a system results from the time development of the interactive behavior of simple constitutive parts (such as swarm). Especially successful are new network models of complex systems [32.108] where the focus is on the properties of various network structures in an emerging network science. Commenting on the current development of computing, [32.109] declare:

“Multiplicities, flux, materialities, heterogeneities, and co-construction are features that are becoming increasingly evident within new configurations of computing.”

Those features are particularly suitable in modeling of living (cognitive) systems.

### 32.6.2 Computational Aspects of Model-Based Reasoning in Science

Nancy Nersessian [32.110] searches for “the cognitive basis of model-based reasoning in science,” especially for model-based creative reasoning that results in “representational change across the sciences,” thus investigating the central issue of creativity in science asking “how are genuinely novel scientific representations created, given that their construction must begin with existing representations?”

Nersessian addresses methodological issues in scientific cognition, and the nature of model-based reasoning in science in order to give an account of their cognitive basis and the role they play in representational change. She introduced the term *model-based reasoning* to denote the construction and manipulation of representations, both sentential, and those related to external mediators [32.111, 112]. Model-based reasoning is applied to among others thought experiments, visual representations, and in analogical reasoning [32.113, 114].

As Giere [32.115] emphasizes, models are not only tools but they also play a central role in the construction of knowledge. “Models are important, not as expressions of belief, but as vehicles for exploring the implications of ideas (McClelland 2010)” [32.105]. One of the ways of acquiring knowledge besides deduction and induction is abduction that leads to knowledge discovery. Along with sentential and model-based theoretical abduction, [32.116] identifies *manipulative abduction* as thinking and learning through doing. Manipulative abduction is thus situated in the domain of extended cognition and presents an extra-theoretical behavior developed through manipulation of artifacts, such as written notes, diagrams, experimental set-ups, visual and other simulations, etc. One of the illustrative examples of extended cognition is diagrammatic reasoning [32.117, 118], [32.119]:

“What is interesting about diagrammatic reasoning is the interaction between the diagram and a human with a fundamentally pattern-matching brain. Rather than locating all the cognition in the human brain, one locates it in the system consisting of a human together with a diagram. It is this system that performs the cognitive task, for example, proving the Pythagorean theorem.”

Giere provides further examples of reasoning with pictorial representations and reasoning with physical and abstract models [32.96, p. 237]

“The idea of distributed cognition is typically associated with the thesis that cognition is embodied. In more standard terms, one cannot abstract the cognition away from its physical implementation.”

This agrees with *Fresco's* conclusions from his book *Physical Computation and Cognitive Science* [32.54].

Based on the idea that “*a complex system, as the cognitive one, and its transformations, can be described in terms of a configurational structure.*” [32.120], morphodynamical abduction is then the abduction expressed through the geometrical framework of configurational structure. *Magnani* in [32.121] explains that

“different mental states are defined by their geometrical relationships within a larger dynamical environment. This suggests that the system, in any given instant, possesses a general morphology we can study by observing how it changes and develops.”

This suggests the possibility of representing aspects of abductive reasoning in the framework of dynamical systems, as processes of natural computation.

In [32.121] the authors argue that:

“Creative and selective abduction can be viewed as a kind of process related to the transformations of the attractors responsible of the cognitive system behavior. In the context of naturalized phenomenology we have described anticipation and abduction

in the light of catastrophic rearrangement of attractors.”

This insight about the character of major qualitative shifts in understanding can be extended to aspects of scientific discovery as *Thagard* addressed for conceptual change and scientific revolutions [32.122].

In the context of model-based reasoning it is instructive to take software as an example of executable computational models that enables transformations from system requirements to design and implementation. Software is used as a cognitive tool of extended cognition, which facilitates cognitive information processing by automation of reasoning, performing cognitive tasks of computation, search, control, etc., that already has resulted in radically new conceptualizations such as cyber-physical systems and Internet of things.

## 32.7 Model-Based Reasoning and Computational Automation of Reasoning

Models have been historically used in scientific and engineering disciplines to handle large-scale, complex research and development enterprises. Modeling as an inherent human skill is tightly coupled with the use of natural language for communication [32.123]:

“Though we share the same earth with millions of kinds of living creatures, we also live in a world that no other species has access to. We inhabit a world full of abstractions, impossibilities, and paradoxes. We alone brood about what didn’t happen, and spend a large part of each day musing about the way things could have been if events had transpired differently.”

Models are typically exploited to design solutions as a preliminary step of the realization of an engineering project: the underlying aim is to anticipate relevant properties, design pitfalls, and other pertinent information that would be extremely expensive if discovered at advanced stages of a project. In modern times, none would build a bridge or a house without a model, at least to know in advance their expected mechanical properties and costs and time for completion, just to mention a few relevant parameters. The fact that most of the techniques exploit pictorial representations has to be directly interconnected to what is said so far, that is, modeling is an intrinsic human activity and our brain is exceptionally skilled in matching images with concepts, as argued in [32.117, 118].

Nowadays at the heart of scientific and engineering projects is software that is largely responsible for its function. The growing adoption of computational

devices and tools controlled by software in all aspects of human’s everyday life and its exploitation in mission-critical tasks inevitably contributed to the enormous magnification of nowadays software complexity. If the adoption of more and more advanced programming languages and techniques (first, second, third, and even fourth generation) can be conceived as progressive steps toward more abstract methods to develop software, they still meet the obstacles of code-centric approaches [32.124]:

“Conventional programming involves the manual translation of domain concepts living in the programmer’s head into a source program written in a concrete, general-purpose programming language such as C++ or Java. This translation inevitably causes important design information to be lost due to the semantic gap between the domain concepts and the language mechanisms available in the programming language. [...] Today, most designs are expressed in a concrete programming language meaning that the larger share of the design information is lost.”

In this respect, one of the main drawbacks of code is a limited understandability of the produced artifacts (even by the authors themselves), which seriously hinders the maintainability of the system. By considering that a complex software systems can be made up of millions of lines of code, have a life span of decades, and often merge in the solution expertise from diverse domains, it is clear that without proper modeling, the system will quickly run out of control, if it would be re-



alizable at all. Therefore, even though programming can be seen as a form of modeling of a system and its behavior, here we will refer to models as higher abstraction level representations, very often pictorial, of a software system [32.125]:

“Modeling, in the broadest sense, is the cost-effective use of something in place of something else for some cognitive purpose. It allows us to use something that is simpler, safer or cheaper than reality instead of reality for some purpose. A model represents reality for the given purpose; the model is an abstraction of reality in the sense that it cannot represent all aspects of reality. This allows us to deal with the world in a simplified manner, avoiding the complexity, danger and irreversibility of reality.”

From the 1980s, modeling techniques in software engineering have become more and more widespread. Their initial exploitation was confined to sole documentation/communication purposes, while from the late 1990s they have been progressively adopted to provide some form of automation. Interestingly, computing is perhaps the only domain (except for logic) that has *meta-products*, that is, it realizes products that are developed by means of the same methods used to realize the design/modeling tools (unlike, e.g., civil engineering where physics and computer graphics are exploited to model the final physical product, houses and bridges). It is not surprising that modeling techniques can be conceived as a derived product of description logics (DL), that is, a family of formal languages used for knowledge representation more expressive than propositional logic. However, as [32.126] argues, “these diagrammatic modeling languages provide no extensional, mathematical semantics, nor any automated reasoning facilities.”

The reasons of an evolution toward automated mechanisms are often connected with consistency problems: if a software system is documented through models and then implemented by hand, none can guarantee that what was coded is compliant with what was designed. Even more important, in future maintenance activities it will be very difficult to keep the consistency between models and code. These issues find their reasons in the gap between domain-specific description of the problem and programming language encoding of a solution, as interpreted by the programmer and narrowed down by the constructs available in the language (see the notion of agency explained later on). As a consequence, a paradoxical situation will appear in which modeling artifacts have been introduced as an additional effort that, however, leaves the initial problem unsolved – aiding in the maintenance of complex systems.

Modeling languages like the UML (unified modelling language), (object management group (OMG) [32.127]) are formal enough to be used for automated reasoning. They exploit pictorial, or better, diagrammatic representations of a system to document its structure and functions. They are typical examples of previously mentioned diagrammatic reasoning where diagrams are objects supporting distributed cognition. With UML being formal here we mean that models conforming, for example, to the UML have to adhere to a well-defined set of rules and constraints to produce models. Being more precise, a model has to adhere to a set of constraints defined by means of a metamodel, that is, a modeling language definition.

The most common way of defining a modeling language is to specify concepts and relationships between them. In general, language elements are interconnected through associations with multiplicities restricting possible source(s) and target(s) of a certain relation. Moreover, elements can contain other elements, and elements can specialize other elements. If a model adheres to all the production rules defined in the metamodel, it is said to conform to the metamodel. The metamodel itself has to adhere to precise rules to specify a language (like those we informally listed so far), which are typically referred to as the meta-metamodel. The abstraction hierarchy is limited at this point by prescribing that the meta-metamodel is defined by itself.

In order to improve the usability and readability of models, very often language concepts are not directly shown to users. On the contrary, they are in general amalgamated into information units and rendered in a more compact/intuitive way. Such utility elements, typically called *syntactic sugar* in programming languages terminology, introduce the distinction between abstract and concrete syntax of models. The former defines the structure of the model in terms of the metamodel it conforms to, while the latter represents how the model is rendered to the user. In this respect, the metamodel defines conformance rules in terms of the abstract syntax, while a certain abstract syntax can give place to multiple concrete syntaxes.

Modeling languages can be divided into two main categories: general purpose (GPL) and domain-specific (DSL). The former category refers to languages that have not been defined with a particular application domain in mind, and hence exploit general concepts to represent a certain real-life phenomena in abstract terms. The latter category instead is typically built bottom-up and exploits the vocabulary specific of a certain applicative field. When linking those categories to model-based reasoning, it is intuitively easier to grasp the *message* carried by a model when produced by a DSL, since it exploits concepts specific of a certain domain, very

often with the help of an adequate (pictorial) concrete syntax. On the contrary, GPLs' models are less intuitive since they re-encode certain phenomena through generic modeling concepts and their concrete syntax has to exploit very general modeling elements. However, the specificity of DSLs is also their Achilles' heel: whenever concepts need to be added and/or refined vast revisions can be required, in general to change users' rendering of elements, attached semantics, and so forth. For example, for a tiny language supporting the graphical definition of mathematical expressions, if we consider elementary school usage we could define the language taking into account only natural numbers. As soon as we want to extend the language to support real numbers everything needs to be revised: notably, negative or rational numbers not only need a different representation, but also enable additional operations.

The concepts exploited to define a language *implicitly establish ontology of an idealized user using models for a certain purpose*. In this respect, the specification of concepts, their relationships, and their graphical rendering through a concrete syntax are all conceived with this goal in mind. In the case of GPLs, ontology is indirectly derived as mapping between generic concepts and corresponding domain-specific items. Notably, the mathematical language mentioned above could be represented as a *package* containing four *classes*, each one identifying an arithmetical operation, and hence called *Addition*, *Subtraction*, *Multiplication*, and *Division*, respectively. In turn, each class would contain *attributes*, called, for example, *operand\_1*, *operand\_2*, and *result*. Moreover, each class would contain a *method do\_calculation* performing the appropriate operation on the operands and returning the result. It is worth noticing that in this case the GPL offers concepts like package, class, and so forth, by which it is encoding reality. On the contrary, a DSL could define a *Calculator* including the four operations, each of which requiring two input parameters and returning an output. If either exploiting a GPL or using a DSL, it is supposed that the user has a clear idea of the correspondence between the sign + and the addition operation, and so forth. Even more important, the user should make correct use of operands putting them in the correct order, especially when performing subtractions or divisions.

When talking about model-based reasoning we usually refer to semantics, that is, the meaning that is associated with concepts and relationships constituting a model [32.128]:

“a language consists of a syntactic notation (syntax), which is a possibly infinite set of legal elements, together with the meaning of those elements, which is expressed by relating the syntax to a se-

mantic domain. Thus, any language definition must consist of the syntax semantic domain and semantic mapping from the syntactic elements to the semantic domain.”

Semantics is distinguished in two main categories, namely structural and behavioral semantics. The former is based on the metamodel definition itself, while the latter focuses on the execution of models conforming to a given metamodel. Therefore, for instance if we consider again the mathematical expression language, we could define the subtraction operator as taking two natural numbers and producing one natural number. At this point we can also impose that the first input shall be greater than the second. Despite that we did not specify what a user would do with a model conforming to such a metamodel, it is quite intuitive that we are defining semantics by the metamodel, since we are narrowing down inputs and output to natural numbers (and we are also constraining the maximum number of inputs to two). Moreover, from an ontological perspective the terms *subtraction*, *operand*, and *result* should be self-explanatory about the concepts provided by the language.

Behavioral semantics adds the dynamics of model elements to the structural part, or the definition of the process related to the ontology defined with the language. Such description could be done in informal ways, even by means of natural language (this is for instance done for the specification of large portions of the UML). Therefore again taking into account the math example, the semantics of subtraction concept would be defined as *The subtraction operation subtracts the value of operand\_2 from operand\_1*. The main issues with informal approaches are their proneness to ambiguous interpretations and the impossibility to exploit them for automation purposes (e.g., interpretation, analysis, execution). A second possibility to give semantics to metamodel concepts is by means of a behavioral sublanguage embedded in the language itself. In some cases the structural part can be decorated with behavioral descriptions (e.g., scripts in a specific programming language); in other cases some portions of the language can be devoted to model behavior (for instance, the UML specification includes behavioral diagrams like state machines. Moreover, there exist UML extensions enabling behavioral description through so-called action languages). The model-driven research area introduced model transformations as another possible alternative, described in the following [32.129]:

“the definition of the semantics of a language can be accomplished through the definition of a mapping between the language itself and another language with well-defined semantics such as Abstract State

Machines, Petri Nets, or rewriting logic. These semantic mappings between semantic domains are very useful not only to provide precise semantics to DSLs, but also to be able to simulate, analyze or reason about them using the logical and semantic framework available in the target domain. The advantage of using a model-driven approach is that these mappings can be defined in terms of model transformations.”

Here model transformations are automatic mechanisms mapping models toward other models as well as streams of characters. In this respect, a transformation performs a semantic anchoring between the structural concepts defined in the metamodel and the corresponding elements generated as target of the transformation execution. From an ontological perspective, the transformation in this case conveys the definition of the process.

## 32.8 Model Transformations and Semantics: Separation Between Semantics and Ontology

Historically, the semantics of a model has been given through corresponding constructs in a programming language, and in some cases modeling languages have been even conceived bottom-up as abstraction of a programming approach. The underlying reason is that for software experts it was easier and less ambiguous to interpret a model by means of the corresponding algorithm implementing what is designed. Moreover, software was mainly used for pure computational purposes that did not act directly on the real world. Therefore, semantics in software engineering had an *indirect* relationship to the real world since it was revealed in terms of *data structures* and *computation actions*.

More recently, embedded software systems are becoming ubiquitous and have been gradually replacing hardware functions, thus requiring a variety of modeling approaches. Nowadays, physical behavior is often mixed with software models since embedded systems directly interact with the world – often referred to as cyber-physical systems. In turn, semantics also evolved from pure computation to combination of computation and interactions with the physical world. For example, a model of a power window in a car contains both a state machine to show how buttons are related with the position of the window, and a physical model of the force the motor applies to move the window. If the force exceeds a certain value the motor has to be stopped and buttons deactivated, since an object could be obstructing window movements.

Another relevant aspect of semantics is that *a model could have more than one interpretation* depending on the user’s (both human and machine) point of view. In other words, a model may carry multifaceted information that could be equivalently used to both perform analysis and to generate code. Notably, in the power window example, models serve both to generate the code for commanding the motor through buttons, and to verify that the window will never harm passengers in its movements.

Model transformations are the means to manipulate models, both to generate other models and to derive textual documents (including code). They are related to model-based reasoning because the correspondences established through transformations draw interconnections among concepts pertaining to different domains. Therefore, they can be exploited to map the concepts expressed in a source model toward a corresponding target for which the semantics is well known (e.g., code generation gives a precise operational semantics to the source model). This notably allows performing more or less automated considerations about the quality attributes of a certain system and to refine its specification if needed.

Despite the existence of model transformations, modeling gives inherently place to *ambiguities* and *problems of interpretation*. In the most trivial case, the agent writing the transformation is not the modeler, opening already at that point space for erroneous interpretations. One of the relevant ambiguity problems is related to *multi-view based modeling*: given the complexity of nowadays systems, they are often modeled by means of different points-of-view that describe the system from corresponding domain-specific perspectives. In general these different views are not completely orthogonal, rather they *overlap and combine* for certain aspects. Such overlaps introduce interpretation problems due to the fact that the same concept may have different semantics depending of the point-of-view it is used in. Even more, the combination of concepts coming from different views could constitute a higher order concept that is not representable in any one of the single views separately.

Further important issue is related to the management of *evolution*, which is one of the peculiarities of software systems in general. When dealing with models and their evolution, some problems come into play due to the level of abstraction. First of all, it is difficult to compare two versions of the same model at differ-

ent points in time and recognize precisely the kind of evolution they have been subject to. Notably, it is very difficult to distinguish between the deletion of an element and subsequent addition of a new one versus the simple modification of the same element from the older toward a newer version. Even if this could appear as a minor aspect, it has relevant consequences. In fact, in the former case whatever was derived from the removed element should be deleted as well, whereas in the latter situation interconnected elements should be tracked and updated accordingly. In turn, the *propagation of updates* poses interpretation problems as well. By recalling the simple evolution example related to the subtraction operation (from natural to real numbers) mentioned before it is not difficult to foresee what kind of problems developers will incur when trying to reflect the impact of change on different related elements. Notably, if a concept in the language was renamed, the question is should its concrete syntax also be revised? How? And what about transformation mappings written on the base of the previous version of the modified concept?

What can be seen as a promising trend to alleviate the issues described above is the introduction of a separation between what have been called as *structural* and *ontological* aspects of a certain language. The structural aspects refer to generic syntactic rules that are invariant with respect to the information carried by the language. On the contrary, the ontological aspects are domain specific and are strictly related to a particular application context [32.130]. In general, structural aspects are simpler to manage, since there is no domain semantics involved. On the contrary, ontological aspects embed the semantics of a certain domain, and therefore typically require user guidance. If we take an example of natural language, structural aspects are grammar rules for sentence construction, while the ontological part is the argument discussed in the text. In this respect, a better denomination would be *informative* and *cognitive* aspects of a language, respectively. In such a way, it would not only be clearer that the latter aspects include semantics details, but it would also make it possible to introduce the notion of *agency and agent*, to locate possible multiple interpretations due to agency (something that gets mixed in the term ontology).

A separation between semantics and ontology, by introduction of the model of *agent-centered ontology*, is due to the fact that ontology (all that exists) *depends on to whom it exists* (see the later discussion). For example for humans, invisible microscopic layers of physical reality did not exist before the invention of microscopes; birds collide with windows because they do not see glass – for them it does not exist before the collision. For them much of human civilization does not

exist otherwise than as physical objects. The simpler an agent in the world is, the simpler is its ontology, that is, all that exists and can exist for that agent. Living agents self-organize at increasing number of layers of agency [32.131]. The core is the basic physical structure of elementary particles and forces from which new layers of organization emerge – from physics to chemistry, biology, and cognitive layer. Humans, on the top of hierarchy, have the highest known number of levels of organization. Starting with the *model of layered agency*, from the basic physical primitives to cognitive functions we can define semantics for an agent. For an agent it is characteristic that it is capable of *acting in the world*.

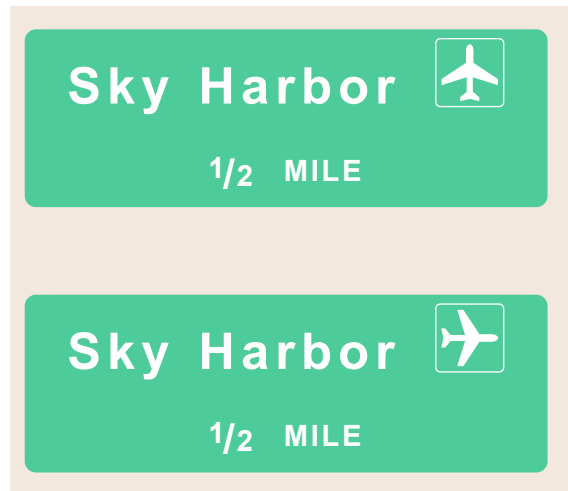
The most fundamental is *physical agency* where agents act completely automatically by simply obeying physical laws. Chemical agency is a level above and it builds on basic physical agency of elementary physical constituents, taken in bigger chunks of molecules that interact with each other. Biological agency emerges from chemical ones when chemical structures and cycles are established, which enable stable self-sustaining and self-reproducing formations [32.84] like first viruses (in between crystal and living organism) and bacteria as simplest cellular organisms. Symbol-manipulation-based agency arises first with organisms possessing nervous system, which helps them to model and predict their environment. Social-level distributed cognition emerges in networks of agents, from simplest living organisms to humans and anticipated intelligent machinery that is currently being developed. On the highest level of organization, there are languages shared by a community of users, used for coordinating actions and control of environment. Here even programming languages belong.

Software is based on programming languages, which constitute logical framework and syntactic rules. With respect to agency and agents in modeling for SE, some consideration is in place. From the earlier sections of this chapter, it is clear that our cognitive mechanisms are shaped by our embodied pattern recognition capability, which in turn has been trained through individual life and experiences as well as based on evolutionary developments of the species. In this respect, model users are agents – the ones who design the language to specify models, the ones that abstract a certain real-life phenomena through the model, and the ones who map the information carried by a model back to a corresponding real-life domain. Agency depends on the background experience of a user that influences the interpretation of concepts represented through models. Notably, domain experts may have a deeper knowledge of certain aspects of the system that would introduce

ambiguities in (or erroneous interpretation of) the models exploited to design systems in their domain (earlier mentioned in the example of the arithmetic modeling language).

As an illustration take the picture shown in Fig. 32.3. If we consider its content as an abstraction of reality, typical human agent would give the interpretation of it as two road signs, both indicating an airport in half a mile. However, there would be some agents also noticing that the shape of the plane resembles an arrow. This added detail couldn't be ignored, since it gives place for further interpretations of the figure. Did the designer intend to convey that to reach the airport the driver should follow the direction pointed by the plane? In other words, is the orientation of the airplane making a difference in driving direction? If the agent designing such a picture did not know anything about the usage scenarios of those signs, of course all those further interpretations would be the erroneous product of agency. In this respect, designing a language has intrinsic pitfalls due to the fact that the agents creating a modeling language most probably will not be the ones using it to model a certain system. In turn, the agents modeling the system will most probably differ from the ones exploiting those models for other activities (testing, implementation, documentation, etc.). As a consequence, models will be inherently ambiguous unless a precise semantics is given. This is one of the main reasons why model-based approaches can miss their potentials if code generation (or other forms of precise semantics) is not provided. To avoid that, automatic derivation of code from model establishes a well-defined semantics free from misinterpretation, and once the mapping between models and code is proved as correct, it would guarantee the consistency between design and implementation.

By considering the distinction between *informative* and *cognitive* aspects of a language it would be possible to alleviate the inherent problems due to ambiguities carried by models. By approaching the problem in a model-based way, it would be necessary to introduce a new language able to express agency, or analogously to define a precise *agent-based ontology* clarifying the level of expertise of a certain agent. In this respect however, the current approach of model-based techniques goes in the direction opposite to a natural process of abstraction. In fact, in the latter, simple components are interconnected to build up higher order cognitive



**Fig. 32.3** Two similar road signs with possible ambiguous interpretations

concepts, while in the former a language is first defined and subsequently the user tries to represent certain reality through the concepts offered by the language. Therefore, a more natural approach would be to let the user model reality in the desired way and incrementally build the underlying language based on the concepts exploited in the modeling activity.

Moreover, partial model transformations could be defined for better specifying some of the cognitive aspects included in a model. Google search engine can be mentioned as an example of current proposal. The engine has been enhanced through the years and currently, thanks to user tracking (which is a machine learning process) it has the capability of anticipating search queries of a given user (agent). It is often enough to type the first three/four characters of a word to get as suggestion the exact sentence one was thinking about. Moreover, with the same accuracy the search engine is able to suggest a potentially more relevant query based on the one written by the user. In the same way, a modeling language should be able to create a customized profile for each user, by learning from the users' modeling operations, a corresponding adequate ontology. In this way, the tool could anticipate some of the users' needs and aid them in correct modeling choices. In other words, the tool would be able to close the gap between ontology of the idealized user of a modeling language and the agent-based one.

## References

- 32.1 G. Dodig-Crnkovic: *Investigations into Information Semantics and Ethics of Computing* (Mälardalen Univ. Press, Västerås 2006)
- 32.2 C. Ess, R. Hagengruber (eds.): The computational turn: Past, presents, futures?, Proc. IACAP 2011 Conf. (Monsenstein Vannerdat, Münster 2011)
- 32.3 D.M. Berry: The computational turn: Thinking about the digital humanities, *Cult. Mach.* **12**, 1–22 (2011)
- 32.4 P. Denning: Structure and organization of computing. In: *Computing Handbook. Computer Science and Software Engineering*, ed. by T. Gonzalez, J. Diaz-Herrera, A. Tucker (Chapman Hall/CRC, Boca Raton 2014)
- 32.5 P. Denning, P. Rosenbloom: The fourth great domain of science, *ACM Commun.* **52**(9), 27–29 (2009)
- 32.6 S. Abramsky, B. Coecke: Physics from computer science: A position statement, *Int. J. Unconv. Comput.* **3**(3), 179–197 (2007)
- 32.7 P.J. Denning: The great principles of computing, *Am. Sci.* **98**, 369 (2010)
- 32.8 S. Wolfram: *A New Kind of Science* (Wolfram Media, Champaign 2002)
- 32.9 J.M. Wing: Computational thinking, *ACM Commun.* **49**(3), 33–35 (2006)
- 32.10 M. Burgin, G. Dodig-Crnkovic: A taxonomy of computation and information architecture, Proc. 2015 Eur. Conf. Softw. Archit. Workshops (ECSAW '15) (ACM, New York 2015), Article 7
- 32.11 G. Dodig-Crnkovic: Shifting the paradigm of philosophy of science: Philosophy of information and a new Renaissance, *Minds Mach.* **13**(4), 521–536 (2003)
- 32.12 M. Burgin: *Theory of Information: Fundamentality, Diversity and Unification* (World Scientific, Singapore 2010)
- 32.13 J. van Benthem, P. Adriaans: *Philosophy of Information* (North Holland, Amsterdam 2008)
- 32.14 J. van Benthem: Logic and the dynamics of information, *Minds Mach.* **13**(4), 503–519 (2003)
- 32.15 J. van Benthem: Logical pluralism meets logical dynamics?, *Australas. J. Logic* **6**, 182–209 (2008)
- 32.16 J. van Benthem: *Logical Dynamics of Information and Interaction* (Cambridge Univ. Press, Cambridge 2011)
- 32.17 P. Allo: Logical pluralism and semantic information, *J. Philos. Logic* **36**(6), 659–694 (2007)
- 32.18 B. Cantwell Smith: *On the Origin of Objects* (MIT Press, Cambridge 1998)
- 32.19 M. Tedre: *The Science of Computing: Shaping a Discipline* (CRC Press/Taylor Francis, Boca Raton 2014)
- 32.20 BusinessDictionary: <http://www.businessdictionary.com/definition/computing.html#ixzz3NZLB7QAD> (2014), visited 02.08.2016
- 32.21 G. Rozenberg, T. Bäck, J.N. Kok (Eds.): *Handbook of Natural Computing* (Springer, Berlin, Heidelberg 2012)
- 32.22 H. Zenil (Ed.): *A Computable Universe. Understanding Computation & Exploring Nature As Computation* (World Scientific/Imperial College Press, Singapore 2012)
- 32.23 G. Dodig-Crnkovic, R. Giovagnoli: *Computing Nature* (Springer, Berlin, Heidelberg 2013)
- 32.24 A. Church: Abstract No. 204, *Bull. Amer. Math. Soc.* **41**, 332–333 (1935)
- 32.25 A. Church: An unsolvable problem of elementary number theory, *Amer. J. Math.* **58**, 354 (1936)
- 32.26 G. Kampis: *Self-Modifying Systems in Biology and Cognitive Science: A New Framework for Dynamics, Information, and Complexity* (Pergamon Press, Amsterdam 1991)
- 32.27 S. Navlakha, Z. Bar-Joseph: Distributed information processing in biological and computational systems, *Commun. ACM* **58**(1), 94–102 (2015)
- 32.28 G. Dodig-Crnkovic: Significance of models of computation from Turing model to natural computation, *Minds Mach.* **21**(2), 301–322 (2011)
- 32.29 E. Eberbach, D. Goldin, P. Wegner: Turing's ideas and models of computation. In: *Alan Turing: Life and Legacy of a Great Thinker*, ed. by C. Teuscher (Springer, Berlin, Heidelberg 2004) pp. 159–194
- 32.30 M. Burgin: *Super-Recursive Algorithms* (Springer, New York 2005)
- 32.31 D. Baltimore: How biology became an information science. In: *The Invisible Future*, ed. by P. Denning (McGraw-Hill, New York 2001) pp. 43–56
- 32.32 S.B. Cooper, J. van Leeuwen: *Alan Turing. His Work and Impact* (Elsevier Science, Amsterdam 2013)
- 32.33 M. Burgin, G. Dodig-Crnkovic: From the closed classical algorithmic universe to an open world of algorithmic constellations. In: *Computing Nature*, ed. by G. Dodig-Crnkovic, R. Giovagnoli (Springer, Berlin, Heidelberg 2013) pp. 241–253
- 32.34 L. Floridi: Open problems in the philosophy of information, *Metaphilosophy* **35**(4), 554–582 (2004)
- 32.35 G. Dodig-Crnkovic, S. Stuart: *Computation, Information, Cognition: The Nexus and the Liminal* (Cambridge Scholars Pub., Newcastle 2007)
- 32.36 G. Dodig-Crnkovic, V. Müller: A dialogue concerning two world systems: Info-computational vs. mechanistic. In: *Information and Computation*, ed. by G. Dodig-Crnkovic, M. Burgin (World Scientific, Singapore 2011) pp. 149–184
- 32.37 L. Floridi: What is the philosophy of information?, *Metaphilosophy* **33**(1/2), 123–145 (2002)
- 32.38 C. Hewitt, P. Bishop, P. Steiger: A universal modular ACTOR formalism for artificial intelligence, Proc. 3rd Int. Joint Conf. Artif. Intell. IJCAI, ed. by N.J. Nilsson (William Kaufmann, Standford 1973) pp. 235–245
- 32.39 C. Hewitt: Actor model for discretionary, adaptive concurrency, *CoRR* (2010) <http://arxiv.org/abs/1008.1459>
- 32.40 C. Hewitt: What is computation? Actor model versus Turing's model. In: *A Computable Universe*,

- Understanding Computation & Exploring Nature As Computation*, ed. by H. Zenil (World Scientific/Imperial College Press, London 2012)
- 32.41 Visuwords online graphical dictionary: <http://www.visuwords.com>, visited 02.08.2016
- 32.42 L. Wittgenstein: *Philosophical Investigations*, transl. by G.E.M. Anscombe, R. Rhees (Blackwell, Oxford 2001)
- 32.43 G. Deleuze, F. Guattari: *A Thousand Plateaus: Capitalism and Schizophrenia* (Univ. Minnesota Press, Minneapolis 2005)
- 32.44 B. Cooper: The mathematician's bias – and the return to embodied computation. In: *A Computable Universe, Understanding Computation & Exploring Nature As Computation*, ed. by H. Zenil (World Scientific/Imperial College Press, Singapore 2012) p. 125
- 32.45 A. Regev, E. Shapiro: Cellular abstractions: Cells as computation, *Nature* **419**, 343 (2002)
- 32.46 J. Fisher, T.A. Henzinger: Executable cell biology, *Nat. Biotechnol.* **25**(11), 1239–1249 (2007)
- 32.47 D. Thompson: *On Growth and Form* (Cambridge Univ. Press, Cambridge 1961)
- 32.48 A. Seilacher, A.D. Gishlick: *Morphodynamics* (CRC, Boca Raton 2014)
- 32.49 R.V. Jean: *Phyllotaxis: A Systematic Study in Plant Morphogenesis* (Cambridge Univ. Press, New York 1994)
- 32.50 P. Prusinkiewicz, A. Lindenmayer: *The Algorithmic Beauty of Plants* (Springer, Berlin, Heidelberg 1990)
- 32.51 G. Rozenberg, L. Kari: The many facets of natural computing, *Commun. ACM* **51**, 72–83 (2008)
- 32.52 J. Crutchfield, W. Ditto, S. Sinha: Introduction to focus issue: Intrinsic and designed computation: Information processing in dynamical systems – Beyond the digital hegemony, *Chaos* **20** (2010) doi:<http://dx.doi.org/10.1063/1.3492712>
- 32.53 G. Dodig-Crnkovic: Dynamics of information as natural computation, *Information* **2**(3), 460–477 (2011)
- 32.54 N. Fresco: *Physical Computation and Cognitive Science* (Springer, Berlin, Heidelberg 2014)
- 32.55 A.M. Turing: The chemical basis of morphogenesis, *Philos. Trans. R. Soc. Lond.* **237**(641), 37–72 (1952)
- 32.56 R. Pfeifer, F. Iida: Morphological computation: Connecting body, brain and environment, *Jpn. Sci. Mon.* **58**(2), 48–54 (2005)
- 32.57 H. Hauser, R.M. Füchslin, R. Pfeifer: *Opinions and Outlooks on Morphological Computation* (Univ. Zurich, Zurich 2014), e-book <http://www.merlin.uzh.ch/publication/show/10528>
- 32.58 L. Cardelli: Abstract machines of systems biology, *Bull. EATCS* **93**, 176–204 (2007)
- 32.59 L. Cardelli: Artificial biochemistry. In: *Algorithmic Bioprocesses*, ed. by A. Condon, D. Harel, J.N. Kok, A. Salomaa, E. Winfree (Springer, Heidelberg 2009) pp. 429–462
- 32.60 G. Chaitin: Epistemology as information theory: From Leibniz to  $\Omega$ . In: *Computation, Information, Cognition – The Nexus and The Liminal*, ed. by G. Dodig-Crnkovic (Cambridge Scholars Pub., Newcastle 2007) pp. 2–17
- 32.61 G. Dodig-Crnkovic: Modeling life as cognitive info-computation. In: *Computability in Europe 2014*. LNCS, ed. by A. Beckmann, E. Csuhaj-Varjú, K. Meer (Springer, Berlin, Heidelberg 2014) pp. 153–162
- 32.62 A. Clark: *Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing* (MIT Press, Cambridge 1989)
- 32.63 M. Mitchell: A complex-systems perspective on the 'computation vs. dynamics' debate in cognitive science, *Proc. 20th Annu. Conf. Cogn. Sci. Soc. Cogsci98*, ed. by M.A. Gernsbacher, S.J. Derry (1998) pp. 710–715
- 32.64 M. Scheutz: *Computationalism new Directions* (MIT Press, Cambridge 2002)
- 32.65 G. O'Brien: Connectionism, analogicity and mental content, *Acta Anal.* **22**, 111–131 (1998)
- 32.66 G. O'Brien, J. Opie: How do connectionist networks compute?, *Cogn. Process.* **7**(1), 30–41 (2006)
- 32.67 R. Trenholme: Analog simulation, *Philos. Sci.* **61**(1), 115–131 (1994)
- 32.68 G. Basti: Intelligibility and reference: Formal ontology of the natural computation. In: *Computing Nature*, ed. by G. Dodig-Crnkovic, R. Giovagnoli (Springer, Berlin, Heidelberg 2013) pp. 139–159
- 32.69 W.J. Freeman: The neurobiological infrastructure of natural computing: Intentionality, *New Math. Nat. Comput.* **5**(1), 19–29 (2009)
- 32.70 H. Maturana, F. Varela: *Autopoiesis and Cognition: The Realization of the Living* (Reidel, Dordrecht 1980)
- 32.71 E. Ben-Jacob, I. Becker, Y. Shapira: Bacteria linguistic communication and social intelligence, *Trends Microbiol.* **12**(8), 366–372 (2004)
- 32.72 E. Ben-Jacob, Y. Shapira, A.I. Tauber: Seeking the foundations of cognition in bacteria, *Physica A* **359**, 495–524 (2006)
- 32.73 E. Ben-Jacob: Learning from bacteria about natural information processing, *Ann. NY Acad. Sci.* **1178**, 78–90 (2009)
- 32.74 W.-L. Ng, B.L. Bassler: Bacterial quorum-sensing network architectures, *Annu. Rev. Genet.* **43**, 197–222 (2009)
- 32.75 S. Schauder, B.L. Bassler: The languages of bacteria, *Genes Dev.* **15**, 1468–1480 (2001)
- 32.76 J.A. Shapiro: Bacteria as multicellular organisms, *Sci. Am.* **256**(6), 82–89 (1988)
- 32.77 J.A. Shapiro: Natural genetic engineering in evolution, *Genetica* **86**, 99–111 (1992)
- 32.78 J.A. Shapiro: Bacteria are small but not stupid: Cognition, natural genetic engineering and sociobacteriology, *Stud. Hist. Philos. Biol. Biomed. Sci.* **38**, 807–819 (2007)
- 32.79 R. Xavier, N. Omar, L. de Castro: Bacterial colony: Information processing and computational behavior, 3rd World Cong. Nat. Biol. Inspir. Comput. (NaBIC) (2011) pp. 439–443
- 32.80 L. Magnani, E. Bardone: Sharing representations and creating chances through cognitive niche construction. The role of affordances and abduc-

- tion. In: *Communications and Discoveries From Multidisciplinary Data*, ed. by S. Iwata, Y. Oshawa, S. Tsumoto, N. Zhong, Y. Shi, L. Magnani (Springer, Berlin, Heidelberg 2008) pp. 3–40
- 32.81 S. Stepney, S.L. Braunstein, J.A. Clark, A.M. Tyrrell, A. Adamatzky, R.E. Smith, T. Addis, C. Johnson, J. Timmis, P. Welch, R. Milner, D. Partridge: Journeys in non-classical computation I: A grand challenge for computing research, *Int. J. Parallel Emerg. Distr. Syst.* **20**, 5–19 (2005)
- 32.82 S. Stepney, S.L. Braunstein, J.A. Clark, A.M. Tyrrell, A. Adamatzky, R.E. Smith, T. Addis, C. Johnson, J. Timmis, P. Welch, R. Milner, D. Partridge: Journeys in non-classical computation II: Initial journeys and waypoints, *Int. J. Parallel Emerg. Distr. Syst.* **21**, 97–125 (2006)
- 32.83 G. Piccinini: Computation in physical systems. In: *The Stanford Encyclopedia of Philosophy*, ed. by E. N. Zalta (Fall 2012 Edition) <http://plato.stanford.edu/archives/fall2012/entries/computation-physicalsystems/>
- 32.84 T. Deacon: *Incomplete Nature. How Mind Emerged from Matter* (Norton, New York, London 2011)
- 32.85 C.E. Maldonado, A.N. Gómez Cruz: Biological hypercomputation: A new research problem in complexity theory, *Complexity* **20**(4), 8–18 (2015)
- 32.86 F. Hernandez Quiroz: Computational and human mind model. In: *The Computational Turn: Past, Presents, Futures?*, ed. by C. Ess, R. Hagengruber (Monsenstein Vannerdat, Münster 2011) pp. 104–106
- 32.87 O. Bournez, M. Cosnard: On the computational power and super-Turing capabilities of dynamical systems, *Theor. Comput. Sci.* **168**, 417–459 (1996)
- 32.88 H.T. Siegelmann: Turing on super-Turing and adaptivity, *Prog. Biophys. Mol. Biol.* **113**(1), 117–126 (2013)
- 32.89 A.M. Turing: Systems of logic based on ordinals, *Proc. Lond. Math. Soc.* **s2-45**, 161–228 (1939)
- 32.90 C.E. Maldonado, A.N. Gómez Cruz: Biological hypercomputation: A concept is introduced, <http://arxiv.org/abs/1210.4819> (2012)
- 32.91 S. Inoue, T. Matsuzawa: Working memory of numerals in chimpanzees, *Curr. Biol.* **17**(23), R1004–R1005 (2007)
- 32.92 R. Bshary, W. Wickler, H. Fricke: Fish cognition: A primate’s eye view, *Anim. Cogn.* **5**(1), 1–13 (2002)
- 32.93 E.L. MacLean, L.J. Matthews, B.A. Hare, C.L. Nunn, R.C. Anderson, F. Aureli, E.M. Brannon, J. Call, C.M. Drea, N.J. Emery, D.B.M. Haun, E. Herrmann, L.F. Jacobs, M.L. Platt, A.G. Rosati, A.A. Sandel, K.K. Schroepfer, A.M. Seed, J. Tan, C.P. van Schaik, V. Wobber: How does cognition evolve? Phylogenetic comparative psychology, *Anim. Cogn.* **15**(2), 223–238 (2012)
- 32.94 H. Bergson: *Creative Evolution* (Dover, New York 1998)
- 32.95 G. Dodig-Crnkovic: Knowledge generation as natural computation, *J. Syst. Cybern. Inf.* **6**(2), 12–16 (2008)
- 32.96 R.N. Giere: Models as parts of distributed cognitive systems. In: *Model-Based Reasoning: Science, Technology, Values*, ed. by L. Magnani, N. Nersessian (Kluwer, Dordrecht 2002) pp. 227–241
- 32.97 D.E. Rumelhart, J.L. McClelland, PDP Research Group: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations* (MIT Press, Cambridge 1986)
- 32.98 L. Suchman: *Plans and Situated Actions* (Cambridge Univ. Press, Cambridge 1987)
- 32.99 F. Varela, E. Thompson, E. Rosch: *The Embodied Mind: Cognitive Science and Human Experience* (MIT Press, Cambridge 1991)
- 32.100 E. Hutchins: *Cognition in the Wild* (MIT Press, Cambridge 1995)
- 32.101 A. Clark: *Being There: Putting Brain, Body and World Together Again* (Oxford Univ. Press, Oxford 1997)
- 32.102 A. Clark: *Supersizing the Mind Embodiment, Action, and Cognitive Extension* (Oxford Univ. Press, Oxford 2008)
- 32.103 A. Clark, D.J. Chalmers: The extended mind, *Analysis* **58**(1), 7–19 (1998)
- 32.104 T. Froese, T. Ziemke: Enactive artificial intelligence: Investigating the systemic organization of life and mind, *Artif. Intell.* **173**(3/4), 466–500 (2009)
- 32.105 T.T. Rogers, J.L. McClelland: Parallel distributed processing at 25: Further explorations in the microstructure of cognition, *Cogn. Sci.* **38**, 1024–1077 (2014)
- 32.106 R. Giere: Scientific cognition as distributed cognition. In: *The Cognitive Basis of Science*, ed. by P. Carruthers, S.P. Stich, M. Siegal (Cambridge Univ. Press, Cambridge 2002) p. 285
- 32.107 W. Bechtel: What knowledge must be in the head in order to acquire knowledge? In: *Communicating Meaning: The Evolution and Development of Language*, ed. by B.M. Velichkovsky, D.M. Rumbaugh (Lawrence Erlbaum, New Jersey 1996)
- 32.108 A.-L. Barabasi: The architecture of complexity, *IEEE Control Syst. Mag.* **27**(4), 33–42 (2007)
- 32.109 I. Cafezeiro, C. Gadelha, V. Chaitin, I.C. da Marques: A knowledge-construction perspective on human computing, collaborative behavior and new trends in system interactions, *Lect. Notes Comput. Sci.* **8510**, 58–68 (2014)
- 32.110 N.J. Nersessian: The cognitive basis of model-based reasoning in science. In: *The Cognitive Basis of Science*, ed. by P. Carruthers, S. Stich, M. Siegal (Cambridge Univ. Press, Cambridge 2002) pp. 133–153
- 32.111 N.J. Nersessian: Should physicists preach what they practice? Constructive modeling in doing and learning physics, *Sci. Educ.* **4**, 203–226 (1995)
- 32.112 N.J. Nersessian: Model-based reasoning in conceptual change. In: *Model-Based Reasoning in Scientific Discovery*, ed. by L. Magnani, N. Nersessian, P. Thagard (Kluwer/Plenum, New York 1999) pp. 5–22
- 32.113 L. Magnani, N.J. Nersessian, P. Thagard (Eds.): *Model-Based Reasoning in Scientific Discovery* (Kluwer/Plenum, New York 1999)



- 32.114 L. Magnani, N.J. Nersessian: *Model-Based Reasoning: Science, Technology, Values* (Springer, New York 2002)
- 32.115 R.N. Giere: Using models to represent reality. In: *Model-Based Reasoning in Scientific Discovery*, ed. by L. Magnani, N. Nersessian, P. Thagard (Kluwer/Plenum, New York 1999) pp. 41–57
- 32.116 L. Magnani: Model-based and manipulative abduction in science, *Found. Sci.* **9**(3), 219–247 (2004)
- 32.117 N.H. Narayanan, M. Suwa, H. Motoda: Hypothesizing behaviors from device diagrams. In: *Diagrammatic Reasoning: Cognitive and Computational Perspectives*, ed. by J. Glasgow, N.H. Narayanan, B. Chanrasekaran (MIT Press, Cambridge 1995) pp. 501–534
- 32.118 D. Wang, J. Lee, H. Zeevat: Reasoning with diagrammatic representations. In: *Diagrammatic Reasoning: Cognitive and Computational Perspectives*, ed. by J. Glasgow, N.H. Narayanan, B. Chanrasekaran (MIT Press, Cambridge 1995)
- 32.119 R.N. Giere: *Scientific Perspectivism* (Univ. Chicago Press, Chicago 2006)
- 32.120 L. Magnani: *Abductive Cognition: The Epistemological and Eco-Cognitive Dimensions of Hypothetical Reasoning* (Springer, Berlin, Heidelberg 2009)
- 32.121 L. Magnani, M. Piazza: Morphodynamical abduction. Causation by attractors dynamics of explanatory hypotheses in science, *Found. Sci.* **10**, 107–132 (2005)
- 32.122 P. Thagard: *Conceptual Revolutions* (Princeton Univ. Press, Princeton 1992)
- 32.123 T. Deacon: *The Symbolic Species: The Co-Evolution of Language and the Brain* (Norton, New York, London 1997)
- 32.124 K. Czarnecki: Generative Programming. Principles and Techniques of Software Engineering Based on Automated Configuration and Fragment-Based Component Models, Ph.D. Thesis (Department of Computer Science and Automation, Technical University of Ilmenau, Ilmenau 1998)
- 32.125 J. Rothenberg: The nature of modeling. In: *Artificial Intelligence, Simulation, and Modeling*, ed. by L.E. William, K.A. Loparo, N.R. Nelson (Wiley, New York 1989) pp. 75–92
- 32.126 F. Dau, P.W. Eklund: A diagrammatic reasoning system for the description logic ALC, *J. Vis. Lang. Comput.* **19**(5), 539–573 (2008)
- 32.127 Object Management Group (OMG): UML 2.4.1 Superstructure Specification. OMG Document Number: formal/2011-08-06 (OMG 2011), <http://www.omg.org/spec/UML/2.4.1/>
- 32.128 D. Harel, B. Rumpe: Meaningful modeling: What's the semantics of "semantics"?, *IEEE Comput.* **37**(10), 64–72 (2004)
- 32.129 A. Vallecillo: A journey through the secret life of models, *Persp. Workshop: Model Eng. Complex Syst. (MECS), Dagstuhl Sem. Proc.* 08331. (2008)
- 32.130 T. Kühne: Matters of (meta-) modeling, *Soft. Syst. Model.* **5**(4), 369–385 (2006)
- 32.131 G. Dodig-Crnkovic: Info-computational constructivism and cognition, *Constr. Found.* **9**(2), 223–231 (2014)

# 33. Computational Scientific Discovery

Peter D. Sozou, Peter C.R. Lane, Mark Addis, Fernand Gobet

Computational scientific discovery is becoming increasingly important in many areas of science. This chapter reviews the application of computational methods in the formulation of scientific ideas, that is, in the characterization of phenomena and the generation of scientific explanations, in the form of hypotheses, theories, and models. After a discussion of the evolutionary and anthropological roots of scientific discovery, the nature of scientific discovery is considered, and an outline is given of the forms that scientific discovery can take: direct observational discovery, finding empirical rules, and discovery of theories. A discussion of the psychology of scientific discovery includes an assessment of the role of induction. Computational discovery methods in mathematics are then described. This is followed by a survey of methods and associated applications in computational scientific discovery, covering massive systematic search within a defined space; rule-based reasoning systems; classification, machine vision, and related techniques; data mining; finding networks; evolutionary computation; and automation of scientific experiments. We conclude with a discussion of the future of computational scientific discovery, with consideration of the extent to which scientific discovery will continue to require human input.

33.1	<b>The Roots of Human Scientific Discovery</b> .....	720
33.2	<b>The Nature of Scientific Discovery</b> .....	721
33.3	<b>The Psychology of Human Scientific Discovery</b> .....	722
33.4	<b>Computational Discovery in Mathematics</b> .....	723
33.4.1	Logic Theorist .....	723
33.4.2	AM and EURISKO .....	724
33.4.3	GRAFFITI .....	724
33.5	<b>Methods and Applications in Computational Scientific Discovery</b> ..	725
33.5.1	Massive Systematic Search Within a Defined Space .....	726
33.5.2	Rule-Based Reasoning Systems .....	726
33.5.3	Classification, Machine Vision, and Related Techniques .....	727
33.5.4	Data Mining .....	727
33.5.5	Finding Networks .....	727
33.5.6	Evolutionary Computation .....	728
33.5.7	Automation of Scientific Experiments ..	729
33.6	<b>Discussion</b> .....	730
	<b>References</b> .....	731

Science is concerned with characterizing and explaining observations and phenomena. For most of history, it has been an exclusively human activity. However, the development of computers has had a substantial impact on science. The assessment and testing of scientific models has seen the application of computational methods, often with spectacular success. Among these major successes have been the development of numerical and simulation methods to compute the predictions of scientific models [33.1, 2].

A more ambitious endeavour is the use of computational methods to represent, in some sense, the formu-

lation of scientific ideas – the characterization of phenomena and the generation of scientific explanations, in the form of hypotheses, theories, and models. This is the focus of this review. What computational methods have been developed so far, and how have they been applied? What scope is there for further developments and applications?

We begin by discussing the evolutionary biological and anthropological roots of scientific discovery, and the establishment of scientific discovery as a human endeavor. We then set out important features of the nature of the scientific discovery process, and the

different forms of discovery. This provides a basis for considering the psychology of scientific discovery. This is followed by a discussion of computational discovery methods in mathematics, which, because of the close links between mathematics and the theoretical sciences, provides a useful prelude to computational discovery

in science. We then survey methods and applications in computational scientific discovery. We finally draw conclusions, with particular attention paid to what type of problems in scientific discovery computational methods are best suited to, and the future of computational scientific discovery.

### 33.1 The Roots of Human Scientific Discovery

Humans are capable of learning, that is, acquiring and processing information, so as to influence their future actions. This includes an ability to make inferences from limited observations. Scientific discovery, it has been claimed, uses similar inferential principles [33.3]. Why did natural selection favor the capacity for learning in animals? A major factor is likely to have been environmental variability [33.4]. There may have been something of a self-sustaining evolutionary process at work, with arms races between the learning capabilities of predators and prey, and competition for resources between members of the same species favoring the best learners. Animals with highly developed brains, particularly primates and corvids, exhibit the ability to innovate and to make tools to overcome new challenges [33.5–7]. Humans, however, stand out from other animals in the extent of their development of tools.

The precise evolutionary basis of human language is uncertain [33.8], but language facilitated the transmission and synthesis of useful information. Pre-agricultural communities are believed to have had a great deal of useful knowledge, largely gathered through trial and error. They demonstrated representational skills with the creation of art objects such as cave art (dating from approximately 40 000 years ago in Europe and Indonesia). Plant domestication is thought to have first occurred in the fertile crescent of Western Asia; the Sumerians of Mesopotamia are strong contenders for being the first to invent writing, somewhat before 3000 BC [33.9]. These developments led to highly ordered and stratified societies, with specialization of labor, significant technologies, and empirical knowledge of many important processes. The form of enquiry recognized today as general scientific explanation came later; the first search for the *causes and principles* of the natural world is attributed to Thales of Miletus around the beginning of the 6th century BC [33.10]. This notion was further developed by Aristotle, explaining what is less well known by means of what is better known and more fundamental, and

the requirement for explanations to be consistent with observations [33.11]. Aristotle's approach is a direct forerunner of the model of scientific discovery that is broadly accepted today, in which scientists propose new hypotheses and then devise experimental tests for them [33.12].

*Ernest Rutherford* remarked that “All science is either physics or stamp collecting” [33.13]. Although this could be said to be simplistic, it captures the essential truth that a great deal of science involves what may be termed *direct observational discovery*: Observing and characterizing phenomena, without the need for deep theoretical explanations. Thales of Miletus is credited with the discovery that when amber is rubbed with fur, it then attracts light objects – what is known today as static electricity [33.14]. The development of optical microscopy in the sixteenth and seventeenth centuries led to the discovery of bacteria, biological cells, and spermatozoa. These were clearly scientific discoveries of vital importance, but they were not in themselves deep theories. Some research in astronomy also has this character, such as the discovery in 1930 of Pluto by Clyde Tombaugh from examination of photographic plates. A deep theoretical understanding of gravitation was important for this discovery – anomalies in the motion of Uranus, not fully explained by the gravitational effects of Neptune, suggested the existence of a hitherto undiscovered planetary body – but the discovery itself was a direct observational discovery: It did not constitute the discovery of an observational law or of a theory.

These discoveries typify situations in which scientific discovery follows directly from technological advances. Thus, the development of x-ray crystallography [33.15] enabled the structure of proteins [33.16] and of deoxyribonucleic acid (DNA) [33.17] to be established. More recently, modern studies in molecular biology have led to a vast amount of data becoming available. This makes it worthwhile to consider data-driven scientific investigation [33.18], which is frequently computational and can complement the hypothesis-driven approach.

## 33.2 The Nature of Scientific Discovery

Scientific discovery involves a range of activities and methods. One general form, as described above, is direct observational discovery: simply finding and characterizing a phenomenon, such as the aforementioned discovery of what became known as static electricity [33.14]. *Langley* [33.19] describes the formation of taxonomies as the first stage of the discovery process. However, we do not consider direct observational discovery to be limited to the formation of taxonomies, as it may include discoveries that fit within existing taxonomies, such as the discovery of Pluto described above. Additionally, the formation of taxonomies is not always the first stage of scientific discovery; for example, the classification of organisms according to how recently they last shared a common ancestor depends on the prior theory that the organisms in question are descended from a common ancestor. In short, there are cases where direct observational discovery does not involve the formation of taxonomies, and cases where the formation of taxonomies is a late stage in the discovery process. We therefore consider direct observational discovery to be a basic form of scientific discovery. It generally occurs at the birth of a scientific field, but often continues as the field matures.

A different form of discovery, which generally comes after a field has matured sufficiently to have accepted terms for describing observations and quantitative measures of these observables, is concerned with finding empirical rules. Such a rule is a useful description of some aspect of the world to which a number of observations conform. Prominent examples are Kepler's laws of planetary motion, Weber's law (also known as the Weber–Fechner law) in psychology, which states that the just noticeable difference between two stimuli is proportional to the size of the stimuli, and the constancy of the speed of light in any reference frame, established by the Michelson–Morley experiment of 1887. This form of discovery can, if desired, be divided into those that involve qualitative rules on the one hand and quantitative on the other; an additional distinction concerns whether or not the rules involve unobserved entities [33.19].

Another category of discovery is theories. A theory is an underlying explanation, accounting for a set of observations by means of a causal process. For example, Newton's theory of gravitation explains Kepler's observations by means of a deeper, causal principle. The theory of evolution by natural selection, conceived by Darwin and Wallace, explains a wide range of observations regarding organisms' adaptations, and forms the basis of the modern science of behavioral ecology.

It should be recognized, however, that the distinction between different categories of scientific discovery is not absolute [33.3, 20]. Newton's theory of gravitation is a case in point: While we believe it is appropriate to call it a theory, one can also describe it as an empirical rule, albeit a deeper one with more explanatory power than Kepler's laws. A putative chemical structure can be regarded as a theory, and therefore the discovery of a chemical structure (such as the discovery by Watson and Crick of the double helix structure of DNA) can be regarded as discovery of a scientific theory, but it is not clear that there is a clear division between this discovery and direct observational discovery; in contrast, the discovery of the theory of evolution by natural selection is very different from a direct observational discovery.

Philosophers of science have been concerned mostly with the discovery of scientific explanations, that is, of empirical rules and theories. The essential characteristic of a scientific explanation is that it be logically coherent [33.21]. But how are scientific explanations (in the form of empirical rules or theories) generated in the first place? And how should a choice be made between competing empirical rules or theories?

Assessing potential explanations is problematic because there is no agreed objective method to assess a potential explanation against an alternative (nor against the possibility that no explanation so far advanced is a good explanation of the phenomenon in question). There are some generally agreed principles that make a potential explanation more likely to be accepted. These are that it is better if it fits the data more closely, better if it is more parsimonious, and better if it is more plausible. However, while metrics can sometimes be calculated for how well a theory fits the data, and for its parsimony, the question of plausibility is ultimately one for human judgment. There is also the problem of how much weight to put on all these factors. Humans appear to have developed pragmatic, context-specific methods of making acceptable inferences [33.3]: The psychology of scientific discovery is discussed in more detail in Sect. 33.3.

The process of finding an *explanation* for a specified phenomenon of interest has the characteristics of an inverse problem. These problems have the following form. There is some state of the world  $S$  which, in conjunction with certain physical laws, gives rise to a data set  $D$ . Given  $S$ , it is in principle straightforward to calculate  $D$ . This is known as the forward problem. The inverse problem is to calculate  $S$  from  $D$ . There may be no solution  $S$  that generates  $D$  exactly, or there

may be several. An example of such a problem concerns mirage data. A mirage is an optical distortion caused by meteorological conditions, which result in variation in the atmospheric refractive index [33.22]. The forward problem is to calculate what a mirage should look like from the refractive index profile. The inverse problem is to deduce the refractive index profile (and from this the temperature profile) from mirage data [33.23]. One approach involves a form of regularization: finding a refractive index profile to minimize a cost function [33.24]. This cost function includes an error term, which depends on how far the mirage data predicted by the proposed profile differs from the real data; computation of this term involves solving the forward problem for the proposed refractive index profile. Another term penalizes the proposed refractive index profile according to a measure of its implausibility. Vision problems are inverse problems [33.25]. A common characteristic of inverse problems is that, as with problems of scientific explanation, there is generally no clear, objective measure of the *best* solution; rather, ad hoc problem-specific measures are needed. As with the example of the inversion of mirage data [33.24], a feature of almost every computational approach to inverse problems is that a candidate solution is tested by solving the forward problem and calculating the data that would be predicted under this candidate solution.

Most inverse problems do not involve scientific discovery – for example, inverting mirages to find a temperature profile would not be considered scientific discovery – but the similarities between scientific discovery and inverse problems are clear, and indeed scientific discovery can be regarded as an inverse problem. It can be cast as follows: Given a suitable theoretical or empirical account of the relevant aspect of the world, specified in sufficient detail, it is possible to calculate what observations should follow from it: that is the forward problem. The inverse problem is to go from the observations to the empirical rule or the theory. It is therefore pertinent to consider the computational methods used in inverse problems. Computational optimization methods are widely used [33.24, 26]. Evolutionary computational methods have been applied [33.27–29]: These are particularly suited to problems in which there are several local optima, where standard optimization methods often get stuck at a local optimum. We will discuss evolutionary computational methods for scientific discovery in more detail in Sect. 33.5.

It must be remembered, however, that not all scientific discovery involves finding an explanation for a specified problem. Rich data sets allow data-driven research in which new entities are directly discovered, and new empirical rules suggested by data analysis [33.18]; we will describe in Sect. 33.5 how computational methods play a large part in these processes.

### 33.3 The Psychology of Human Scientific Discovery

At this point, it is pertinent to ask what is understood about how humans make scientific discoveries. This allows consideration of whether or not the human computational discovery process can be effectively simulated computationally. If it can, a further question arises: Does computational simulation or replication of the methods used in human scientific discovery give rise to effective practical tools for scientific discovery?

*Gillies* [33.30] describes two contrasting approaches to understanding scientific discovery. *Francis Bacon* [33.31] postulated a central role for induction in scientific discovery: Put simply, he regarded discovery as dependent, in a large part, on the simple application of logical rules to observations, with the development of theory strictly following the collection of data. In contrast, *Karl Popper* argued that observations are generally selective, and made within a theoretical context; he suggested that science proceeds through a process of conjecture and falsification [33.32]. On the question of where such conjectures come from, *Popper* [33.33] suggested that “there is no such thing as a logical method for having new ideas” [33.33, p. 37]. This line of think-

ing would seem to conflict with the modern notion that the human brain is effectively a computational device, though from a practical point of view it would also effectively apply if the human brain is computational, but too complex for understanding or meaningful simulation of brain activities that generate new ideas to be possible.

Where the aim is to explain a specified phenomenon, in the form of finding a suitable empirical rule or theory which is plausible and fits the data, the task can be described as a search process [33.20, 34, 35]. *Campbell* [33.34] and *Simonton* [33.35] consider a largely random search process, followed by a rigorous selection process. *Langley* et al. [33.20] put more emphasis on the concept of heuristic search, that is, using a search method which is intended to give a relatively high chance of finding a good solution quickly, compared to a purely random search of the same search space. In a study of how subjects learn to use an electronic device, *Klahr* and *Dunbar* [33.36] propose the concept of scientific discovery by dual search [33.34], involving a search in both hypothesis space and experi-

ment space. Klahr and Dunbar report that subjects vary in their approach to discovery: The principal approach of one group, the theorists, was to formulate hypotheses and then test them experimentally, while a second group, the experimenters, conducted many experiments without explicit hypotheses. This would seem to provide evidence that humans can use both *Popperian* and *Baconian* approaches to discovery. In the assessment of hypotheses, however, humans are prone to confirmation bias, that is, putting disproportionate weight on studies that would tend to confirm a favored hypothesis [33.37].

*Qin* and *Simon* [33.38] show how human subjects can discover a scientific law (Kepler's third law of planetary motion) by a process of exploring possible, though simple, algebraic relationships between two variables, and using the results of early explorations to make better-informed guesses of the relationship. This approach mimics the workings of the BACON computational scientific discovery method for finding scientific laws [33.39]. However, the participants in the experiment of *Qin* and *Simon* [33.38] were primed to expect that some sort of simple relationship between the variables existed. Kepler did not know for certain that such a relationship existed, so the extent to which the experiment of *Qin* and *Simon* represents the sit-

uation faced by Kepler is questionable. Nevertheless, *Qin* and *Simon*'s results do provide some support for the role of rule-based reasoning in scientific discovery. *Kulkarni* and *Simon* [33.40] developed the KEKADA computational system for explaining general scientific processes; it is able to replicate the discovery of *Hans Krebs*'s hypothesis for the urea–ornithine cycle [33.41], using rule-based methods, again providing some support for the idea that inductive methods are important in human scientific discovery. See Sect. 33.5 for more details about the BACON and KEKADA computational discovery systems.

Notwithstanding the *Baconian* view of discovery, it is generally held that at least some forms of human scientific discovery involve something which can be termed creativity. *Boden* [33.42] suggests that there are three forms of creativity. The first is *combinational* creativity: the new (unexpected) combination of familiar ideas. The second is *exploratory* creativity: the exploration of accepted, structured spaces. The third is *transformational* creativity, which involves ideas outside the rules of an accepted space. A capacity that contributes to creativity, and is particularly important in scientific discovery, is the use of analogy [33.3]; this has been the subject of cognitive modeling, with an emphasis on the role of memory [33.43].

## 33.4 Computational Discovery in Mathematics

Mathematics is not a scientific discipline, as it does not ostensibly deal directly with objects from the real world. However, as a formal language, mathematics is used to represent and discuss concepts in many scientific disciplines. This is most apparent in physics, where theories such as general relativity and the standard model for particle physics rely extensively on mathematical techniques for theoretical understanding and model construction (for an early discussion of this link, see *Wigner* [33.44]).

Many of the techniques found here are precursors of similar ideas used for computational discovery in other sciences, in particular the idea of using heuristic (rule-based) techniques to search a space of candidate solutions. Our selection of systems is intended to be indicative of the general techniques. For more details and coverage of additional systems in this area, see *Colton* [33.45], *Colton et al.* [33.46], and *Larson* [33.47].

### 33.4.1 Logic Theorist

Logic Theorist (LT) [33.48] was a program developed by *Newell et al.* to find proofs in elementary (proposi-

tional) logic. LT is programmed with some axioms and inference rules found in *Whitehead* and *Russell* [33.49], and develops proofs of several theorems found therein. It relies on a method that will also be seen in some systems in the next section: A space of possible solutions – in this case proofs – is explored using a class of methods known as heuristics. Heuristics can be defined as rules of thumb that are likely (but not certain) to provide a correct solution to a problem. Using heuristics has the advantage of cutting down the number of states that are visited during search. This makes it possible to carry out a highly selective search, sometimes called a heuristic search.

When started, LT stores in its memory a list of axioms, a list of operations, and the logical expression to be proved. It uses four rules of inference when searching for a proof. These are: substitution, which allows a variable to be replaced by a new variable or expression; replacement, where expressions, such as *implies*, are replaced by equivalent expressions using other connectives; detachment, which can split up compound statements when one part has been proved; and syllogism (chaining), which enables chains of inferences to be followed.

As in chess, where the legality of a move does not imply its quality, LT uses several proof strategies to make its search efficient. These strategies look for situations where the above rules of inference may be applied: For example, when wanting to show *a implies c* and knowing that *b implies c*, then proving *a implies b* would enable use of the syllogism inference.

Although intended as a program for proving logical statements, *Newell et al.* [33.48] also make some novel claims about the program's ability to simulate the behavior of human problem solvers, and were among the first authors to argue for the use of computer programs to aid in understanding psychological processes. See *Gobet and Lane* [33.50] for a discussion of LT's impact on psychology. Around the same time, a program was developed [33.51] that did not lay any claims to modeling human behavior, but used more intensive computational strategies to prove all the theorems in Chaps. 1–3 of *Whitehead and Russell* [33.49].

### 33.4.2 AM and EURISKO

Automated Mathematician (AM) [33.52, 53] is a program which uses heuristics to discover conjectures (and hence potential theorems) in mathematics. Unlike LT, it is unable to directly prove theorems. Instead, conjectures are created by modifying existing theorems, and tested empirically against specific examples. AM works within the set and number theory areas of mathematics, in part because it is easy for the computer to generate examples of, say, a conjecture about numbers. AM has been criticized [33.54], but serves as a model of how a mathematical discovery system can be designed.

AM begins with a basic set of concepts and can create well-known conjectures in set theory (such as subsets) and number theory (such as prime numbers, or Goldbach's conjecture). AM represents each concept in a frame, holding information including: its algorithm, examples, which other concepts it is related to by generalization or specialization, and other related concepts. A concept is then selected out of the current pool, based on a measure of *interestingness*, and adapted versions of the concept are created. These adaptations may be a specialization of the concept, a restriction of its domain, or similar. In addition, and crucially, a human observer may indicate a concept for the system to work on next.

As AM proved fairly successful in mathematics, a related system, EURISKO, was built to attempt to discover new search heuristics: EURISKO is a meta-discovery system, which discovers new ways for discovery to occur. As *Lenat and Brown* [33.55] explain, one of the discoveries of EURISKO was that AM's success was largely down to how mathematical concepts were represented. Replicating AM's success in other domains requires careful work on formulating the internal representation of the domain. In particular, changing the syntactic form of an expression should be reflected in meaningful changes to the semantics (to the meaning).

### 33.4.3 GRAFFITI

GRAFFITI [33.56] is a system for developing conjectures in an area of mathematics known as graph theory. A graph is a set of nodes interconnected by edges, and graph theory has important applications in many scientific areas, including physics, chemistry and computer science (see Sects. 33.5.1 and 33.5.5).

GRAFFITI, like AM, is used to generate conjectures but is unable to construct proofs. The conjectures are formed from a database of invariants of a graph, such as the diameter (greatest number of edges between any two nodes), rank (number of nodes minus the number of connected components), or chromatic number (the number of colors required to color a graph so that no two adjacent nodes are of the same color). Simple sums of these invariants are then generated, and tested against a database of known graphs.

As the conjectures can require considerable computational time to check, a pair of heuristics are used to try to focus on interesting conjectures. The *beagle* heuristic is used to check that the conjecture is not trivial, for example, that an invariant is less than itself plus 1. The *dalmation* heuristic is used to check that the conjecture is different to ones already in GRAFFITI's database.

If no counterexample can be found by the program, the conjecture is passed to the user. Conjectures are then published, and may be picked up and further analysed by graph theorists. GRAFFITI is one of the few systems proven to make conjectures which mathematicians find interesting, and has helped to advance the field. A substantial number of the conjectures have resulted in publications: a list may be found at [33.57].

## 33.5 Methods and Applications in Computational Scientific Discovery

We now turn to implementations of computational scientific discovery methods. Some of the methods are, in their present implementation, specific to particular domains, such as astronomy, chemistry, evolutionary biology, and psychology. Others are more general. Table 33.1 lists some principal examples of computa-

tional discovery methods. Below, we consider a number of discovery techniques. This is not intended to be an exhaustive list of all techniques and applications in computational scientific discovery. The main aim is to convey a sense of the range of methods and applications.

**Table 33.1** Principal examples of scientific discovery systems

System	Type of discovery	Domain	Main technique
DENDRAL [33.58]	Chemical structures (topological)	Chemistry/biochemistry	Partly rule-based, partly brute force exhaustive consideration of possible structures
MECHEM [33.59]	Chemical reaction pathways	Chemistry	Systematic search using defined components to a defined level of complexity; rule based/reasoning framework for controlling search
BACON [33.39]	Scientific laws	General	Rule-based/reasoning
GLAUBER [33.60]	Qualitative rules	Primarily chemical reactions, but potentially general	Rule-based/reasoning
KEKADA [33.40]	Scientific processes	General	Rule-based system that seeks to explain phenomena by recursively generating hypotheses; can also propose experiments
GOLEM [33.61]	Predicting three-dimensional structure of proteins	Biochemistry	Statistical classifier, using a machine learning method to determine the classification rules
<i>Storrie-Lombardie et al.</i> [33.62]	Classifying galaxies	Astronomy	Classifier based on a neural network, trained using backpropagation
<i>Shamir</i> [33.63]	Classifying galaxies	Astronomy	Classifier based on weighted proximity to a number of descriptors determined from test data
<i>Tiffin et al.</i> [33.64]	Candidate genes for disease causation	Biomedical	Data mining, combining gene expression data and biomedical literature
Warmr [33.65]	Potentially carcinogenic chemicals	Chemistry	Data mining combined with rule-based reasoning
GRAM [33.66]	Co-expressed genes and regulatory networks	Biomedical	Network generation from pairwise measures of expression, then incremental node addition
PC algorithm [33.67]	Causal relationships between variables	General	Network refinement by successive deletion and directional interpretation of edges
<i>Guindon and Gascuel</i> [33.68]	Deducing phylogenies from DNA	Evolutionary biology	Hill-climbing to improve estimated phylogenetic tree, based on maximum-likelihood methods applied to DNA data
<i>Frias-Martinez and Gobet</i> [33.69]	Process-based theories	Psychology	Evolutionary computation (genetic programming)
<i>Schmidt and Lipson</i> [33.70]	Scientific laws	General	Symbolic regression, based on genetic programming
Robot scientist [33.71]	Formulation and experimental testing of simple hypotheses	Biochemistry	Rule-based reasoning and control of a robot



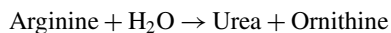
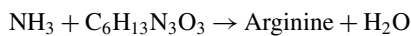
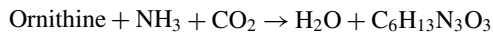
### 33.5.1 Massive Systematic Search Within a Defined Space

The first significant achievement in computational scientific discovery was a system known as heuristic DENDRAL. This began in 1965, with the purpose of automatically finding structures in organic chemistry [33.58, 72]; an important motivation was to provide a test-bed for the applicability of ideas in the emerging field of artificial intelligence (AI). Heuristic DENDRAL relies on the idea of making intensive use of specific knowledge, together with harnessing computational power for searching. It systematically evaluates all the topologically distinct arrangements of a set of atoms consistent with the rules of chemistry. Later versions also consider three-dimensional geometry, and use data beyond mass spectrometry. Alongside heuristic DENDRAL is a sister system called meta-DENDRAL, the purpose of which is to learn the rules of mass spectroscopy from empirical data. While heuristic DENDRAL carries out computational scientific discovery directly, meta-DENDRAL is a method for developing a tool used in scientific discovery. It is therefore heuristic DENDRAL which is of more direct interest to the present study. The core of heuristic DENDRAL is a plan-generate-test algorithm. The planner devises hypotheses that reject or propose certain classes of chemical graph. A key feature of the planner is that it incorporates specialist knowledge to constrain the set of potential solutions considered by the generator. The generator was designed to exhaustively and efficiently generate all the possible chemical structures, within specified constraints. The testing part of the algorithm includes a prediction component. This takes a proposed structure and generates a predicted mass spectrum: This can be compared to the real data. This testing of a possible solution by comparing the predicted data it would generate to the real data involves a principle discussed in the consideration of inverse problems in Sect. 33.2: testing a proposed solution to an inverse problem by solving the forward problem for this solution, and comparing the predicted outcome to the real data.

It should be emphasized that heuristic DENDRAL is not just a number cruncher: The use of specialized knowledge to constrain the set of potential solutions considered – justifying the term *heuristic* – was very important to the success of this system. However, number crunching was also necessary: the capacity to use computational power to systematically evaluate a set of potential solutions which is too large a job to undertake manually.

Another system that makes good use of systematic search within a defined space is MECHEM [33.59]. This is concerned with finding chemical reaction path-

ways, that is, the steps involved in a chemical reaction. An example is a possible pathway in the urea–ornithine cycle, originally proposed by *Krebs* [33.41].



In this proposed pathway, the chemical species has been conjectured: It was not observed, nor was it part of the input data. The hypothesis-formation algorithm used in MECHEM makes use of two complexity parameters, specifying the number of steps and species to be contained in a hypothesis. Then the hypothesis generator finds the possible hypotheses within this constraint. If they are all rejected, then at least one of the complexity parameters is incremented. The MECHEM system also has a higher level (rule-based) decision system, allowing it to indicate that new experimental evidence should be sought, or that the problem be suspended.

### 33.5.2 Rule-Based Reasoning Systems

The BACON research program [33.20, 38], as with DENDRAL described above, originated in artificial intelligence research. BACON is a series of systems for discovering empirical rules, in the form of laws, by uncovering relationships within data sets. It makes use of rule-based induction, looking initially for simple relationships between variables, such as an invariant ratio or product. In what it achieves, BACON has a lot in common with regression, and can be considered a form of dimension reduction. Later versions of BACON go beyond simple regression and dimension reduction by generating properties representing intrinsic properties of entities; an example is the refractive index, generated in the discovery of Snell's law of refraction.

The GLAUBER discovery system [33.20, 60] uses a similar rule-based induction process to BACON, but is concerned with qualitative empirical rules. For example, it can discover the law that every acid combines with every alkali to produce some salt.

The KEKADA system [33.40] is a tool with the purpose of understanding scientific processes; it has been applied to the urea–ornithine cycle, replicating the steps undertaken by *Krebs* [33.41] to formulate his famous hypothesis describing how the cycle operates. In seeking to replicate how scientists act, it includes a *problem chooser* module: This determines which discovery task to attempt when there are several potential tasks on the agenda, according to considerations such as how important a task is, and how accurately it can be studied. It allows for a puzzling experimental finding to be added to the agenda for investigation. A hypothesis genera-

tor creates hypotheses when faced with a new problem. There are also rules to propose experiments whose findings could change confidence in existing hypotheses. The results of experiments feed into the system via hypothesis modifiers and confidence modifiers.

While rule-based reasoning can be considered to be the main discovery technique in the above systems, it should be noted that rule-based reasoning is also an important component of several other scientific discovery systems. Thus, for example, heuristic DENDRAL, described above, has a rule-based system to control which searches it carries out, while GOLEM, described below, uses rule-based reasoning to learn classification rules.

### 33.5.3 Classification, Machine Vision, and Related Techniques

Computerized processes for classification, alongside techniques for object/instance recognition, including machine vision methods, are becoming increasingly important. It is therefore not surprising that such systems have found applications in computational scientific discovery. The GOLEM system has been shown to successfully produce hypotheses about protein structure [33.61]. It uses machine learning to determine rules for predicting the structure. The basic algorithm takes a random sample of pairs of example residues, taken from all the proteins in the system, and computes common properties. These properties are then made into a rule. GOLEM therefore makes use of rule-based techniques, within the envelope of what is in effect a classification system.

In astronomy, data sets may include image data, and computational methods can be applied to problems such as the identification and classification of astronomical objects. *Storrie-Lombardie* et al. [33.62] use a neural network to classify galaxies into five types, based on 13 variables measured by machine, using a backpropagation algorithm to train the network. *Shamir* [33.63] describes the automatic classification of galaxies using a method which first converts each image to a number of low-level descriptors, and then uses discriminant analysis to find the descriptors which are most informative. A weighted distance between two feature vectors can then be computed; the predicted class of a test image is given by the class of the training image that has the smallest weighted distance to it. A similar approach has been applied to the classification of structures in biological images [33.73].

### 33.5.4 Data Mining

Methods used for finding patterns from large amounts of data, sometimes from disparate sources, have been

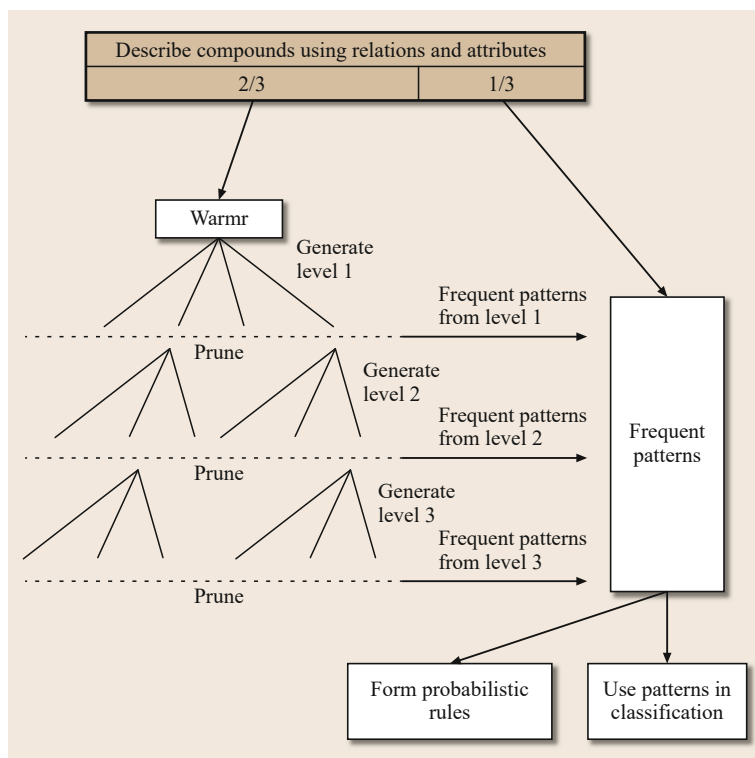
termed data mining [33.74, 75]. Scientific literature can be a useful source of such data [33.76]. For example, candidate genes for causation of disease can be determined from computational discovery of important statistical associations, using literature databases together with protein function [33.77] or gene expression [33.64] data. Computational data-mining techniques have also been used, in combination with rule-based methods, to identify candidate carcinogenic compounds, using a database of carcinogenetic tests of compounds [33.65] (Fig. 33.1). A deep commonality between machine learning (and classification) methods on the one hand and data mining techniques on the other is that both involve finding statistical associations in data. In machine learning, the emphasis tends to be on the techniques for finding associations within a given set of data, while data mining is often more concerned with techniques for extracting useful data from diverse or technically challenging sources. Machine-learning and data-mining approaches can be used together [33.78].

### 33.5.5 Finding Networks

Sometimes a set of entities interact together, in a way that governs some process. This is termed a network. An example is a gene network, that is, a set of genes and proteins that interact to govern a biological process [33.79]. Gene networks can be discovered through a computational process that combines different sources of evidence about interactions between genes and regulator proteins [33.66].

Finding networks frequently involves the use of optimization methods, to find a set of hypothesized interactions that maximizes some objective function, based on statistical measures of the strength of these interactions. The objective function can include a complexity penalty term, aimed at preventing overfitting and ensuring network sparseness [33.80]. Computational implementation of exhaustive search, clustering, or an optimization method which moves individual nodes to increase connectivity, can find networks of interacting molecules [33.81]. The structure and parameters of ecological networks, denoting how species interact in an ecosystem, can be estimated using computational methods based on an analysis of biological flows from prey to predator species [33.82, 83]. Combinatorial optimization techniques can be used to find the most important members of a social network [33.84].

It is often important to characterize probabilistic dependencies between variables in a network, and, where feasible, to seek evidence for causality. Probabilistic dependencies can be represented in a form of directed acyclic graph (i. e., where any relationship between two



**Fig. 33.1** Data-mining methodology for the Warmr system. The inputs are descriptions of compounds. Warmr goes through successive rounds of generating new patterns by adding to existing patterns and pruning patterns that do not occur frequently. At each level another logical condition is added to the pattern. The maximum level searched is limited by preset computer resources (after [33.65])

nodes A and B has a direction and there are no cycles) known as a Bayesian network (BN) [33.85]. In a BN, a node is conditionally independent of its nondescendants given its parents. One of the more successful computational approaches in finding such networks involves first connecting all nodes with edges, then deleting edges between variables that are independent, or conditionally independent on a subset of the remaining variables [33.67]; another step involves inferring the direction of the remaining edges from conditional dependencies. See *Haughton et al.* [33.86] for a review of computational methods. Such networks can provide evidence for causality [33.87, 88], though this approach has its critics [33.89].

Closely related to the discovery of networks is the discovery of ancestral relationships between entities. This is of particular interest in biology, for the discovery of phylogenetic relationships [33.90]. An important technique is the computational application of maximum likelihood methods, which seeks the phylogenetic tree that maximizes the probability of the observations given the tree [33.68, 91].

### 33.5.6 Evolutionary Computation

With the continuing development of faster computers, it has become possible to evaluate increasingly large

sets of candidate hypotheses, models, or theories for those which are consistent with available data, and which are admissible on other grounds such as parsimony and plausibility. How to generate the candidate hypotheses, models, or theories to be tested remains a serious challenge. Evolutionary computation is a process for searching a large space of potential candidate solutions by mimicking the Darwinian process of natural selection. The two main methods of evolutionary computation are genetic algorithms, a method for finding a set of parameters, and genetic programming, a method for finding an algorithm, in the form of a computer program. Here, we consider the application of genetic programming [33.92, 93] to scientific discovery. Genetic programming involves creating an initially random population of programs. In successive iteration cycles, a new population is generated by preferential selection of those individuals which are better with respect to a defined fitness function. In addition, a small amount of random change is introduced into the new programs (mutation); this helps the process explore new regions of the search space. New variants are also created by randomly combining two existing members of the population (crossover).

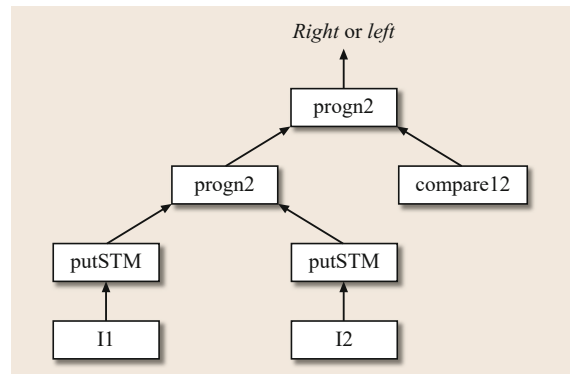
The development of process-based theories in psychology involves finding plausible psychological theories, constructed from basic processes, to explain the

results of psychological experiments. *Frias-Martinez* and *Gobet* [33.69] and *Lane et al.* [33.94] have applied genetic programming to this problem. A theory representation language allows a psychological theory to be represented by a computer program, built up from a set of operators corresponding to elementary psychological processes. Operators specify actions such as storing items in, or retrieving them from, short-term memory, and simple logical operations. This set of operators is based on previous theories in the psychological literature. In the genetic program, the fitness function measures how well predictions resulting from a candidate psychological theory fit a data set; it can also penalize theories that are too complex. *Frias Martinez* and *Gobet* [33.69] and *Lane et al.* [33.94] have shown that the technique can be used to generate theories to explain subjects' behavior in the delayed match to sample task [33.95], in which subjects must match a new image to one of two images they have seen previously. The theories developed using genetic programming generate predictions that fit the empirical data well. In addition, simple and surprising theories can be created [33.69]. Figure 33.2, adapted from [33.69], shows an example of such a theory. Input 1 is the new image, while input 2 is one of the previous images. The operators used in the theory are listed in Table 33.2. This capacity to generate unexpected theories offers the potential for these techniques to provide new insights into the psychological phenomena under study.

*Schmidt* and *Lipson* [33.70] describe a system for finding laws embedded in data sets. It uses a technique known as symbolic regression, and is based on genetic programming. A key test for a candidate law is that it should have good predictive ability, based on partial derivatives between pairs of variables. *Schmidt* and *Lipson* [33.70] represent an equation as a computer program. An initially random population of equations is created, and then a genetic programming algorithm is applied.

### 33.5.7 Automation of Scientific Experiments

A great deal of scientific experimentation involves labor-intensive cycles of setting up experiments, and collecting and analyzing data. The *robot scientist* [33.71, 96] automates this process, with the automation controlled by a computer algorithm. It iteratively collects and analyzes data and generates hypotheses,



**Fig. 33.2** Example of a theory generated by genetic programming. According to this theory, the delayed match to sample task is accomplished by comparing one of the new images to the original image; the second new image is not used (after [33.69])

**Table 33.2** Operators used in the theory shown in Fig. 33.2

Operator	Description
Progn2	Function: executes two inputs sequentially Input: Input1, Input2 Output: the output produced by Input2
PutSTM	Function: writes the input into short term memory (STM) Input: Input1 Output: the element written in STM (Input 1)
CompareI2	Function: compares positions 1 and 2 of STM and returns empty (NIL) if they are not equal or the element if they are equal Input: none Output: NIL or the element being compared

determined directly from data, for applications such as establishing which genes are involved in encoding enzymes. Some previous scientific discovery methods have incorporated in their logic the capacity to propose experiments [33.40, 59], but the robot scientist is probably the first practical application of a fully automated robotic system in which real world, rather than computational, experiments are formulated, executed, and analyzed.

Other automated systems, while arguably not as complete as the robot scientist, have been developed for the automation of the collection and analysis of data in hostile environments [33.97]. This applies in areas such as studies of underwater environments [33.98] and space exploration [33.99].

## 33.6 Discussion

The roots, nature, and psychology of scientific discovery (Sect. 33.1–33.3) provide a context for understanding the potential application of computational discovery systems, and their limitations. Section 33.4 has described examples of computational discovery systems in mathematics; these have characteristics, such as the use of logic, and of searching through large numbers of cases to find possible patterns, in common with scientific discovery systems.

The survey of scientific discovery methods in Sect. 33.5 has shown that computers can *do* scientific discovery, in the form of characterizing phenomena and generating scientific explanations. Yet the wholesale replacement of human scientists by computers is not on the horizon. Quite simply, computers are better suited to some aspects of the scientific discovery process than others.

What computers do particularly well is number crunching: carrying out large numbers of calculations and data operations, with great accuracy and at speed which exceeds human capability by several orders of magnitude. It follows that those areas where computational discovery methods have been most successful tend to be those for which number crunching can most readily be applied to discovery processes. The first significant discovery system, heuristic DENDRAL, illustrates this. The discovery of chemical topologies, and structures, involves consideration of a potentially enormous range of candidate solutions. Computational brute force enables a systematic search to be carried out, albeit within a limited parameter range. Rule-based methods for making the search methods efficient, and the facility to use specific knowledge to constrain the search, are important, but it is the number-crunching capacity of computers that made this approach feasible, even with the available computing power of the mid-1960s, when the system was first proposed. The later MECHEM system for finding chemical pathways similarly depends on number crunching to carry out a systematic search of possible pathways. This is not to suggest that humans use such forms of systematic search. As *Giza* [33.100] has argued, computational scientific discovery systems can proceed in a radically different manner from noncomputational methods, and employ criteria for choosing candidate discoveries that are different from those employed by human scientists.

The systems that are primarily rule-based, such as BACON [33.20], GLAUBER [33.60] and KEKADA [33.40], are those that come closest to attempting to replicate how humans carry out scientific discovery, at least to the extent that humans use inductive rules. It has been argued that this approach is not

successful, in that these methods have not discovered any new and important rules, generalizations, or laws: See *Gillies* [33.30] for a discussion of this claim. This does not detract from the possible use of such systems to shed light on understanding an important part of the basis for how humans make scientific discoveries. It does, however, suggest that systems that are primarily rule-based do not constitute the most practically important tools currently available for computational scientific discovery.

Classification, machine vision and data-mining methods provide important computational discovery tools, which fully utilize a computer's capability to process large amounts of information [33.65, 73]. These methods essentially involve detecting or recognizing patterns, which is similar in principle to finding conjectures in mathematics [33.52]. Automatic determination of network relationships is likely to prove to be increasingly important [33.80], and may benefit from future advances in statistical modeling and optimization techniques, but human input is required for problems where a strong degree of judgment and prior knowledge is required to construct causal relationships [33.89]. In the specific field of determining phylogenetic relationships, using established maximum-likelihood methods, computational discovery has been very successful [33.91].

The use of evolutionary computational methods in scientific discovery is relatively recent, and shows promise. The application to psychological theories [33.69] opens up the possibility of automatically discovering useful models involving a sequence of simple processes. The approach, however, requires human judgment in specifying the characteristics of operators used to construct models, in the extraction of data from published results, and in the interpretation of results. Automation of experimental procedures will continue to be important in hostile environments [33.97], and may find further application for relatively predictable cycles of hypothesis proposal and testing [33.71].

For computers to be really effective at the kinds of discovery methods that humans use will require basic developments in strong artificial intelligence, that is, computers which can replicate general human capacities. Is such a development likely? *Penrose* [33.101] has suggested that human consciousness may be dependent on nonalgorithmic physical processes, and not representable by a computational algorithm, and that this poses a serious barrier to strong AI; by way of example he draws on the way mathematical truth is discovered, arguing that there is no general algorithm for determining the truth of a mathematical proposition. Among the critics of this line of reasoning, however, is

Dennett [33.102], who suggests that people make use of reasoning methods that evolved for reasons related to survival, and that these may happen to work well for assessing mathematical propositions a high proportion of the time. Another possible barrier to strong AI is that inference depends on context [33.103, 104]: So far, humans have proved better than machines in being able to judge how to make good generalizations. Gillies [33.30] argues that humans have a *political superiority* to computers because computers are designed and built by humans in order to carry out human tasks; it follows that

“if a computer is designed to solve problems a, b, c, . . . , it is likely to give rise to further problems x, y, z, . . . which the computer system itself will not be able to solve, but which will need some human thinking for their resolution.”

In any event, the development of strong AI does not appear to be on the immediate horizon. There is no clear consensus on the prospects for strong AI in the medium and long-term future.

A recurring theme in this review has been that the application of computational scientific discovery systems requires human input – in areas such as judging putative causal relationships [33.89], judging the correct context for making generalizations [33.103, 104],

and specifying operators used for constructing evolutionary computational models [33.69]. An important open question is the extent to which such aspects of human judgmental skill could eventually be automated, but there does not seem to be a strong prospect of substantial automation of this sort in the near future.

The immediate trend will be for more development of discovery systems oriented to the kind of things computers do well. This is the case in several modern applications of AI, where problems are solved using the computer’s ability to use large quantities of data, rather than in mimicry of how humans do the same task. For example, *Google translate* does not try to construct models of what a sentence means, but instead scans the internet for data on how each sentence may be translated. In this way, the problem can be solved with a computer, but without employing a method that is human-like. Similarly, it is likely that areas of scientific discovery will be achieved by computers working in computer-like ways, with humans providing added value in terms of synthesis or creative insights. This harnessing of the complementary qualities of humans and machines is likely to increase the rate at which scientific discoveries are made [33.105].

In conclusion, computational scientific discovery methods are an increasingly important tool in science. But the role of the human scientist remains, for the foreseeable future, essential.

## References

- 33.1 H.K. Versteeg, W. Malalasekera: *An Introduction to Computational Fluid Dynamics: The Finite Volume Method* (Pearson Education, Harlow 2007)
- 33.2 D.W. Heermann: *Computer-Simulation Methods in Theoretical Physics* (Springer, Berlin 1990)
- 33.3 J. Holland, K. Holyoak, R. Nisbett, P. Thagard: *Induction: Processes of Inference, Learning, and Discovery* (MIT, Cambridge 1986)
- 33.4 D.W. Stephens: Change, regularity, and value in the evolution of animal learning, *Behav. Ecol.* **2**, 77–89 (1991)
- 33.5 S.M. Reader, K.N. Laland: Social intelligence, innovation, and enhanced brain size in primates, *Proc. Nat. Acad. Sci.* **99**, 4436–4441 (2002)
- 33.6 N.J. Emery, N.S. Clayton: The mentality of crows: Convergent evolution of intelligence in corvids and apes, *Science* **306**, 1903–1907 (2004)
- 33.7 A.M. Auersperg, B. Szabo, A.M. von Bayern, A. Kacelnik: Spontaneous innovation in tool manufacture and use in a Goffin’s cockatoo, *Curr. Biol.* **22**, R903–R904 (2012)
- 33.8 M.H. Christiansen, S. Kirby: Language evolution: Consensus and controversies, *Trends Cogn. Sci.* **7**, 300–307 (2003)
- 33.9 J. Diamond: *Guns, Germs and Steel* (Vintage, London 2005)
- 33.10 P. Curd: Presocratic philosophy. In: *The Stanford Encyclopedia of Philosophy* (Winter 2012 Edition), ed. by E.N. Zalta (2012) <http://plato.stanford.edu/archives/win2012/entries/presocratics/>
- 33.11 C. Shields: Aristotle. In: *The Stanford Encyclopedia of Philosophy* (Spring 2014 Edition), ed. by E.N. Zalta (2014) <http://plato.stanford.edu/archives/spr2014/entries/aristotle/>
- 33.12 J.R. Platt: Strong inference, *Science* **146**, 347–353 (1964)
- 33.13 P.M.S. Blackett: Memories of Rutherford. In: *Rutherford at Manchester*, ed. by J.B. Birks (Heywood, London 1962) pp. 102–113
- 33.14 L.A. Geddes: Looking back how measuring electric current has improved through the ages, *IEEE Potentials* **15**, 40–42 (1996)
- 33.15 W.L. Bragg: The diffraction of short electromagnetic waves by a crystal, *Proc. Camb. Philos. Soc.* **17**, 43–57 (1913)
- 33.16 L. Pauling, R.B. Corey, H.R. Branson: The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain, *Proc. Nat. Acad. Sci.* **37**, 205–211 (1951)

- 33.17 J.D. Watson, F.H. Crick: Molecular structure of nucleic acids, *Nature* **171**, 737–738 (1953)
- 33.18 D.B. Kell, S.G. Oliver: Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis driven science in the post genomic era, *Bioessays* **26**, 99–105 (2004)
- 33.19 P. Langley: The computational support of scientific discovery, *Int. J. Human-Comput. Stud.* **53**, 393–410 (2000)
- 33.20 P. Langley, H.A. Simon, G. Bradshaw, H.A. Simon, J.M. Zytkow: *Scientific Discovery: Computational Explorations of the Creative Processes* (MIT, Cambridge 1987)
- 33.21 A. Machado, F.J. Silva: Toward a richer view of the scientific method: The role of conceptual analysis, *Am. Psychol.* **62**, 671–681 (2007)
- 33.22 R. Greenler: *Rainbows, Halos, and Glories* (Cambridge Univ. Press, Cambridge 1980)
- 33.23 W.G. Rees, C.M. Roach, C.H.F. Glover: Inversion of atmospheric refraction data, *JOSA A* **8**, 330–338 (1991)
- 33.24 P.D. Sozou: Inversion of mirage data: An optimization approach, *JOSA A* **11**, 125–134 (1994)
- 33.25 M. Bertero, T.A. Poggio, V. Torre: Ill-posed problems in early vision, *Proc. IEEE* **76**, 869–889 (1988)
- 33.26 M.V. Afonso, J.M. Bioucas-Dias, M.A. Figueiredo: An augmented Lagrangian approach to the constrained optimization formulation of imaging inverse problems, *IEEE Trans. Image Process.* **20**, 681–695 (2011)
- 33.27 H.Y. Li, C.Y. Yang: A genetic algorithm for inverse radiation problems, *Int. J. Heat Mass Transf.* **40**, 1545–1549 (1997)
- 33.28 C.L. Karr, I. Yakushin, K. Nicolosi: Solving inverse initial-value, boundary-value problems via genetic algorithm, *Eng. App. Artif. Intell.* **13**, 625–633 (2000)
- 33.29 D.K. Karpouzos, F. Delay, K.L. Katsifarakis, G.D. Marsily: A multipopulation genetic algorithm to solve the inverse problem in hydrogeology, *Water Resour. Res.* **37**, 2291–2302 (2001)
- 33.30 D. Gillies: *Artificial Intelligence and Scientific Method* (Oxford Univ. Press, Oxford 1996)
- 33.31 F. Bacon: *Novum Organum* (Open Court, Chicago 1994), ed. by P. Urbach, J. Gibson, originally published in 1620
- 33.32 K.R. Popper: *Conjectures and Refutations: The Growth of Scientific Knowledge* (Routledge and Kegan Paul, London 1963)
- 33.33 K.R. Popper: *The Logic of Scientific Discovery* (Unwin Hyman, London 1990), 14th impression
- 33.34 D. Campbell: Blind variation and selective retention in creative thought as in other knowledge processes, *Psychol. Rev.* **67**, 380–400 (1960)
- 33.35 D. Simonton: *Origins of Genius* (Oxford Univ. Press, Oxford 1999)
- 33.36 D. Klahr, K. Dunbar: Dual space search during scientific reasoning, *Cogn. Sci.* **12**, 1–48 (1988)
- 33.37 K. Dunbar, J. Fugelsang: Scientific thinking and reasoning. In: *The Cambridge Handbook of Thinking and Reasoning*, ed. by K.J. Holyoak, R.G. Morrison (Cambridge Univ. Press, Cambridge 2005) pp. 705–725
- 33.38 Y. Qin, H.A. Simon: Laboratory replication of scientific discovery processes, *Cogn. Sci.* **14**, 281–312 (1990)
- 33.39 P. Langley, G. Bradshaw, H.A. Simon: BACON 5: The discovery of conservation laws, *Proc. 7th Int. Jt. Conf. Artif. Intell.*, Br. Columbia (AAAI, Palo Alto 1981) pp. 121–126
- 33.40 D. Kulkarni, H.A. Simon: The processes of scientific discovery: The strategy of experimentation, *Cogn. Sci.* **12**, 139–175 (1988)
- 33.41 H.A. Krebs: The discovery of the ornithine cycle of urea synthesis, *Biochem. Educ.* **1**, 19–23 (1973)
- 33.42 M.A. Boden: Creativity and artificial intelligence, *Artif. Intell.* **103**, 347–356 (1998)
- 33.43 P. Thagard, K.J. Holyoak, G. Nelson, D. Gochfeld: Analog retrieval by constraint satisfaction, *Artif. Intell.* **46**, 259–310 (1990)
- 33.44 E.P. Wigner: The unreasonable effectiveness of mathematics in the natural sciences. Richard Courant lecture in mathematical sciences delivered at New York University, May 11, 1959, *Commun. Pure Appl. Math.* **13**, 1–14 (1960)
- 33.45 S. Colton: Computational discovery in pure mathematics. In: *Computational Discovery of Scientific Knowledge*, Lecture Notes in Computer Science, Vol. 4660, ed. by S. Džeroski, L. Todorovski (Springer, Berlin Heidelberg 2007) pp. 175–201
- 33.46 S. Colton, A. Bundy, T. Walsh: On the notion of interestingness in automated mathematical discovery, *Int. J. Human-Comput. Stud.* **53**, 351–375 (2000)
- 33.47 C.E. Larson: A survey of research in automated mathematical conjecture-making, *DIMACS Ser. Discrete Math. Theor. Comput. Sci.* **69**, 297–318 (2005)
- 33.48 A. Newell, J.C. Shaw, H.A. Simon: Elements of a theory of human problem solving, *Psychol. Rev.* **65**, 151–166 (1958)
- 33.49 A.N. Whitehead, B. Russell: *Principia Mathematica*, Vol. 1 (Cambridge Univ. Press, Cambridge 1910)
- 33.50 F. Gobet, P.C.R. Lane: Human problem solving: Beyond Newell et al.'s (1958) elements of a theory of human problem solving. In: *Cognitive Psychology: Revisiting the Classic Studies*, ed. by D. Groome, M.W. Eysenck (Sage, Thousand Oaks 2015) pp. 133–145
- 33.51 H. Wang: Toward mechanical mathematics. In: *Automation of Reasoning: Classical Papers on Computational Logic 1957–1966*, ed. by J. Siekmann, G. Wrightson (Springer, Berlin 1983) pp. 244–264
- 33.52 D.B. Lenat: *AM: An Artificial Intelligence Approach to Discovery in Mathematics as Heuristic Search* (Dept. Computer Science, Stanford Univ., Stanford 1976)
- 33.53 R. Davis, D.B. Lenat: *Knowledge-Based Systems in Artificial Intelligence* (McGraw-Hill, New York 1982)
- 33.54 G.D. Ritchie, F.K. Hanna: AM: A case study in AI methodology, *Artif. Intell.* **23**, 249–268 (1984)

- 33.55 D.B. Lenat, J.S. Brown: Why AM and EURISKO appear to work, *Artif. Intell.* **23**, 269–294 (1984)
- 33.56 S. Fajtlowicz: On conjectures of Graffiti, *Ann. Discrete Math.* **38**, 113–118 (1988)
- 33.57 E. Delavina: Bibliography on conjectures, methods and applications of Graffiti (2016), <http://cms.dt.uh.edu/faculty/delavinae/research/wowref.htm>
- 33.58 R.K. Lindsay, B.G. Buchanan, E.A. Feigenbaum, J. Lederberg: DENDRAL: A case study of the first expert system for scientific hypothesis formation, *Artif. Intell.* **61**, 209–261 (1993)
- 33.59 R.E. Valdes-Perez: Theory-driven discovery of reaction pathways in the MECHEM system, *Proc. 10th Natl. Conf. Artif. Intell.*, San Jose (AAAI, Palo Alto 1992) pp. 63–69
- 33.60 J.M. Zytkow, H.A. Simon: Normative systems of discovery and logic of search, *Synthese* **74**, 65–90 (1988)
- 33.61 S. Muggleton, R.D. King, J.E. Sternberg: Protein secondary structure prediction using logic-based machine learning, *Protein Eng.* **5**, 647–657 (1992)
- 33.62 M.C. Storrie-Lombardi, O. Lahav, L. Sodre, L.J. Storrie-Lombardi: Morphological classification of galaxies by artificial neural networks, *Mon. Not. R. Astron. Soc.* **259**, 8P–12P (1992)
- 33.63 L. Shamir: Automatic morphological classification of galaxy images, *Mon. Not. R. Astron. Soc.* **399**, 1367–1372 (2009)
- 33.64 N. Tiffin, J.F. Kelso, A.R. Powell, H. Pan, V.B. Bajic, W.A. Hide: Integration of text-and data-mining using ontologies successfully selects disease gene candidates, *Nucleic Acids Res.* **33**, 1544–1552 (2005)
- 33.65 R.D. King, A. Srinivasan, L. Dehaspe: Warmr: A data mining tool for chemical data, *J. Comput.-Aided Mol. Des.* **15**, 173–181 (2001)
- 33.66 Z. Bar-Joseph, G.K. Gerber, T.I. Lee, N.J. Rinaldi, J.Y. Yoo, F. Robert, D.K. Gifford: Computational discovery of gene modules and regulatory networks, *Nat. Biotechnol.* **21**, 1337–1342 (2003)
- 33.67 P. Spirtes, C. Glymour: An algorithm for fast recovery of sparse causal graphs, *Soc. Sci. Comput. Rev.* **9**, 62–72 (1991)
- 33.68 S. Guindon, O. Gascuel: A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood, *Syst. Biol.* **52**, 696–704 (2003)
- 33.69 E. Frias-Martinez, F. Gobet: Automatic generation of cognitive theories using genetic programming, *Minds Mach.* **17**, 287–309 (2007)
- 33.70 M. Schmidt, H. Lipson: Distilling free-form natural laws from experimental data, *Science* **324**, 81–85 (2009)
- 33.71 R.D. King, J. Rowland, S.G. Oliver, M. Young, W. Aubrey, E. Byrne, M.L. Kata, K. Karkham, P. Pir, L.N. Soldatova, A. Sparkes, K.E. Whelan, A. Care: The automation of science, *Science* **324**, 85–89 (2009)
- 33.72 B.G. Buchanan, E.A. Feigenbaum: DENDRAL and Meta-DENDRAL: Their applications dimension, *Artif. Intell.* **11**, 5–24 (1978)
- 33.73 N. Orlov, L. Shamir, T. Macura, J. Johnston, D.M. Eckley, I.G. Goldberg: WND-CHARM: Multi-purpose image classification using compound image transforms, *Pattern Recognit. Lett.* **29**, 1684–1693 (2008)
- 33.74 U. Fayyad, G. Piatetsky-Shapiro, P. Smyth: From data mining to knowledge discovery in databases, *AI Magazine* **17**, 37 (1996)
- 33.75 X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, D. Steinberg: Top 10 algorithms in data mining, *Knowl. Inf. Sys.* **14**, 1–37 (2008)
- 33.76 L. Hirschman, J.C. Park, J. Tsujii, L. Wong, C.H. Wu: Accomplishments and challenges in literature data mining for biology, *Bioinformatics* **18**, 1553–1561 (2002)
- 33.77 C. Perez-Iratxeta, P. Bork, M.A. Andrade: Association of genes to genetically inherited diseases using data mining, *Nat. Genet.* **31**, 316–319 (2002)
- 33.78 N.M. Ball, R.J. Brunner: Data mining and machine learning in astronomy, *Int. J. Mod. Phys. D* **19**, 1049–1106 (2010)
- 33.79 H. Kitano: Systems biology: A brief overview, *Science* **295**, 1662–1664 (2002)
- 33.80 M. Hecker, S. Lambeck, S. Toepfer, E. Van Someren, R. Guthke: Gene regulatory network inference: Data integration in dynamic models—a review, *Biosys.* **96**, 86–103 (2009)
- 33.81 V. Spirin, L.A. Mirny: Protein complexes and functional modules in molecular networks, *Proc. Natl. Acad. Sci.* **100**, 12123–12128 (2003)
- 33.82 R.E. Ulanowicz: Quantitative methods for ecological network analysis, *Comput. Biol. Chem.* **28**, 321–339 (2004)
- 33.83 P. Kavanagh, N. Newlands, V. Christensen, D. Pauly: Automated parameter optimization for ecopath ecosystem models, *Ecol. Model.* **172**, 141–149 (2004)
- 33.84 S.P. Borgatti: Identifying sets of key players in a social network, *Comput. Math. Organ. Theory* **12**, 21–34 (2006)
- 33.85 Z. Ghahramani: An introduction to hidden Markov models and Bayesian networks, *Int. J. Pattern Recog. Artif. Intell.* **15**, 9–42 (2001)
- 33.86 D. Haughton, A. Kamis, P.A. Scholten: A review of three directed acyclic graphs software packages: MIM, tetrad, and WinMine, *Am. Stat.* **60**, 272–286 (2006)
- 33.87 D.M. Hausman, J. Woodward: Independence, invariance and the causal Markov condition, *Br. J. Phil. Sci.* **50**, 521–583 (1999)
- 33.88 C. Glymour: Learning, prediction and causal Bayes nets, *Trends Cogn. Sci.* **7**, 43–48 (2003)
- 33.89 N. Cartwright: Causation: One word, many things, *Phil. Sci.* **71**, 805–820 (2004)
- 33.90 J.P. Huelsenbeck, F. Ronquist, R. Nielsen, J.P. Bollback: Bayesian inference of phylogeny and its impact on evolutionary biology, *Science* **294**, 2310–2314 (2001)
- 33.91 Z. Yang: PAML 4: Phylogenetic analysis by maximum likelihood, *Mol. Biol. Evol.* **24**, 1586–1591 (2007)



- 33.92 J. Koza: *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, Vol. 1 (MIT, Cambridge Massachusetts 1992)
- 33.93 R. Poli, W. Langdon, N. McPhee: A field guide to genetic programming, <http://www.gp-field-guide.org.uk> (2008)
- 33.94 P.C.R. Lane, P.D. Sozou, M. Addis, F. Gobet: Evolving process-based models from psychological data using genetic programming. In: *AISB50: Selected Papers*, ed. by M. Bishop, K. Devlin, Y. Erden, R. Kibble, S. McGregor, M. Majid al-Rifaie, A. Martin, M. Figueroa, S. Rainey (AISB, London 2015) pp. 144–149
- 33.95 L. Chao, J. Haxby, A. Martin: Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects, *Nat. Neurosci.* **2**, 913–919 (1999)
- 33.96 A. Sparkes, W. Aubrey, E. Byrne, A. Clare, M.N. Khan, M. Liakata, R.D. King: Towards robot scientists for autonomous scientific discovery, *Autom. Exp.* **2**, 1 (2010)
- 33.97 J.G. Bellingham, K. Rajan: Robotics in remote and hostile environments, *Science* **318**, 1098–1102 (2007)
- 33.98 I. Vasilescu, K. Kotay, D. Rus, M. Dunbabin, P. Corke: Data collection, storage, and retrieval with an underwater sensor network, *Proc. 3rd Int. Conf. Embed. Networked Sens. Syst.* (2005) pp. 154–165
- 33.99 J. Schwendner, F. Kirchner: Space Robotics: An overview of challenges, applications and technologies, *KI-Künstliche Intell.* **28**, 71–76 (2014)
- 33.100 P. Giza: Automated discovery systems and scientific realism, *Minds Mach.* **12**, 105–117 (2002)
- 33.101 R. Penrose: *The Emperor's New Mind: Concerning Computers, Minds and the Laws of Physics* (Oxford Univ. Press, Oxford 1989)
- 33.102 D.C. Dennett: Betting your life on an algorithm, *Behav. Brain Sci.* **13**, 660–661 (1990)
- 33.103 G. Gigerenzer: Strong AI and the problem of second order algorithms, *Behav. Brain Sci.* **13**, 663–664 (1990)
- 33.104 M. Addis, P.D. Sozou, P.C. Lane, F. Gobet: Computational scientific discovery and cognitive science theories, *Proc. IACAP*, ed. by V. Müller (Springer, Heidelberg 2016)
- 33.105 C. Glymour: The automation of discovery, *Daedalus* **133**, 69–77 (2004)

# 34. Computer Simulations and Computational Models in Science

Cyrille Imbert

Computational science and computer simulations have significantly changed the face of science in recent times, even though attempts to extend our computational capacities are by no means new and computer simulations are more or less accepted across scientific fields as legitimate ways of reaching results (Sect. 34.1). Also, a great variety of computational models and computer simulations can be met across science, in terms of the types of computers, computations, computational models, or physical models involved and they can be used for various types of inquiries and in different scientific contexts (Sect. 34.2). For this reason, epistemological analyses of computer simulations are contextual for a great part. Still, computer simulations raise general questions regarding how their results are justified, how computational models are selected, which type of knowledge is thereby produced (Sect. 34.3), or how computational accounts of phenomena partly challenge traditional expectations regarding the explanation and understanding of natural systems (Sect. 34.4). Computer simulations also share various epistemological features with experiments and thought experiments; hence, the need for transversal analyses of these activities (Sect. 34.5). Finally, providing a satisfactory and fruitful definition of computer simulations turns out to be more difficult than expected, partly because this notion is at the crossroads of difficult questions like the nature of representation and computation or the success of scientific inquiries (Sect. 34.6). Overall, a pointed analysis of computer simulations in parallel requires developing insights about the evolving place of human capacities and humans within (computational) science (Sect. 34.7).

34.1	<b>Computer Simulations in Perspective</b>	736
34.1.1	The Recent Philosophy of Scientific Models and Computer Simulations	736
34.1.2	Numerical Methods and Computational Science: An Old Tradition	737
34.1.3	A More or Less Recent Adoption Across Scientific Fields	738
34.1.4	Methodological Caveat	738
34.2	<b>The Variety of Computer Simulations and Computational Models</b>	739
34.2.1	Working Characterization	739
34.2.2	Analog Simulations and Their Specificities	740
34.2.3	Digital Machines, Numerical Physics, and Types of Equivalence	741
34.2.4	Non-Numerical Digital Models	741
34.2.5	Nondeterministic Simulations	742
34.2.6	Other Types of Computer Simulations	742
34.3	<b>Epistemology of Computational Models and Computer Simulations</b>	743
34.3.1	Computer Simulations and Their Scientific Roles	743
34.3.2	Aspects of the Epistemological Analysis of Computer Simulations	744
34.3.3	Selecting Computational Models and Practices	746
34.3.4	The Production of 'New' Knowledge: In What Sense?	748
34.4	<b>Computer Simulations, Explanation, and Understanding</b>	750
34.4.1	Traditional Accounts of Explanation	751
34.4.2	Computer Simulations: Intrinsically Unexplanatory?	751
34.4.3	Computer Simulations: More Frequently Unexplanatory?	752
34.4.4	Too Replete to Be Explanatory? The Era of Lurking Suspicion	754
34.4.5	Bypassing the Opacity of Simulations	757
34.4.6	Understanding and Disciplinary Norms	758
34.5	<b>Comparing: Computer Simulations, Experiments and Thought Experiments</b>	758
34.5.1	Computational Mathematics and the Experimental Stance	759
34.5.2	Common Basal Features	759
34.5.3	Are Computer Simulations Experiments?	762
34.5.4	Knowledge Production, Superiority Claims, and Empiricism	765
34.5.5	The Epistemological Challenge of Hybrid Methods	767

34.6	<b>The Definition of Computational Models and Simulations</b> .....	767	34.7.1	The Partial Mutation of Scientific Practices .....	774
34.6.1	Existing Definitions of Simulations.....	768	34.7.2	The New Place of Humans in Science...	774
34.6.2	Pending Issues .....	770	34.7.3	Analyzing Computational Practices for Their Own Sake.....	774
34.6.3	When Epistemology Cross-Cuts Ontology.....	773	34.7.4	The Epistemological Treatment of New Issues .....	775
34.7	<b>Conclusion: Human-Centered, but no Longer Human-Tailored Science</b> ....	773	References.....		775

For several decades, much of science has been computational, that is, scientific activity where computers play a central and essential role. Still, computational science is larger than the set of activities resorting to computer simulations. For example, experimental science, from vast experiments in nuclear physics at the European Organization for Nuclear Research (CERN) to computational genomics, relies heavily on computers and computational models for data acquisition and their treatment, but does not seem to involve computer simulations proper. In any case, there is a great and still proliferating variety of types of computer simulations, which are used for different types of inquiries and in different types of theoretical contexts. For this reason, one should be careful when describing the philosophy of computer simulations and unjustified generalizations should be avoided. At the same time, how much the development of computer simulations has been changing science is a legitimate question. Com-

puter simulations raise questions about the traditional conceptualization of science in terms of experiments, theories and models, about the ways that usual scientific activities like predicting, theorizing, controlling, or explaining are carried out with the help of these new tools and, more generally, how the production of scientific knowledge by human creatures is modified by computer simulations. Importantly, while the specific philosophical analysis of computer simulations is recent (even if it was preceded by the development of the philosophical study of scientific models) and computational science is a few decades old, the development of computational tools and mathematical techniques aimed at bypassing the complexity of problems belongs to a much older tradition. This means that claims about how much computer simulations change science, and how much a closer attention to computer simulations should change our picture of scientific activity, are questions to be treated with circumspection.

## 34.1 Computer Simulations in Perspective

When discussing philosophical and epistemological issues related to computational models and computer simulations, different chronologies should be kept in mind. The blossoming of the philosophy of models and simulations, within the philosophy of science is something recent (Sect. 34.1.1). The development of techniques aimed at extending our inferential and computational powers corresponds to a longer trend, even if the recent invention of powerful digital machines has changed the face of computational science (Sect. 34.1.2). Finally, the acceptance of computer simulations as legitimate scientific tools across the different fields goes at various paces (Sect. 34.1.3). This means that, even if computer simulations do change the face of science, much care is needed when analyzing the aspects of science which are actually changed, and how we should modify our picture of science when we adopt a computer simulation-based perspective (Sect. 34.1.4).

### 34.1.1 The Recent Philosophy of Scientific Models and Computer Simulations

While the use of computer simulations in the empirical sciences, in particular physics, developed after the construction of the (ENIAC) computer during World War II [34.1], and started changing how the empirical sciences were practiced, for decades computer-related discussions among philosophers were primarily focused on the development of artificial intelligence and the analysis of human cognition. Particularly active were debates in philosophy of mind regarding the question of the *computational theory of the mind*, that is, whether the mind can be likened to a digital computer, and in particular to a classical machine employing rules and symbolic representations [34.2–6]. However, within the mainstream philosophy of science, continued interest for computational science, compu-

tational models, and digital simulations of empirical systems as such did not really start until the early 1990s, with articles by *Humphreys* [34.7, 8], *Rohrlich* [34.9] or *Hartmann* [34.10]. (Such a description of the field is necessarily unfair to earlier works about the use of computer simulations in the empirical sciences. Particular mention should be given to the works of *Bunge* [34.11] or *Simon* [34.12].) An article by *Hughes* about the investigations of the Ising model [34.13], a special issue of *Science in Context* edited by *Sismondo* [34.14] and works by *Winsberg* [34.15–17], who completed his Ph. D. in 1999 about computer simulations, also contributed to the development of this field. Finally, in 2006, the *Models and Simulations* Conference took place, which was the first of what was to become a still active conference series, which has contributed to making the issue of computational science one of the fields of philosophy of science.

Philosophical works about scientific models, a very close field, were not significantly older. The importance of the notion of set-theoretic model had been emphasized by partisans of the model-theoretic view of theories in the 1970s, but, if one puts aside works by pioneers like *Black* [34.18] or *Hesse* [34.19], this did not launch investigations about scientific models proper. Overall, the intense epistemological study of models did not start until the 1980s, with in particular a seminal article by *Redhead* about scientific models in physics [34.20]. Members of the *Stanford School* also argued against the view that science was unified and that theories played a dominant role in scientific activities such as the selection and construction of models [34.21], and conversely emphasized the autonomy of experimental and modeling practices. This context was appropriate for an independent investigation about the role of models in science, which bloomed at the end of the 1990s [34.22] and was further fed by a renewal of interest for the question of scientific representation [34.23–25]. These investigations of models paved the way for new studies focused neither on theories nor on experiments. However, while the difficulty to explore a model was already acknowledged in works by *Redhead* and *Cartwright*, interest for the actual modes of its exploration, in particular by computer simulations, was not triggered. Indeed, the focus remained on the effects of the complexity of the inquiry on scientific representations, with studies about simplifications, approximations, or idealizations (Even *Laymon's* 1990 paper [34.26], in spite on its apparent focus on computer simulation, mainly deals with the nature of approximation and what it is to accept or believe a theory.), or how to articulate the model-theoretic view of theories and the uses of models and representations in actual scientific practices, by taking into

account scientific users, qua intentional cognitive creatures [34.27, 28], and their cognitively constrained ways to handle models by means of inferences, graphs, pictures or diagrams (*Kulvicki* [34.29], *Giardino* Chap. 22, this volume; *Bechtel* Chap. 27, this volume). Overall, in spite of the close connection within scientific practice between the uses of models and their computational explorations, the issue of computational models and computer simulations was not seen clearly as a fruitful field of inquiry of its own, this trend of thought being explicitly and vividly brought to the fore in 2008 in a deliberately provocative paper by *Frigg* and *Reiss* [34.30].

### 34.1.2 Numerical Methods and Computational Science: An Old Tradition

The second relevant chronology is that of the advancement in attempts to solve complex mathematical problems by developing computing machines and mathematical methods. Importantly, while the development of digital computers in the mid-twentieth century changed the face of scientific computation, humans did not wait for this decisive breakthrough to extend their mathematical and computational powers. Further, as *Mahoney* wrote it, “the computer is not one thing, but many different things, and the same holds true of computing” [34.31], and it is only in the twentieth century that different historical strands related to logic, mathematics, or technologies came together. On the one hand, early mathematical procedures, like Newton’s method to find the roots of real-valued functions, or Euler’s method to solve ordinary differential equations, were developed to provide numerical approximations for problems in numerical analysis. This field was already important to investigate physical systems but, with the advent of digital computers, it became a crucial part of (computational) science. On the other hand, mechanical calculating tools, such as abacuses or slide rules, were used from the Antiquity through the centuries. The invention by Pascal of a device (the *Pascaline*) to perform additions and subtractions, and the conceptualization by Babbage of mechanical computing systems fed by punched cards, were important additional steps. *Human computers* were also used. For example, in 1758, Clairaut predicted the return of Halley’s comet, by dividing the computational work with other colleagues [34.32]. Gaspard de Prony produced the logarithm and trigonometric tables in the early nineteenth century by dividing the computational tasks into elementary operations, which were carried out by unemployed hairdressers with little education. Human computers were used during World

War I to compute artillery tables and World War II to help with the Manhattan project [34.33, 34]. Finally, mechanical analog computers were developed for scientific purposes by engineers and scientists like Thomson or Kelvin, in the late nineteenth century, Vannevar Bush, between the two World Wars, or Enrico Fermi, in 1947, and such computers were used till the 1960s. Finally, even in the digital era, new technological change can have a large impact. For decades, access to computational resources was difficult and only possible in the framework of big projects. Typically, *Schelling's* first simulations of residential segregation [34.35] were hand made. An important recent step has been the development of personal computers, which has brought more flexibility and may have triggered the development of new modeling practices [34.36].

### 34.1.3 A More or Less Recent Adoption Across Scientific Fields

The development of computational science and the use of computational models and simulation methods vary from one field to another. Since the 1940s onward, computer simulations have been used in physics, and computers were also used in artificial intelligence as early as the late 1950s. However, some fields have resisted such methods, and still do, as far as commonly accepted mainstream methods are concerned. Typically, the development of computational models and computer simulations in the human and social sciences, with the possibility of analyzing diachronic interactions between agents (versus static models describing equilibria) is much more recent. As emphasized earlier, *Schelling's* initial dynamic model of segregation was first run manually in 1969. Attempts to use computational science to predict social and economic behavior were globally met with suspicion in the 1960s and 1970s, all the more since these studies were often carried out by scholars who did not belong to well-entrenched traditions in these fields (such as scientists studying complexity, including human behavior, in institutions like the Santa Fe Institute). Overall, in economics, computer simulations are still not accepted [34.37]. Similarly, the development of a specific (and still somewhat distinct) subfield using computational methods to analyze social phenomena is recent, with the edition by *Hegselmann et al.* of the volume *Modelling and Simulation in the Social Sciences from the Philosophy of Science Point of View* [34.38], the need felt to create, in 1998, the *Journal of Artificial Societies and Social Simulation* and the publication in 2005 of the handbook *Simulation for the Social Scientist* by *Gilbert and Troitzsch* [34.39].

### 34.1.4 Methodological Caveat

These different chronological perspectives call for the following comments.

First, philosophers should be careful when developing an epistemology of computational models and computer simulations. Modeling and simulating practices have been developed in various epistemic contexts in scientific fields in which well-entrenched theories are more or less present and which have different methodological and scientific norms. Thus, the role of computer simulations and their epistemological assessment can significantly differ from one case to another, and bold generalizations should be carefully justified or avoided. As just mentioned, the use of computer simulations is central and accepted in fields like climate science (even if it raises important problems) but is still regarded with great suspicion in fields like economics [34.37, 40].

Second, how much computational models and computer simulations correspond to epistemologically different practices, which should be described in terms of some *computational turn*, cannot be assumed, but should be investigated on a case-by-case basis regarding all potentially relevant aspects. This can be illustrated with the question of the tractability of scientific models. *Humphreys*, in his 2004 book *Extending Ourselves*, proposes the following two principles to analyze science: “It is the invention and deployment of tractable mathematics that drives much progress in the physical sciences”; and its converse version: “most scientific models are specifically tailored to fit, and hence are constrained by, the available mathematics” [34.41, pp. 55–56]. These two principles suggest both a continuist and discontinuist reading of the development of science. First, students of science need to assess which precise aspects of scientific practices have been changed by the development of computers and whether these changes should be seen as a scientific revolution, or simply as an extension of existing modes of reasoning [34.42]. In this perspective, questions about the tractability and complexity of models can no longer be ignored, and may be crucial to an understanding of how new branches of modeling and computational practices can develop and of how the dynamics of science can be qualitatively different [34.43]. At the same time, scientific practices were also constrained by the available mathematics before the advent of computers, and new findings in mathematics already paved the way for the development of new scientific practices. For example, *Lakatos* emphasizes that [34.44, p. 137]

“the greatness of the Newtonian programme comes partly from the development – by Newtonians of

classical infinitesimal analysis which was a crucial precondition for its success.”

From this point of view, a continuist reading is also required.

Third, the computational perspective may require partly revising the philosophical treatment of questions about science, and scientific representation in particular. Computer simulations are *actual* activities of investigation of scientific models, and, for this reason, the tractability and computational constraints that they face can hardly be ignored when analyzing them. They force us to adopt an in practice perspective, where what matters is not the logical content of representations (that is, the information which agents can access in principle, with unlimited resources), but the results and conclusions which agents can in practice reach with the inferential resources they have [34.41, §5.5]. By contrast, traditional analyses of scientific models adopt an in-principle stance: the question of their exploration and of the tractability of the methods used to explore them is one question among others, and is implicitly idealized away when discussing other issues. This implies surreptitiously smuggling in the unjustified claim that the distinction between what is possible in principle and what is possible in practice can be ignored for the investigation of these other issues, which may sometimes be controversial.

At the same time, philosophers of science draw their examples from the scientific literature, which, by

definition, presents successful investigations of models which must have been found to be, one way or another, tractable enough regarding the inquiries pursued. In brief, discussions about the scientific models which are found in scientific practices are ipso facto concerning computationally tractable models, or models having computationally tractable versions.

How much these remarks imply that existing analyses about scientific models have been discretely skewed, or on the contrary that the constraints of tractability have already been taken into account, needs to be ascertained, and the answer may be different depending on the question investigated. For example, for decades the question of the relations between fields has mainly been treated in terms of relations between theories. While this perspective is in part legitimate, recent investigations suggest that tractable models may also be a relevant unit to analyze scientific theoretical, methodological or practical transfers between fields [34.41, §3.3], [34.45, 46]. In any case, when discussing questions related to scientific representation, explanation, or confirmation, philosophers of science must watch out that answers may sometimes differ for the models that scientists work with daily (and which more and more require computers to be investigated), and for simple analytically solvable models, which philosophers more naturally focus upon, and which may have a specific scientific status regarding the construction of knowledge and the development of families of models in each field.

## 34.2 The Variety of Computer Simulations and Computational Models

Computer simulations involve the use of computers to represent and investigate the behavior of physical systems (Sect. 34.2.1). Beyond this simple characterization, various types of computer simulations can be met in science, each with its specificities, and, it is important to distinguish them to avoid undue extrapolations. Differences can be met at various levels of description. Computing machines can be digital or analog (Sect. 34.2.2). Digital computers are usually used to carry out numerical computations (Sect. 34.2.3), even if all computer simulations do not involve operations on numbers (Sect. 34.2.4). In both cases, computations may be deterministic or nondeterministic (Sect. 34.2.5). Finally, various types of mathematical and physical computational models can be met across science, such as equation-based models, agent-based models, coupled models or multiscale models, but, not all important computational methods or mathematical

frameworks are used to carry out computer simulations (Sect. 34.2.6).

My purpose in this section is to present and characterize different important types of simulations, which are used in scientific practice and will regularly be referred to in the following sections, and to highlight some specific epistemological questions related to them.

### 34.2.1 Working Characterization

In science, computer simulations are based on the use of computers. A computer is a physical apparatus which can reliably be used to carry out logical and mathematical operations. A computer simulation corresponds to the actual use of a computer to investigate a physical system *S*, by computationally generating the description of some of the states of one of the potential

trajectories of  $S$  in the state space of a computational model of  $S$  (*working characterization*).

This characterization is not meant as a full-blown definition (Sect. 34.6) but as a synthetic presentation of important features of computer simulations.

First, it emphasizes that the development of computers is a central step in the recent evolution of science, which was made possible by steady conceptual and technical progresses in the twentieth century. It can therefore be expected that computational aspects are often, though not necessarily always, central for the epistemological analysis of computational science and computer simulations (Sect. 34.3). Second, the working definition is meant to emphasize that all uses of computers in science cannot be seen as computer simulations. Typically, the use of computers to analyze big data is not considered as a computer simulation since the dynamics of the target system is not represented. Third, the characterization remains neutral regarding the question of whether in science there are simulations that are not based on the use of computers (whatever these could be). It is not incompatible with the claim that computer simulations are some sort of experimental activity, even if people willing to endorse such claims need to explain and justify in which sense the corresponding uses of computers can be considered as *experimental* (Sect. 34.5). Finally, since different types of computers exist, computer simulations may correspond to various types of objects. The working definition emphasizes that, in order to analyze actual science, the emphasis should be primarily on models of computations that can have an *actual* physical realization, and on physical systems that can be *used in practice* for scientific purposes – even if investigations about potential machines, and how some physical systems could instantiate them, may be relevant for foundational issues.

I now turn to the description of important types of computer simulations that have been, or still are, used in science and that figure in epistemological discussions about computer simulations.

### 34.2.2 Analog Simulations and Their Specificities

Analog computers were important tools for scientific computing till the late 1960s, during which with handbooks of analog computation were still being written [34.47, 48], and attempts were made in the early 1970s to link analog and digital computers. Analog simulations and physical analog systems are still occasionally used to investigate physical systems.

An analog computer is a physical machine which is able to carry out algebraic and integrodifferential

operations upon continuous physical signals. Thus, operations that would be difficult to program on a digital computer are immediately possible on an analog machine. The specificity of analog machines is that they contain physical elements whose dynamics decisively contribute to the dynamic instantiation of these mathematical operations. For a machine to be used as an analog computer, its physical dynamics must be explicitly known and completely under control so that there is no uncertainty about the operations which are carried out. While systems like wind tunnels cannot be made to compute several different dynamics, mechanical analog computers like the differential analyzer and electrical analog computers can be used as general-purpose computational tools.

Even if analog computers and analog simulations are seldom used nowadays, understanding them is epistemologically important. For instance, while quantum computation is an extension of classical digital computation, quantum analog computing, which involves no binary encoding, may prove useful for the purpose of the quantum simulation of physical systems [34.49]. Analog computers are considered to be potentially more powerful than digital machines *and* to be actually instantiated by physical systems, even if we are unable to use them to the full extent of their capacities because of analog noise or the impossibility of precisely extracting the information they process. The analysis of analog computability is also important for foundational studies aimed at determining which types of actual computers devices could be used for the purposes of computer simulations, how much resources we may need to simulate physical systems or what natural systems can compute [34.50–52]. For example, the General Purpose Analog Computer was introduced by Shannon as a model of the differential analyzer, which was used from the 1930s to 1960s.

Finally, understanding how analog computers work is important to understand analog simulations and how they differ from digital simulations. As is pitifully emphasized by *Arnold* [34.53, p. 52], the failure to distinguish properly between digital computer simulations and analog simulations can be (and has recently been) a major source of error in the philosophical discussions of computer simulations. Analog computers physically instantiate the mathematical dynamics which they are used to investigate. Therefore, the analog computational model that is analyzed is instantiated both by the physical computer and by the target system that is being simulated. Thus, the simulating and simulated processes share a common structure and are isomorphic [34.54], which need not be the case for digital simulations (Sect. 34.5.3). Importantly, this common structure is purely mathematical, and involves dimen-

sionless quantities [34.55, Chap. 8]. While the need to describe systems in terms of dimensionless quantities is a general one in the empirical sciences [34.56–58], and is also crucial for digital simulations, here it is specifically important to understand the type of reasoning involved in analog simulations. Indeed, the physical description of the simulating and simulated systems matter only in so far as one needs to justify that they instantiate a common dimensionless dynamical structure. In brief, such analogical reasoning does not involve any direct comparison between the physical material properties of the simulating and simulated systems: the mathematical structure mediates the comparison. In other words, even with analog simulations, an analysis of the similarities of the two systems is irrelevant once one knows which analog computation is being carried out by both systems.

### 34.2.3 Digital Machines, Numerical Physics, and Types of Equivalence

In digital machines, information is processed discretely, coded in binary digits (1 or 0), and stored in transistors. Computations involve the transition between computational states. These transitions are described in terms of logical rules between the states. If these rules can be described in a general form, they may be described in terms of equations involving variables. Digital computers can have various types of architecture with different computational performances. Traditionally, software was written for sequential computation, in which one instruction is executed at a time. In contrast, modern supercomputers are designed to solve tasks in parallel, and parallelism can be supported at different levels of architecture, which often implies the need to adapt algorithms, if not models, to parallel computation [34.59].

Digital machines can be used to develop different types of computer simulations. Much computational science is numerical: binary sequences code for numbers and computers carry out numerical operations on these numbers by processing the binary strings. Since computers can only devote limited memory to represent numbers (e.g., with floating-point representation), numerical science usually involves numerical approximations. In other words, computer simulations do not provide exact solutions to equations – even if the notion of an exact solution is not as straightforward as philosophers usually take it to be [34.60].

Different types of equivalence between computations, and, by extension, computer simulations, should be distinguished beyond equivalence at the bit level [34.61]. Logical and mathematical expressions and algorithms can be mathematically equivalent when they refer to, or compute, the same mathematical object

or some of its properties. Because of floating-point representation, round-off errors cannot be avoided in simulations. When algorithms result in small cumulative errors, they are stable and two such stable algorithms may be considered as numerically equivalent – although they need not be computationally equivalent in terms of their computational efficiency. Finally, based on the type of inquiry pursued, wider notions of representational equivalence can be defined at the computational model or computer simulation level. Typically, two computations yielding the same average quantity, or describing the same topology of a trajectory, may be considered as equivalent. Overall, this shows that analyses of the failures and predictive or explanatory successes of computer simulations must often be rooted in the technical details of computational practices [34.62]. From this point of view, an important part of computational science can be seen as the continuation of the numerical analysis tradition presented in Sect. 34.1.2.

### 34.2.4 Non-Numerical Digital Models

A large part of science gives a central role to scientific theories couched in terms of differential equations relating continuous functions with their derivatives. For this reason, much of computational science is based on finite-difference equations aimed at finding approximate numerical solutions to differential equations. However this theory- and equation-oriented picture does not exhaust actual practices in computational science. First, computer simulations can be carried out in the absence of theories – which turns out to be a problem when it comes to the explanatory value of computer simulations (Sect. 34.4). Second, even when equation-based theories exist, computational models are not necessarily completely determined by these theories and by mathematical results describing how to discretize equations appropriately (Sect. 34.3.2). Finally, even when well entrenched, equation-based, theories exist, digital, but non-numerical, computer simulations can be developed. This perspective was advocated in the 1980s by computer scientists like Fredkin, Toffoli, or Margolus. Building on the idea previously expressed by *Feynman*, that maybe “nature, at some extremely microscopic scale, operates exactly like discrete computer logic” [34.63], they wanted to develop “a less round-about way to make nature model itself” [34.64, p. 121] than the use of computers to obtain approximate numerical solutions of equations. The idea was to represent more directly physical processes by means of *physically minded* models, with interactions on a spatial lattice providing an emulation “of the spatial locality of physical law” [34.65] and to use exact models obeying discrete symbolic dynamics to dispense with numer-



ical approximations. In practice, this resulted in the renewed development of cellular automata (hereafter CA) studies and their use for empirical investigations. A CA involves cells in a specific geometry; each cell can be in one of a finite set of states, and evolves following a synchronous local update rule based on the states of the neighboring cells. The field of CA was pioneered in the 1940s by Ulam's works on lattice networks and von Neumann's works on self-replicating automata. It was shown over the decades that such models, though apparently over-simplistic, can not only be successfully used in fields as different as the social sciences [34.66] and artificial life [34.67], but also physics, in which they were shown in the late 1970s and 1980s to be mesoscopic alternate to Navier–Stokes macroscopic equations [34.68].

### 34.2.5 Nondeterministic Simulations

Another important distinction is between deterministic and nondeterministic algorithms. From the onset, computers were used to execute nondeterministic algorithms, which may behave differently for different runs.

Nondeterministic simulations involve using random number generators, which can be based on random signals produced by random physical processes, or on algorithms producing pseudorandom numbers with good randomness properties. Overall, the treatment of randomness in computer simulations is a tricky issue since generating truly random signals, with no spurious regularities which may spoil the results by introducing unwanted patterns, turns out to be difficult.

Monte Carlo methods, also called Monte Carlo experiments, are a widely used type of nondeterministic simulations. They were central to the Manhattan project, which led to the production of the first nuclear bombs and contributed heavily to the development of computer simulations. They can be used for various purposes such as the calculation of mathematical quantities like Pi or the assessment of average quantities in statistical physics by appropriately sampling some interval or region of a state space. These practices are hard to classify and, depending on the case, seem to correspond to computational methods, experiments, or full-blown computer simulations. *Metropolis* and *Ulam* [34.69] is a seminal work, *Galison* [34.70, 71] correspond to historical studies, and *Humphreys* [34.8], *Beisbart* and *Norton* [34.72]) to epistemological analyses.

### 34.2.6 Other Types of Computer Simulations

It is difficult to do justice to all the kinds of simulations that are seen in scientific practice. New com-

putational methods are regularly invented, and these often challenge previous attempts to provide rational typologies. Further, the features presented in the previous sections are often mixed in complex ways. For example, CA-based methods in fluid dynamics, which were not originally supposed to involve numbers or offer exact computations, were finally turned into lattice Boltzmann methods, which involve making local averages [34.73]. Here, I shall merely present types of computer simulations that are widely discussed in the philosophical literature.

#### Agent-Based Methods

Agent-based methods involve the microlevel description of agents and their local interactions (in contrast to global descriptions like balance or equilibrium equations), and provide tools to analyze the microscopic generation of phenomena. They are often opposed to equation-based approaches, but the distinction is not completely sharp, since equations do not need to describe global behaviors and, when discretized, often yield local update rules. Agent-based models and simulations are used across fields to analyze artificial, social, biological, etc., agents. CA models like the Schelling model of segregation can be seen as agent-based models even though most such agent-based also involve numbers in the descriptions of local interactions. Because they rely on microscopic descriptions, agent-based simulations are often at the core of debates about issues such as emergence [34.74], explanation [34.75], or methodological individualism in science [34.76].

#### Coupled and Multiscale Models

Extremely elaborate computational models, developed and studied by large numbers of scientists, are increasingly used to investigate complex systems such as Earth's atmosphere, be it for the purpose of precise predictions and weather forecasting or for the analysis of larger less precise trends of climate studies. While in fluid dynamics, it is sometimes possible to carry out *direct simulations*, where the whole range of spatial and temporal scales from the dissipation to the integral scale are represented [34.77, Chap. 9], such methods are too costly for atmosphere simulations, in which sub-grid models of turbulence or cloud formation need to be included (see *Edwards* [34.78] and *Heymann* [34.79] for accessible and clear introductions). Also, different models sometimes need to be coupled like in the case of global coupled ocean-atmosphere general circulation models.

These complex computer simulations raise a number of epistemological issues. First, in the case of multiscale or coupled models, the physical and computational compatibility of the different models can be

a tricky issue, and one must be careful that it does not create spurious behavior in the computer simulation (see *Winsberg* [34.16, 80], *Humphreys* [34.41, 81] for more analyses about such models). Second, since there are various ways of taking into account subgrid phenomena, pluralism in the modeling approaches cannot be avoided [34.82]. Importantly, the existence of different incompatible models need not be seen as a problem, and scientists can try to learn by comparing their results or elaborate ensemble methods to try to deal with uncertainties [34.83]. The development of investigations of such large-scale phenomena requires collective work, both within and between research teams. Typically, not only the interpretation of the models, their justification, the numerical codes [34.84], but also the standard of results [34.78, 85] must be shared by members of the corresponding communities. An important but still unexplored question is how much the collective dimension of these activities influences epistemologically how they are carried out. From this point of view, the epistemology and philosophy of computational models and computer simulations can be seen as another chapter of the analysis of the collective dimension of science.

## 34.3 Epistemology of Computational Models and Computer Simulations

Epistemologists analyze whether and how much knowledge claims are justified. In this case, it requires analyzing the specific roles played by computer simulations in the production and generation of items of knowledge (Sect. 34.3.1). Different levels of description and analysis can be relevant when investigating the epistemology of computer simulations, in addition to that of the computational model and how it is theoretically or experimentally justified (Sect. 34.3.2). Importantly, how computer simulations are justified, and why specific computational models are used by scientists, are overlapping (though not identical) questions. For example, field- or inquiry-specific explanations of the use of computer simulations fail to account for cross-disciplinary recurrences in the use of computational models, which may have more general mathematical or computational explanations (Sect. 34.3.3). Overall, computer simulations are one of the main sources of knowledge and data in contemporary science, even if the sense in which they produce new data and knowledge is often misunderstood (Sect. 34.3.4).

### Computational Methods versus Computer Simulations

Not all major families of mathematical and computational methods are used to produce computational models or computer simulations of empirical systems. Evolutionary algorithms are used for the investigation of artificial worlds, or of foundational issues about evolution, and they have important applications in the field of optimization methods. Artificial neural networks are used in the field of machine learning and data learning and to predict the behavior of physical systems out of large databases. Bayesian networks are helpful to model knowledge, develop reasoning methods, or to treat data. Overall, all these computational methods have clear practical applications. They can be used for scientific tasks, sometimes concurrently with computer simulations in the case of predictive tasks. However, no genuine representations of physical systems and their dynamics seem to be attached to their use – even if, as the development of CA-based simulations has shown, novel formal settings may eventually have unexpected applications for modeling purposes in the empirical sciences.

### 34.3.1 Computer Simulations and Their Scientific Roles

Science, as an institution, aims to reach epistemic goals of various sorts, both propositional (like reaching some epistemic states, typically justified true beliefs) and practical (like being able to reliably manipulate some physical systems). Epistemologically analyzing science requires the study of the reliability and efficiency of scientific procedures to reach these goals. Accordingly, to develop the epistemology of computer simulations, one first needs to single out their different scientific goals.

Even if they also serve pedagogical or expository purposes, most computer simulations can be described as activities aimed to develop knowledge. There exist various types of scientific knowledge (see *Humphreys* [34.81] for an overview), which raise specific problems, and, conversely, various types of knowledge can be produced by computer simulations.

Typically, items of knowledge may differ in how they are justified (theoretically, experimentally, inductively, etc.), and whether they were reached by a priori or a posteriori investigations. They may also differ regarding the activities needed to produce them and the type of information that they provide. For exam-

ple, predictive or explanatory knowledge, or knowledge about how systems behave and can be controlled are of different types. Some scientific roles can be general (*predicting*) and others are very specific. For example, computer simulations are used to develop evidential standards in physics by simulating detection procedures and identifying patterns of data (*signatures*) [34.86]. Overall, developing a coherent and fine-grained epistemology of computer simulations would require drawing a map of their various roles to see how much their epistemological features are general or contextual and role specific.

Let us now be more specific. In the twentieth century, the role of experiments, as sources of empirical evidence about nature and guides in the selection of theories, was repeatedly, if not exclusively, emphasized by empiricist philosophers of science. Conversely, activities which did not provide such evidence were mainly seen as serving theoretical purposes. Typically, models were first seen as being primarily of a theoretical nature [34.20, §5]. In this perspective, *Models as Mediators*, in 1999, represented a significant advance. Morgan and Morrison, by presenting a more precise “account of what [models] do” in science [34.20, p. 18], offered a more nuanced epistemological picture, where models were shown to have functions as diverse as investigating theories and the world, intervening, helping for measurement purposes, etc. Since an important role of computer simulations is to demonstrate the content of models [34.13] or unfold well-defined scenarios [34.87], computer simulations can be expected to have, or contribute to, similar roles to those described by Morgan and Morrison and to potentially share these roles with other demonstrative activities like argumentation or mental simulations.

Importantly, such a description of science, where items or activities as diverse as theories, models, computer simulations, thought experiments, or experiments may serve partly overlapping purposes, remains compatible with empiricism provided that experiments are seen as, in the architecture of knowledge, the only ultimate source of primary evidence about the nature of physical systems. It is also compatible with the claim that secondary, derived sources of knowledge, like theories, models, or simulations, can sometimes be more reliable than experiments to provide information about how systems behave, in particular in cases in which experimental evidence is hard to come by (Sect. 34.5.4).

Overall, it is unlikely that there is such a thing as *the* epistemology of computational models and simulations. If the various roles of computer simulations are specific cases of general functions, like demonstrating or unfolding, there may be such a thing as a general, but incomplete, epistemology of computer

simulations, corresponding to the general epistemological problems raised by such general functions. In any case, to complete the picture, one needs to go deeper into the analysis of the roles that computer simulations serve within scientific practices and how they fulfill these roles in various types of contexts. This program is not incompatible with the philosophical perspectives of some of the advocates of the so-called *practice turn* in science [34.88], and in particular of authors who put contextually described scientific activities at the core of their description of science [34.89, 90].

### 34.3.2 Aspects of the Epistemological Analysis of Computer Simulations

#### A Multilayered Epistemology

Epistemology analyzes the types of justifications that we have for entertaining knowledge claims, and investigates how and why we epistemically fail or succeed. In the case of computer simulations, failure may take place at various levels, from the material implementation of the computation to the physical model that is at the core of the inquiry, and at all the intermediate semantic levels of interpretation that are needed to use computers for the investigation of models (see *Barberousse* et al. [34.91] for a general description and *Grim* et al. [34.92] for a discussion of some specific failures found in computer simulations). Overall, the epistemology of computer simulations involves discussing the reasons that we have for claiming:

1. The computers that we use work correctly.
2. The programs or algorithms do what we want them to do.
3. The computer simulations, *qua physical representations*, correctly depict what we want them to represent.

Steps 1 and 2 correspond to questions related to engineering and computer science. I shall not discuss these at length but will simply illustrate them to show how serious they are in this context. For example, at the architectural level, parallel computing requires coordination the different cores of computers so that all potential write-conflicts in the memory are solved. At the program level, when trying to solve a problem P with an unknown solution S, scientists need to prove the correctness of the algorithms they use and to verify that the programs written do indeed execute these algorithms. Many such verification problems are undecidable, which means that no general versatile procedure can be found to make this verification for all cases. However, this does not imply that proofs of the correctness of the algorithm cannot sometimes be pro-

vided for specific problems. Overall, scientists in this field still actively investigate and debate how much algorithms can be verified (see *Fetzer* [34.93], *Asperti et al.* [34.94] and *Oberkampff and Roy* [34.95] for discussions). At a higher mathematical level, as we saw earlier, many computational methods provide numerical methods for approximately solving problems, and the stability of algorithms can be a source of concern, which means that analyzing computational errors is part of the epistemology of simulations [34.62].

Finally, one needs to assess whether the approximations in the solution, as well as the representational inadequacies of the model, are acceptable regarding the physical inquiry pursued. At this interpretational level, because of the variety of theoretical contexts in which computer simulations are carried out, there is no single and general way in which the reliability of the results they provide can be analyzed. The credentials of computer simulations will be different depending on whether a sound theory is being used, how much uncertainty there is about the initial conditions, how complex the target system is, whether drastic simplification assumptions have been made etc. Also, depending on what the simulation is used for, and what type of knowledge it is meant to provide, the justificatory requirements will be more or less stringent. It takes different arguments to justify that based on a simulation one knows how to represent, control, predict, explain, or understand the behavior of the system (see Sect. 34.4 for a discussion of the last two cases, and [34.96] for similar analyses). Similarly, precise quantitative spatial-temporal predictions are in need of much pointed justifications than computer simulations aimed at studying average quantities or qualitative behaviors of systems. Importantly, this discussion of the reliability of computer simulations overlaps significantly with that of the epistemology of physical models, and with how the results issued from approximate, idealized, coarse grained, or simply partly false models can still be scientifically valuable (see *Portides* Chap. 2, this volume; *Frigg and Nguyen* Chap. 3, this volume). However, in the present context, it is important to emphasize that, even if the content of models obviously constrains the reliability of the information that can be extracted from them, models do not by themselves produce results – only procedures which investigate them do. In this perspective, the epistemology of computer simulations is a reminder that reliability primarily characterizes practices or activities that produce knowledge and that models, taken alone, are not such practices. In other words, epistemological discussions about the reliability of models as knowledge providers make sense only by explicitly reintroducing such practices or when it can be assumed that reliably

extracting all their content is possible, an assumption that, in the framework of computational science, is often not plausible.

#### From Theoretical to Empirical Justifications

Computer simulations have often been viewed as ways of exploring theories by hypothetico-deductive methods. This characterization captures a part of the truth, since existing theories are often a starting point for the construction of computer simulations. In simple cases, computer simulations can mainly be determined by theories, like in the case of *direct simulations* [34.77] in fluid dynamics, which derive from Navier–Stokes equations, and in which all relevant scales are simulated and no turbulence model is involved.

However, taken as a general description, this view misrepresents how computer simulations are often produced and their validity justified. As emphasized by *Lenhard* [34.97], even when theoretical equations are in the background, computer simulations often result from some cooperation between theory and experiments. For example, in 1955 when Norman Phillips managed to reproduce atmospheric wind and pressure relations with a six-equation model, which arrangement of equations could lead to an adequate model of the global atmosphere was uncertain and the need for experimental validation was primordial to confirm his speculative modeling assumptions. Overall, the role of empirical inputs in simulation studies is usually crucial to develop phenomenological modules of models, parameterize simulations, or investigate their reliability based on their empirical successes [34.15, 17].

At the same time, since computer simulations are used precisely in cases where empirical data are absent, sparse, or unreliable [34.16], sufficient data to build up and empirically validate a computational model may be missing. In brief, in some cases, computer simulations can be sufficiently constrained neither by theories nor by data and are somewhat autonomous. From an epistemological point of view, this potential situation of theoretical and experimental under-determination is not something to be hailed, since it undermines the scientific value of their results (see also Sect. 34.4.2).

#### The Epistemology of Complex Systems

Because computer simulations are generally used to analyze complex systems, their epistemology partly overlaps that of complex systems and their modeling. It involves the analysis of simplification procedures at the representational or demonstration levels and of how various theoretical or experimental justifications are often used concomitantly. Overall, when it comes to investigating complex systems, obtaining reliable knowledge is difficult. Thus, any trick or procedure that

works is welcome and the result is often what Winsberg has labeled a *motley* epistemology [34.16].

At the same time, sweeping generalizations should be avoided. Philosophers studying computer simulations have too often cashed in their epistemology in terms of that of the most complex cases, such as computer simulations in climate science, which are characterized by extreme uncertainties and the complexity of their dynamics. But computer simulations are used to investigate systems that have various types and degrees of complexity, and whose investigation meets different sorts of difficulties. It is completely legitimate, and politically important, that philosophers epistemologically analyze computational models and computer simulations in climate science (see, e.g., [34.98] for an early influential article). However, to obtain a finer grained and more disentangled picture of the epistemology of computer simulations, and not to put everything in the same boat, a more analytic methodology should be applied. More specifically, one should first analyze how the results are justified in more simple cases of computer simulations where specific scientific difficulties are met independently. In a second step, it can be analyzed how adding up scientific difficulties changes justificatory strategies and when exactly more holistic epistemological analyses are appropriate [34.99]. In this perspective, much remains to be done.

### Epistemic Opacity

Epistemic opacity is present in computer simulations to various degrees and has various origins.

Models are often said to be white, gray, or black boxes depending on how they represent their target system. White-box models describe the fundamental laws or causal mechanisms of systems whereas black-box models simply correctly connect different aspects of their behavior. This distinction partly overlaps with that of theoretical and phenomenological models (see *Barberousse* and *Imbert* [34.100, §3.2] for sharper distinctions about these last notions). Computer simulations can be based on all types of such models, which may affect the understanding that they yield [34.101] (see also Sect. 34.4).

Opacity can also be present at the computational model or computational process level. *Global epistemic opacity* may arise from the complexity of the computation when it is not possible for an agent to inspect and verify all steps of the process [34.41, §3.7], [34.102]. It is in part contingent since it is rooted in our limitations as epistemic creatures, but it may be in part intrinsic in the sense that the complexity of the computation may be irreducible (see Sect. 34.4.3). Importantly, it is compatible with *local epistemic transparency*, when *any* step of the process can be inspected by a human mind – which

may prove useful in cases in which problems can be located by testing parts of the process and applying a dichotomy procedure. Local transparency requires that all details of the physical models and computational algorithms used be transparent, which may be more or less the case. Usually, computer simulations make heavy use of mathematical resource libraries such as code lines, routines, functions, algorithms, etc. In applied science, more or less opaque computational software can be proposed to simulate various types of systems, for example, in computational fluid dynamics [34.91, p. 567]. This raises epistemological problems since black-box software is built on physical models with limited domains of physical validity, and results will usually be returned even when users unknowingly apply such software outside these domains of validity.

Another form of epistemic opacity for individual scientists arises from the fact that investigating natural systems by computer simulations may require different types of experts, both from the empirical and mathematical sciences. As a result, no single scientist has a thorough understanding of all the details of the computational model and computational dynamics. Such type of opacity is not specific to computer simulations, since it is a consequence of the epistemic dependence between scientists within collaborations [34.103].

### 34.3.3 Selecting Computational Models and Practices

How do individual scientists decide to pursue specific theories, and, in particular, what types of sociological, psychological, or epistemic factors play a role in such processes? Conversely, do selected theories share specific features or properties? *Mutatis mutandis*, similar questions can be asked about other elements of science, such as research programs, experiments, models, practices, and, in the present case, computational models and computer simulations. Philosophers have mainly analyzed these questions by focusing on the explicit scientific justifications of individual practices, and the content of the representations involved. As we shall see, this is only a part of the story.

#### Explanation of Uses versus Justification of Uses

A helpful distinction is that between the explanation (and context) of use of a practice and its scientific justification within a scientific inquiry aimed at reaching specific purposes. To use words close to *Reichenbach's* [34.104, pp. 36–37], while the latter deals with the objective relation that scientists consciously try to establish between these given <activities> (simula-

tions, experiments, etc.) and the conclusions that are obtained from them, other aspects of material, computational, cognitive, or social natures, potentially unknown to the scientific agents involved in the inquiry, may play a role to explain that these activities were actually carried out. For example, in the case of the Millennium Run (a costly simulation in astrophysics), the results were made publicly accessible. Scientists who were not involved in the process leading to the decision to carry out this simulation could try to make the best of it since it was already there and milk it as much as possible for different purposes. Or, some scientists may decide to study biological entities like proteins or membranes by means of Monte Carlo simulations, because members of their teams happen to be familiar with these tools. However, once they have decided to do so, they must still justify how their computer simulations support their conclusions.

In the perspective of explaining actual scientific uses, one also needs to distinguish between explanations aimed to account for specific uses (e.g., *Why was the millennium simulation carried out in 2005 by the Virgo consortium?*) and those aimed to explain more general patterns, corresponding to the use of practices of a given type, within or across several fields of science (e.g., *Why are Monte Carlo simulations regularly used in this area of physics?*, *Why are they used regularly in science?*). Importantly, since different instantiations of a pattern may have different explanations, the aggregated frequency of a scientific practice, like that of the use of the Ising model across science, may be the combined effect of general transversal factors and of inquiry- or field-specific features [34.105].

#### Field-Specific versus Cross-Disciplinary Explanations

A tempting move has often been to answer that scientific choices are primarily, if not completely, theory driven – and are therefore field specific. After all, theories guide scientists in their predictive and explanatory activities by fueling the content of their representations of natural systems. However, a reason to look for additional elements of explanations is that the spectrum of actual modeling and computational practices is smaller than our scientific knowledge and goals would allow [34.21, 41, 106]. For example, why do the harmonic oscillator, the Ising model, the Lotka–Volterra model, Monte Carlo simulations, etc., play such prominent roles throughout science?

As highlighted by *Barberousse* and *Imbert* [34.105], a variety of significantly different explanations of the greater or lesser use of models of a given type, and of scientific practices, can be found, beyond the straightforward suggestion that there are

regularities in nature, which are mirrored by modeling and computational practices.

#### Local Factors

The explanation may be rooted in the specificities of modeling and computational activities. In particular, if explaining is better achieved by limiting the number of (types) of (computational) models [34.21, pp. 144–5], or explanatory argument patterns [34.107], it is no surprise that often the same computational models and practices are met. Also, scientists may feel the need to avoid dispersion of their efforts in cases when research programs need to be pursued for a long time before good results can be reached and it is more profitable to exploit a local mine than to go digging somewhere else [34.106, Chaps. 4 and 5], [34.21, pp. 143–4]. More generally, the recurrence of computational practices may be viewed as another example of the benefits of adopting scientific standards [34.108]. One may also, in the Kuhnian tradition, put the emphasis on the education of scientists, who are taught to see new problems through the lens of specific problems or exemplars [34.106, p. 189], and emphasize that this education has a massive influence on the lineages of models or practices which are later developed. This story can have a more or less contingentist version, depending on why the original models or practices at the lineage seeds are adopted in the first place, and why these uses are perpetuated and scientists do not emancipate from them after schooling.

Theories may also play an indirect role in the selection of computational models. For example, models naturally couched in the standard formalism of a theory may be easier to use, even if the same physics can also be put to work by using other models. *Barberousse* and *Imbert* [34.100] analyze the case of lattice methods for fluid simulations in depth, which, though significantly different from approaches based on Navier–Stokes differential equations, can be used for the same purposes, even if this requires spending time learning and harnessing new methods and formalisms, which physicists may be reluctant to do.

#### Computational and Mathematical Explanations

As seen in Sect. 34.1.4, *Humphreys* [34.41, 81], suggests that most scientific models are tailored to fit the available mathematics, hence the importance in scientific practice of tractable models (see *Humphreys*'s notion of computational template [34.41, §3.4], and further analyses in [34.45]). Even if one grants the potential importance of such mathematical and computational factors, cashing out in detail the corresponding explanation is not straightforward. *Barberousse* and

*Imbert* [34.105] emphasize that there are various computational explanations. The *objective computational landscape* (how intrinsically difficult problems are, how frequent easy problems are) probably influences how science develops, even if knowing exactly what it looks like and how it constrains scientific activity is of the utmost difficulty. However, the *epistemic computational landscape* (scientists' beliefs about the objective computational landscape) may just be as important since it frames modeling choices made by scientists.

Other potentially influential factors may also include how difficult it is to explore the objective landscape (and the corresponding beliefs regarding the easiness of this exploration), how much scientists, who try to avoid failure, are prone to resort to tractable models, or which techniques are used to select such tractable models (since some specific techniques, like polynomial approximations, may repeatedly select the same models within the pool of tractable models). Finally, modeling conservativeness may also stem from the computational and result pressure experienced by scientists, that is, how scarce computational resources are in their scientific environment and how much scientists need to publish results regularly.

#### Universality, Minimality, and Multiple Realizability

Other explanations may be offered in terms of how weak the hypotheses are to satisfy a model or a distribution. For example, the Poisson distribution is often met because various types of stochastic processes satisfy the few conditions that are required to derive it [34.41, pp. 88–89]. Relationships between models and how models approximate to each other may also be important. Typically, the Gaussian distribution is the limit of various other distributions (see however, *Lyon* [34.109] for a more refined analysis and the claim that in Nature Gaussian distributions are common, but not pervasive). More generally, models that capture universal features of physical systems and are rooted in basic properties, such as their topology, can be expected to be met more often. Therefore, for reasons having to do with the mathematics of coarse-grain descriptions, and the explanation of multiple realizability, many systems fall into the same class and have similar descriptions [34.110–112] when minimal, macro-level, or simply qualitative models are built and explored.

Importantly, all the above explanations are not exclusive. Typically, the emphasis on tractability may be a general one in the sense that models always need to be tractable if they are to be used by scientists.

### 34.3.4 The Production of 'New' Knowledge: In What Sense?

#### Be Careful of Red Herrings!

It is commonly agreed that computer simulations produce *new knowledge*, *new data*, *new results*, or *new information* about physical systems (*Humphreys* [34.41], *Winsberg* [34.113, pp. 578–579], *Norton* and *Suppe* [34.114, p. 88], *Barberousse* et al. [34.91, p. 557], *Beisbart* [34.115]). This can be considered as a factual statement, since contemporary science, which is considered to produce knowledge, relies more and more heavily on computer simulations.

At the same time, the notion of knowledge should not be a red herring. It is commonly considered that experiments, inferences, thought experiments, representations, or models can bring knowledge, which then generates the puzzle that widely different activities have similar powers. The puzzle may be seen as somewhat artificial since knowledge, especially scientific, can be of different types [34.81], and when *new* knowledge is produced, the novelty can also be of different types. In this perspective, it may be that what is produced by each of these activities falls under a general identical concept but is significantly different. From this point of view, the real question concerning computer simulations is not whether they produce *knowledge*, but in which particular sense they produce knowledge, what kind of knowledge they produce, what is specific to the knowledge produced by computer simulations, and what type of novelty is involved.

A comparison can be made with thought experiments, for which the question of how they can produce new knowledge has also been debated. Both activities correspond to the exploration of virtual situations, and do not involve direct interactions with the systems investigated. From this point of view, computer simulations and thought experiments can be seen as platonic investigation of ideas, with this difference that, for computer simulations, the mind is assisted by computers [34.41, p. 115–116]. Overall, computer simulations have been claimed to sometimes play the same role of unfolding as thought experiments [34.87], have sometimes been equated with some types of thought experiments [34.116], and it has been suggested that computational modeling might bring the end of thought experiments [34.117]. Importantly, even if thought experiments are perhaps less used in science than formerly, this latter claim seems implausible. The reason is that there are different kinds of thought experiments, and many reveal conceptual possibilities that have little to do with computational explorations. Arguably, the possibility to set up computer simulations would have added nothing to famous thought experiments such as

those made by Galileo, Einstein, Podolsky and Rosen, or Schrödinger. (I am grateful to Paul Humphreys for emphasizing this point.) In any case, a satisfactory account of these activities should account for both similarities and differences in how they work epistemologically and how they are used.

In any case, the question of how and what we can learn about reality by using these methods arises, even if the sources of puzzlement do not exactly touch the same points in each case. Indeed, how mental thought experiments work is more opaque than how computer simulations do. For this reason, their rational reconstruction as logical arguments [34.118, p. 354] is more controversial than that of computer simulations [34.115], and it is less clear whether their positive or negative epistemic credentials are those of the corresponding reconstructed argument [34.119]. (For example, if certain thought experiments are reliable because mental reasoning capacities about physical situations have been molded by evolution, development, or daily experiments, it is not clear that their logical reconstruction will more vividly make clear why they are reliable.) The situation is clearer for computer simulations since the process is externalized and is based on more transparent mechanisms (see however Sect. 34.3.2). Then, if computer simulations are nothing else than (computationally assisted) thinking corresponding to the application of formal rules, and their output is somewhat contained in the description of the computational model, how knowledge is generated is clearer but the charge of the lack of novelty is heavier.

#### The Need for an Adequate Notion of Content

Suppose that a physical system  $S$  is in a state  $s$  at time  $t$  and obeys deterministic dynamics  $D$ . Then, the description of  $D$  and  $s$  characterizes a mathematical structure  $M$ , which is the trajectory of  $S$  in its phase space and is known as such. If a computer simulation unfolds this trajectory, then it explicitly shows which states  $S$  will be in. At the same time, any joint description of one of these states and of the dynamics denotes the same structure  $M$ , which is known to characterize the evolution of  $S$ . So, from a logical point of view, no new content has been unraveled by the computer simulation, which can at best be seen as a means of producing new descriptions of identical contents. In brief, if knowledge is equated with that of logical content, computer simulations do not seem to be necessarily producing new knowledge. We may even be tempted to describe computer simulations as somewhat infertile and thereby perpetuate a tradition according to which formal or mechanical procedures to draw inferences, and rules of logic in particular, are sterile, as far as discovery is concerned, and can at best be used to present pieces of

knowledge that have already been found – a position defended by *Descartes* in 1637 in the *Discours de la Méthode* [34.120]. This kind of puzzle, though particularly acute for computer simulations, is not specific to them and is nothing new for philosophers of language – Frege and Russell already analyzed similar ones. However, this shows that, *pace* the neglect for linguistic issues in the present philosophy of science, without an adequate theory of reference and notion of content that would make clear what exactly we know and do not know when we make a scientific statement, we are ill-equipped to precisely analyze the knowledge generated by computer simulations [34.41, 121].

Computational science may also remain somewhat mysterious if one reasons with the idealizations usually made by philosophers of science. As emphasized in Sect. 34.1.4, idealizing away the practical constraints faced by users is characteristic of much traditional philosophy of science and theories of rationality. In the present case, it is true that “*in principle*, there is nothing in a simulation that could not be worked out without computers” [34.122, p. 368]. Nevertheless, adopting this *in principle* position is unlikely to be fruitful here since, when it comes to actual computational science, which scientific content can be reached *in practice* is a crucial issue if one wants to understand how computational knowledge develops and pushes back the boundaries of science (see *Humphreys* [34.41, p. 154] and *Imbert* [34.102, §6]).

Overall, it is clear that present computational procedures and computer simulations do contribute to the development of scientific knowledge. Thus, it is incumbent on epistemologists and philosophers of sciences to develop conceptual frameworks to understand how and in what sense computer simulations extend our science and what type of novelty is involved.

#### Computer Simulations and Conceptual Emergence

Computer simulations unfold the content of computational models. How to characterize the novelty of the knowledge that they bring us? Since the notion of novelty is also involved in discussions about emergence, the literature about this latter notion can be profitably put to work here.

Just as emergence may concern property instances and not types [34.123, 124, p. 589], the notion of novelty needed here should apply to tokens of properties instantiated in distinctive systems and circumstances, or to specific regularities the scope of which covers such tokens and circumstances. For example, the apparition of vortices in fluids is in a sense nothing new, since the behavior of fluids is covered by existing theories in fluid dynamics, no new concept is involved, and other phe-



nomena of this type are already understood for some well-studied configurations. At the same time, finding out that patterns of vortices emerge in configurations of a new type is a scientific achievement and the discovery of some new piece of knowledge.

Importantly, as emphasized by *Barberousse* and *Vorms* [34.125, p. 41], the notion of novelty should be separated from that of surprise. When the exact value of a variable is precisely learnt and lies within the range that is enabled by some physical hypothesis or principle, we have a kind of *unsurprising novelty*. *Barberousse* and *Vorms* give an example from experimental science, but computer simulations may also provide exact values for quantities, which agree with general laws (e.g., laws of thermodynamics) and are therefore partly expected.

In addition, computer simulations can provide cases of *surprising novelty*, concerning behaviors that are covered by existing theories like chaotic behavior for classical mechanics. Indeed, Lorenz attractor and behaviors of a similar type were discovered by means of computer simulations of a simplified mathematical model initially designed to analyze atmospheric convection, and this stimulated the development of chaos theory [34.125, p. 42].

This leads us to a type of novelty, related to what *Humphreys* calls conceptual emergence. Something is conceptually emergent relative to a theoretical framework *F* when it requires a conceptual apparatus that is not in *F* to be effectively represented [34.41, p. 131], [34.123, p. 585]. The conceptual apparatus may require new predicates, new laws and sometimes the introduction of a new theory. Importantly, conceptual emergence is not merely an epistemic notion. It does not depend on the concepts we already possess and the conceptual irreducibility is between two conceptual frameworks. Further, even if instances of the target pattern can be described at the microlevel without the conceptually emergent concepts, the description of the pattern itself, if it is made without these novel concepts, is bound to be a massive disjunction of microproperties, which misrepresents the

macro-pattern qua pattern. Also, the same conceptually emergent phenomena may arise in different situations and its description may therefore require an independent conceptual framework, just like the regularities of special science require new concepts, unless one is prepared to describe their content in terms of a massive disjunction of all the cases they cover [34.126].

Interestingly, various phenomena investigated by computational science are conceptually emergent. Even if computer simulations are sufficient to generate them, identifying, presenting, and understanding them may require further analyses of the simulated data, re-descriptions at higher scales and the development of new theoretical tools. For example, traffic stop-and-go patterns in CA models of car traffic, emergent phenomena in agent-based simulations, and much of the knowledge acquired in classical fluid dynamics seem to correspond to the identification and analysis of conceptually emergent phenomena. Effectively, it is by conceptually representing these phenomena in different frameworks that one manages to gain novel information about these systems, above and beyond our blind knowledge of the microdynamics that generates them.

It is important to emphasize that different types of novelty described above are also met in experiments exploring the behavior of systems for which the fundamental physics is known. In other words, the potential novelty of experimental results should not be overemphasized. Even if only experiments can *confound* us [34.127, pp. 220–221] through results which are not covered by our theories or models, many of the new empirical data that these experiments provide us with are no more novel than those produced by computer simulations. The statements describing these results are not *strongly referential*, in the sense that no unknown aspects of the deep nature of the corresponding systems would be unveiled by a radically new act of reference [34.87, pp. 3463–3464]. These statements derive from what we already know about the physical systems investigated, and the experimental systems unravel them for us. In this sense, they are merely *weakly referential*.

## 34.4 Computer Simulations, Explanation, and Understanding

Can scientists provide explanations by simulating phenomena? If the answer is based on the explanatory requirements corresponding to the existing accounts of explanation, it is hard to see why some computer simulations could not be explanatory (Sect. 34.4.1). Why the specificities of computer simulations should necessarily deprive them for their explanatory potential is also unclear (Sect. 34.4.2), which is compatible

with the claim that computer simulations are used for inquiries whose results are, on average, less explanatory (Sect. 34.4.3). Be this as it may, because they heavily rely on informational and computational resources, computer simulations challenge our intuitions about explanatoriness, and in particular the expectation that good explanations should enable scientists to enjoy first-person objective understanding of the sys-

tems they investigate (Sect. 34.4.4). Even if computer simulations fail to meet these expectations because of their epistemic opacity, understanding may sometimes be regained by appropriately visualizing the results or studying phenomena at a coarser level (Sect. 34.4.5). In any case, scientific judgments about such issues are influenced by disciplinary norms, which may sometimes evolve with the development of computational science (Sect. 34.4.6).

### 34.4.1 Traditional Accounts of Explanation

Philosophers of science have discussed intensively the issue of scientific explanation over the last decades. The seminal works of Hempel were published in the 1940s, when computational science started to develop. However, until recently, discussions about computer simulations and explanations did not interfere with each other – which could suggest that for theorists of explanation, *how* explanations are produced does not in fact matter. While it is true that many of the examples of explanatory inquiries analyzed in the literature are simple and, at least in their most elementary versions, do not belong to computational science, it is hard to see why computer simulations could not in some cases satisfy the requirements corresponding to major accounts of explanations. According to the deductive-nomological (hereafter DN) model, one explains a phenomenon when a sentence describing it is logically deduced from true premises essentially containing a scientific law [34.128, pp. 247–248]. For example, the explanation of the trajectory of a comet, by means of a computer simulation of its trajectory based on the laws of classical (or relativistic) mechanics together with the initial positions of all bodies significantly influencing its trajectory, seems to qualify as a perfect example of DN explanation – provided that computer simulations can be seen as deductions [34.91, 115].

Analog statements can be made concerning the causal and unificationist models of explanations. The computer simulation of the comet’s trajectory is a way to trace the corresponding causal processes, described in terms of mark transmission [34.129] or of conserved quantities such as energy and momentum [34.130]. Other causal theorists of explanation like Railton have claimed that explanatory progress is made by detailing the various causal mechanisms of the world and all the nomological information relevant to the investigated phenomenon; the corresponding “ideal explanatory text” is thereby slowly unveiled [34.131]. But, one should note that, because such ideal explanatory texts are necessarily complex, their investigation is almost inevitably made by computational means.

Similarly, computer simulations can sometimes be instantiations of argument patterns that are part of what *Kitcher* describes as the explanatory store unifying our beliefs [34.107]. For example, the computation of the comet’s trajectory can be seen as an instantiation of “the Newtonian schema for giving explanations of motions in terms of underlying forces” [34.132, p. 121, p. 179].

Be this as it may, computer simulations have often been claimed, both by scientists and philosophers, to be somewhat problematic concerning explanatoriness and lacking some of the features that are expected to go with the fulfillment of explanatory requirements. This reproach of unexplanatoriness can be understood in several senses.

### 34.4.2 Computer Simulations: Intrinsically Unexplanatory?

One may first claim that computer simulations in general, or some specific types of them, do not meet one’s favorite explanatory requirements. For example, agent-based simulations may be described as not usually involving covering laws nor providing explanatory causal mechanisms or histories [34.75, 133]. However, one should not ascribe to computer simulations reproaches that should be made to the field itself. If a field does not offer well-entrenched causal laws and one is convinced that explanations should be based on such laws, then the computer simulations made in such fields are not explanatory, but this has nothing to do with computer simulations in general. Also, some computer simulations are built with scientific material like phenomenological regularities, which potentially makes them unexplanatory, but this material could also be used in the context of explanatory inquiries involving arguments or closed form solutions to models. Thus, the problem comes from the use of this material and not from the reliance on one or another mode of demonstration – and claiming that computer simulations are unexplanatory is like blaming the hammer for the hardness of the rock.

For this reproach to be meaningful (and specific to computer simulations), it should be the case that other inquiries based on the same material are indeed explanatory, but that the corresponding explanations based on computer simulation are not, because of specific features of computer simulations or some types of them. It is not completely clear how this can be so. Computer simulations are simply means of exploring scientific models and hypotheses by implementing algorithms, which provide information about tractable versions of these models or hypotheses. Therefore, their explanatory peculiarity, if any, should be an effect of

specific features like the use of algorithms, coding languages, or external computational processes.

There is no denying that the need to format scientific models and hypotheses into representations that are suitable for computational treatment comes with constraints. For example, a recent challenge has been to adapt coupled circulation models and their algorithms to the architecture of modern massively parallel supercomputers. Similarly, when one uses CA models for fluid dynamics, the physical hypotheses must be expressed in the straightjacket of up-to-date rules between neighboring cells on a grid. Beyond these genuine constraints on computational practices, one should remember that, computational languages, provided they are rich enough, are content neutral in the sense that any content that can be expressed with some language can also be expressed with them. Similarly, computational devices like the computers we use daily are universal machines in the sense that any solution to a computational problem (or inference) that can be produced by other machines can also be produced by them. For these reasons, it is hard to see why, *in principle*, computer simulations should be explanatorily limited, since the theoretical content and inferences related to other means of inquiries can also be processed by them.

The case of CA models abovementioned exemplifies nicely this point. For several decades, CA models have been used under various names in various fields; from Schelling's investigations about spatial segregation in neighborhoods, analysis of shock waves in models of car traffic, models of galaxies, investigations of the Ising model, to fluid dynamics (see *Ilachinski* [34.134] for a survey). Because existing theories and scientific laws are not expressed in terms of CA, some philosophers have claimed that CA-based simulations were merely phenomenological [34.135, pp. 208–209], [34.9, p. 516]. Nevertheless, *Barberousse* and *Imbert* [34.100] have argued that such bold *general* statements do not resist close scrutiny. They present the case of lattice gas models of fluids and argue that, beyond their unusual logical nature, from a physical point of view, such mesoscopic models and computer simulations make use of the same underlying physics of conserved quantities as more classical models, and can be seen as no less theoretical than concurrent computer simulations of fluids based on macroscopic Navier–Stokes equations. Therefore, there is no reason why such computer simulations could not be usable for similar explanatory purposes.

Overall, there is no denying that *some* (and possibly many) computer simulations are not explanatory. Providing various examples of unexplanatory computer simulations is scientifically valuable, but it says nothing general about their general lack of explanatory power,

unless one shows why unexplanatoriness stems from specific features of (some types of) computer simulations qua simulations. In the absence of such conceptual analyses, one can simply conclude that some scientific uses of computer simulations, or some computational practices, turn out to be unexplanatory.

### 34.4.3 Computer Simulations: More Frequently Unexplanatory?

A different claim is that, given the current uses of computer simulations in science, they are more often unexplanatory than other scientific items or activities, even if this is partly a contingent matter of fact. The explanatoriness of computer simulations can be threatened in various ways. Computational models may be built on false descriptions of target systems or may lack theoretical support and simply encapsulate phenomenological regularities; they may have been spoiled by the approximations, idealizations, and modeling tricks used to simplify models and make them tractable; they may depart from the well-entrenched explanatory norms in a field or may not correspond to accepted explanatory methods. Clearly, none of these features is specific to computer simulations. However, it may be the case that because of their current uses in science, computer simulations more frequently instantiate them.

#### The Lure of Computational Explorations

Because they are powerful heuristic tools, and because other means of exploration are often not available, computer simulations are more often used to toy and tinker with hypotheses, models, or mechanisms and, more generally, to *experiment on theories* [34.135, 136]. This may especially be the case in fields where there is no well-established theory to justify (or invalidate) the construction of models, or where collected evidence is not sufficient to check that the simulated mechanisms correspond to actual mechanisms in target systems. For example, in cognitive science, competing theories of the mind and its architecture coexisted for decades, and even modern techniques of imaging like fMRI (functional magnetic resonance imaging), though empirically informative, do not provide sufficient evidence to determine how the brain works precisely in terms of causal mechanisms. Accordingly, in this field, developing a model that is able to simulate the cognitive performances of an agent does not imply that one has understood and explained how her brain works, and more refined strategies that constrain the functional architecture must be developed if one wants to make explanatory claims [34.4, Chap. 5]. The issue is all the more complex in this specific field since the inquiry may also involve determining (verses assuming)

whether neural processes are computations [34.137]. Similarly, in the social sciences, empirically validating a simulation is far from being straightforward and as a result the epistemic and, in particular, explanatory value of computer simulations is often questionable [34.138].

Overall, since computer simulations offer powerful tools to investigate hypotheses and match phenomena, it is a temptation for scientists to take a step further and claim that their computer simulations have explanatory value. In brief, computer simulations offer a somewhat natural environment for such undue explanatory claims.

### The Worries of Under-Determination

In the case of computer simulations, the higher frequency of inappropriate explanatory claims may be reinforced by the combination of several factors.

When toying with hypotheses, scientists are often interested in trying to reproduce some target phenomenology, so they often do not tinker in a neutral way. The specific problem with computer simulations is that, in many cases, getting the phenomenology right is somewhat too easy, and the general problem of under-determination of theoretical claims by the evidence is particularly acute.

First, computer simulations are often used in cases where data are scarce, incomplete, or of low quality (see, e.g., [34.78, Chap. 10] for the case of climate data and how making data global was a long and difficult process). The scarcity of data can also be a primary motivation to use computer simulations to inquire about a system for which experiments are difficult or impossible to carry out, like in astrophysics [34.139]. Furthermore, knowledge of the initial and boundary conditions out of which the computer simulations should be fed may also be incomplete, which leaves more latitude for scientists to fill in the blanks and possibly match data. As a result, confidence in the result of computer simulations like the Millennium Run and in their representational and explanatory success is in part undermined [34.139].

Second, computer simulations usually involve more variables and parameters than theories. For example, for a  $10 \times 10$  grid with cells characterized by three variables, the total number of variables is already 300. This raises the legitimate suspicion that, by tuning variables in an appropriate way, there is always a means to obtain the right phenomenology. (Ad hoc tuning is of course not completely straightforward, since the many variables involved in a computer simulation are usually jointly constrained. Typically, in a fluid simulation, all cells of the grid obey the same update rule and are correlated.) This possibility of tuning variables and parameters is indeed used in fields like machine

learning, which can be based, for example, on the use of artificial neurons. In such fields, one first combines a limited number of elementary mathematical functions (e.g., artificial neurons) that, when adequately parameterized, reproduce potentially complex behaviors found in databases (the learning phase). In a second step, one uses the parameterized functions (e.g., the trained neural network) on new cases in the hope that extrapolation and prediction are possible. In such cases, even if the right phenomenology is reproduced, and extrapolation partly works, it is clear that the trained neural network and the corresponding mathematical functions do not explain the phenomena. Overall, this means that the ability to reproduce some potentially complex phenomenon is far from being sufficient to claim that the corresponding computer simulation has explanatory power (see also [34.140] for the issue of the over-fitting of computer simulations to data).

Third, when scientists do succeed, they may be subject, as other human creatures, to confirmation biases, overweigh their success and tend to ignore the fact that various mechanisms or laws can produce the same data (or that other aspects of their computer simulations do not fit). While such biases are not specific to computational inquiries, they are all the more epistemologically dangerous since matching phenomena is easy.

### Complex Systems Resist Explanation

Because they are very powerful tools, computer simulations are specifically used for difficult investigations, which usually have features that may spoil their explanatory character [34.141, 142]. Typically, in the natural sciences, computer simulations and computational methods are centrally used for the study of so-called complex systems [34.143, 144], see also Chap. 35. Realistically investigating complex systems would imply taking into account many interrelated nonlinear aspects of their dynamics including long-distance interactions and, in spite of the power of modern computers, the corresponding models are usually intractable. Therefore, drastic simplifications need to be made in both the construction of the model and its mathematical treatment, which often threatens the epistemic value of the results.

Importantly, for the above reasons, the problem of the explanatory value of computer simulations can arise even in fields like fluid dynamics where the underlying theories are well known. It is no surprise that this problem is more acute in fields, such as the human and social sciences, in which no such theories are available, the investigated objects are even more complex, sound data are more difficult to collect and interpret, and the very nature of what counts as a sound explanation and genuine understanding is more debated [34.145, 146] especially in relation to computer simulations [34.133].

For these reasons, even if there are good arguments for claiming that computer simulations do not fare worse than other methods like analytic models or experiments (see [34.40] for the case of economics), it is not surprising that their potential explanatory power is undervalued.

Overall, it is plausible that often computer simulations have less explanatory power than other methods, and that this does not stem from their nature but from the type of uses they usually have in science. If this is the case, the question of the explanatory power of computer simulations is to be treated on a case-by-case basis by using the same criteria as for assessing the explanatory power of other scientific activities, *pace* the distrust that shrouds the use of computer simulations.

#### 34.4.4 Too Replete to Be Explanatory? The Era of Lurking Suspicion

Theories of explanation should capture our intuitions about what is explanatory. From this point of view, it is interesting to see whether computer simulations meet these intuitions, especially when they fulfill the explanatory requirements described by theories of explanations.

##### Computer Simulations and Explanatory Relevance

Good explanations should not include explanatorily irrelevant material. While determining whether some piece of information is explanatorily relevant to explain some target fact is a scientific task, finding a satisfactory notion of explanatory relevance is a task for philosophers. Despite progresses concerning this problem, current accounts of explanation still fall short of capturing this notion [34.147, 148]. At the same time, existing results are sufficient to understand why computer simulations raise concerns regarding explanatory relevance.

Scientific information, in particular causal laws, accounts for the behavior of phenomena. Thus, it is legitimate, when trying to explain some phenomenon, to show that its occurrence can be derived from a scientific description of the corresponding system. Nevertheless, even then, one may fall short of satisfying the requirement of explanatory relevance. This is clearly explained by *Salmon* in his 1989 review of theories of explanation, where he asks “Why are irrelevancies harmless to arguments but fatal to explanations?” and further states that “irrelevant premises are pointless, but they have no effect whatever on the validity of the argument” [34.149, p. 102]. While philosophers have mainly focused on the discussion of irrelevant unscientific premises, the problem actually lies deeper. Parts of the content of

laws or mechanisms, essentially involved in explanatory arguments, can be irrelevant to the explanation of aspects of phenomena that are covered by these laws or mechanisms [34.148]. So the problem is not simply to discard inessential (unscientific or scientific) premises, but also to determine, within the content of the scientific premises that are essentially used in explanatory derivations, what is relevant and what is not [34.102, 148, 150].

This problem is especially acute for computer simulations. Take a computer simulation that unfolds the detailed evolution of a system based on the description of its initial state and the laws governing it. Then all aspects of the computational model are actually used in the computational derivation. At the same time, all such aspects are not necessarily explanatorily relevant with respect to all facets of the computed behavior. Typically, some aspects of the computed behavior may simply depend on the topology of the system, on symmetries in its dynamics or initial conditions, on the fact that some initial quantity is above some threshold, etc.

Accordingly, the following methodological maxim may be proposed: *the more an explanation (resp. an argument) contains independent pieces of scientific information, the more we are entitled to suspect that it contains irrelevancies (regarding the target behavior).*

At the same time, one should remain aware that explaining some target phenomenon may sometimes irreducibly require that all the massive gory details involved in the simulation of the corresponding system are included. For example, as chaos theory shows it, explaining the emergence and evolution of a hurricane may essentially require describing the flapping of a butterfly’s wings weeks earlier.

An additional problem is that there is no general scientific method to tell whether a premise, or some part of the information it conveys, is relevant. Contrarily to what the hexed salt example [34.151] may perhaps suggest, irrelevant pieces of information within an explanation do not wear this irrelevance on their sleeves and are by no means easy to identify. This is the *problem of the lack of transparency, or of opacity, of irrelevant information.*

Overall, since they are based on informationally replete descriptions of their target systems, computer simulations legitimately raise the suspicion of being computational arguments that contain many irrelevancies, and therefore of being poor explanations – even when they are not.

##### Computer Simulations, Understanding, and Inferential Immediacy

Mutatis mutandis, similar conclusions can be reached regarding the issue of computational resources. Since

this issue is closely related to the question of how much computer simulations can bring about understanding, things shall be presented through the lens of this latter notion.

It is usually expected that explanations bring understanding. Theorists of understanding, while disagreeing on the precise nature of this notion, have explored its various dimensions, which provides a good toolkit to analyze how computer simulations fare on this issue.

Hempel cashes in the notion of understanding in terms of nomic expectability. From this point of view, taken as explanatory arguments, computer simulations seem able to provide understanding since, like other scientific representations, they can rely on nomological regularities. Further, in contrast to sketchy explanations, they make the nomic dependence of events explicit. Consider the explanation analyzed by *Scriven* that “the impact of my knee on the desk caused the tipping over of the inkwell” [34.152]. The *hidden strategy* described by *Woodward* [34.153] is to claim that the value of this latter nonnomological explanation is to be measured against an *ideal* explanation, which is fully deductive and nomological and describes the detailed succession of events that led to the stain on the carpet, even if this complete explanation is often inaccessible. From this point of view, a computer simulation can offer a way to approach such an ideal explanation, by providing an explicit deduction of the lawful succession of events that brought about the explanandum. However, an epistemic problem is that, once such a computer simulation has been carried out (and properly stored), it is possible to explicitly highlight any part of it, but it is not possible to scrutinize all parts because there are too many of them. This is one of the reasons why computer simulations are intrinsically opaque to human minds [34.41, §5.3], see also Sect. 34.3.2.

Be this as it may, causal theorists of explanation should agree that computer simulations often contribute significantly to developing our understanding by reducing uncertainty about the content of causal ideal explanatory texts, as requested in [34.131].

Computer simulations also seem to be able to provide unificatory understanding. For unificationists like *Kitcher*, understanding is a matter of “deriving descriptions of many phenomena using the same pattern of derivation again and again” [34.107, p. 423]. Since computer simulations offer more ways of deriving phenomena, by providing new patterns of derivation or instantiating existing patterns in more complex cases, at least some of them contribute to unification.

Things are less straightforward with *Woodward’s* account of explanation and understanding. *Woodward* argues that a good explanation provides “understanding

by exhibiting a pattern of counterfactual dependence between explanans and explanandum” [34.154, p. 13]. From this point of view, computer simulations fare well since, if one does not go beyond their domain of validity, they provide general patterns of counterfactual dependence between their inputs  $I$  and outputs  $O$ , which are obtained by applying  $t$  times their update algorithms ( $UA$ ), that is, more formally,  $O(t, I) = UA^t(I)$ .

Is there a philosophical catch? *Woodward* also requires that the pattern of counterfactual dependence be described in terms of a functional relation. But what is to count as a *function* in this context? Functions can be defined explicitly (by means of algorithms) or implicitly (by means of equations). The advantage of computer simulations is that they provide algorithmic formulations based on elementary operations of how the explanandum varies with the explanans. From this point of view, computer simulations are more explicit than models, which simply provide equations linking the explanans and the explanandum. However, the problem is that with computer simulations any kind of functional immediacy is lost, since it is computationally costly to carry out the algorithm. Indeed, *Woodward* usually describes straightforward examples of functional dependence like  $Y = 3X_1 + 4X_2$ . With such functions, we may feel that the description of the counterfactual dependence is *just there*, since, by *simply instantiating* the variables and carrying out the few operations involved, specific numerical relations are accessible. In such simple cases, a human mind can do the work by itself and answer the corresponding what-if-things-had-been-different (what-if) questions. In contrast, with a computer simulation, computing the output takes much computational power. So the tentative conclusion is that computer simulations provide understanding in *Woodward’s* sense, but this understanding is not immediately accessible, the degree of (non)-immediacy being described by the computational resources it takes to answer each what-if question. Importantly, an equation-based model may give the illusion of immediacy, since the equation presents a short description of how the variables are correlated. However, one should watch out that short equations can be unsolvable, and short descriptions of algorithms (like  $O(t, I) = UA^t(I)$ ) with simple inputs can yield complex behaviors that are computationally costly to predict [34.155].

Similar conclusions can be reached if one focuses on analyses of understanding proper. *De Regt* and *Dieks* propose to analyze understanding in terms of intelligibility, where this latter notion implies the ability to recognize qualitative characteristic consequences without performing exact calculations [34.156]. In this sense, understanding seems to be a matter of immediacy, as was already suggested by *Feynman*, who described it as

the ability to foresee the behavior of a system, at least qualitatively, or the consequences of a theory, without solving exactly the equations or performing exact calculations [34.157, Vol. 2, 2–1].

Depending on the cases, foreseeing consequences requires logical and cognitive operations to a greater or lesser extent. Thus, the above ideas may be rephrased in a more gradualist way, by saying that the less inferential or computational steps one needs to go through to foresee the behavior of a system or the consequences of a theory, the better we understand it. In this perspective, computer simulations fare terribly badly, since they involve going through many gory computational steps and, even once these have been carried out, scientists usually end up with no simple picture of the results and no inferential shortcuts that could exempt them from this computational stodginess for future similar investigations.

#### Understanding: What Do We Lose with Computer Simulations?

Before the advent of computational science, explanatory advances in science were always the direct product of human minds and pen-and-rubber methods. Therefore, any actual scientific explanation that satisfied the requirements for explanatoriness was also human sized, and the epistemic benefits logically contained within such explanations could actually be enjoyed by competent and informed epistemic agents. In [34.158, p. 299], *Hempel* states that an explanatory argument shows that “the occurrence [of an event] was to be expected” and he adds “in a purely logical sense.” This addition emphasizes that expectation should not be understood as a psychological notion nor refer to the psychological aspects of the activity of explaining. In the case of computer simulations, this addition is somewhat superfluous. Nomic expectability remains for scientists, since, based on computer simulations, they may know that they can entertain the belief that an event should happen. However, this belief is completely *cold*. Since the activity of reasoning is externalized in computers, it is no longer part of the proper cognition of scientists and does not come with the psychological side-effects associated with first-person epistemic activities, such as emotions or feelings of expectation, impressions of certainty and clarity, or the oft-mentioned aha or eureka feeling which usually comes with first-person experiences of understanding. In other words, with computer simulations, the mind is no longer the carrier of the activity of explanation, and simply records what it should believe. Unfortunately, epistemic benefits associated with the individual ability to carry out this activity are also lost. Since the explanatory argument can no longer be surveyed by a human mind, the details of the rela-

tions between the premises of the explanatory argument and its conclusion are opaque. Therefore, scientists are no longer able to encompass *uno intuitu* all aspects of the explanation and how they are related, to develop expectations about counterfactual situations (in which similar hypotheses are met), and the unificatory knowledge that only global insights can provide is also lost. Overall, with computer simulations the objective intelligibility that is enclosed in explanations and can be accessed by first-person epistemic appropriation of the explanatory arguments can no longer be completely enjoyed by scientists (see also [34.159] for further analyses about epistemic opacity in this context). In this perspective, the problem of computer simulations is not that they have less explanatory value but that *we cannot have epistemic access* to this explanatory value. In brief, this problem would not pertain to the logic of computer-simulation-based explanations but to their epistemology.

#### New Standards for Understanding?

The gradualist description regarding the need of cognitive and logical operations to foresee consequences (see Sect. 34.4.4 Computer Simulations, Understanding, and Inferential Immediacy) suggests that the boundary between cases where intelligibility is present or is lost is not completely sharp. Importantly, the ability to foresee consequences depends on various factors such as the knowledge of physical or mathematical theorems to facilitate deductions, the knowledge of powerful formalisms to facilitate inferences, how much the intuition of scientists has been trained to anticipate consequences of a certain type and has somewhat internalized inferential routines, etc., [34.102, §6.4]. In other words, at least in some cases, the frontiers of what has a computational explanation, but remains unintelligible to a human mind, can be pushed back to some extent.

This raises the question of how much the frontiers of intelligibility can be extended and whether the ideal of inferential or computational briefness for explanations should be considered as a normative standard. Two positions are possible. One may claim that genuine explanations should *always* yield the possibility for human subjects to access the corresponding understanding. Or one may claim that, as shown by computational science, we have gone beyond human-sized science, not all good explanations can be comprehended by human minds, and this is not a defect of *our* science, even if it is clearly an epistemic inconvenience.

A motivation for endorsing the former claim is that the lack of intelligibility of explanations often stems from epistemic flaws of the agents producing them and can be corrected. Typically, in science, results are often

laboriously proved and, with the advance of scientific understanding, shorter and clearer proofs, or quicker algorithms, are found.

Overall, it seems sound to adopt the following methodological maxim: *the more resources we need to produce (or check) an explanation (resp. an argument, a proof), the more we are entitled to suspect, in the absence of contrary evidence, that the explanation is unduly complex*. From this point of view, computer simulations do not seem flawless, since they make abundant use of computational and inferential resources. Accordingly, it is legitimate to suspect computer simulations of providing unduly complex explanations, which have simpler versions yielding the expected accessible understanding.

Nevertheless, this philosophical stance may be inappropriate in many cases. There is a strong suspicion that explaining phenomena often requires using an irreducible amount of resources. This idea of computational irreducibility has been vocally advanced, though not clearly defined, by Wolfram [34.155], and philosophers have toyed with close intuitions in recent discussions about emergence [34.74, 123, 124, 160–162]. Capturing the idea in a clear, robust and fruitful definition is a difficult on-going task [34.163]. However, there seems to be an agreement that this intuitive notion is not empty, which is what matters for the purpose of the present discussion. Overall, this means that in all such cases, asking for computationally simple explanations does not make sense, since such explanations do not exist. In this perspective, tailoring our explanatory ideals to our human capacities is wishful thinking, since in many cases, the inaccessibility of the usual epistemic benefits of explanations does not stem from our epistemic shortcomings.

This suggests that we may have to bite the bullet and say that, sometimes, computer simulations do bring full-fledged explanation and objective understanding, even if, because of our limited cognitive capacities, we cannot enjoy this understanding and the epistemic benefits harbored by such explanations. In other words, both of the above philosophical options are correct, though in different cases.

Ideally, one would like to be able to know when each of these two options should be adopted. Unfortunately, determining whether a computational process can be shortcut or a computational problem solved by quicker algorithms, seems to be in practice opaque (*problem of the lack of transparency of the optimality of the computational process*). This means that in most cases, when facing a computational explanation of a phenomenon, one does not know whether there are computationally or inferentially shorter versions of this explanation (and we are to be epistemically blamed

for being so explanatorily laborious), or whether one cannot do better (and the process is intrinsically complex).

Overall, because determining whether explanations are informationally minimal (regarding the use of relevant information) and whether arguments or computations are optimal is opaque, computer simulations are doomed to remain shrouded in suspicion about their explanatoriness, *even in cases in which there is no better (that is, shorter or less informationally replete) explanation*. In brief, the era of suspicion regarding the explanatoriness of computer simulations will not end soon.

#### 34.4.5 Bypassing the Opacity of Simulations

Even when computer simulations are epistemically opaque, some strategies can be tried to regain predictive power, control, and potentially understanding regarding the corresponding inquiries.

##### Understanding, Control and Higher Level Patterns

As emphasized by Lenhard [34.159], by *manipulating* computational models and observing which behavior patterns are obtained, scientists can try to control the processes involved and develop “a feeling for the consequences.” Lenhard suggests that this *understanding by control*, which is oriented toward design rules and predictions, corresponds to a pragmatic account of understanding, which is also involved in the building of reliable technological artifacts.

Other authors have emphasized that, even if the details of computer simulations cannot be followed by human minds, one may sometimes still obtain valuable insights by building coarse-grained representations of the corresponding target systems and analyzing whether macro-dynamics emerge when microinformation is thrown away [34.164]. Surprisingly, the existence of coarse-grained dynamics seems to be compatible with complex, potentially computationally irreducible, dynamics at the microlevel [34.165, 166], even if this by no means warrants that control or understanding can always be regained at the macro-level. Thus, the question arises as to when and how much epistemic virtues like predictive power, control, and potentially understanding, which are somewhat lost at the microlevel, can be partly recovered at the macro-level, and how the corresponding patterns can be detected. The treatment of such questions requires the analysis of logical and mathematical relations between descriptions of systems at different scales and, for this reason, it should gain from ongoing debates and research in the philosophical



and scientific literature about the emergence of simple behavior in complex systems.

### Visualization and Understanding

Another important issue is how to exploit macro-level patterns that are present in computer simulations to restore partial cognitive grasp of the simulated systems by humans. Given the type of creatures that we are, and in particular the high visual performance of the brain, using visual interfaces can be part of the answer. Indeed, the format of scientific representations partly determines what scientists can do with them – whereas, as emphasized by [34.41, p. 95], philosophers have often considered the logical content of a representation to be the only important element to analyze them. To go further into these issues, sharp analyses of representational systems and their properties are required. Tools and concepts developed in the Goodmanian tradition prove to be extremely useful [34.167]. For example, *Kulvicki* [34.29] highlights how much graphs and images can present information more immediately than symbolic representations can. This notion of immediacy is cashed in in terms of semantic salience, syntactic salience or extractability. *Vorms* further shows how taking into account formats of representation in the analysis of scientific reasoning is crucial, since inferences have different cognitive costs depending on the format of representation [34.168]. *Jebeile* [34.169] applies similar concepts to computational models and argues that visualization tools can have a specific explanatory role since they do not merely present computational data in more accessible ways, but also suggest interpretations that are not contained in the original data, highlight relations between these data, and thereby point at elements of answers to what-if questions.

Overall, the issue of how much visualization can convey objective understanding remains debated. For

example, *Kuorikoski* [34.164] acknowledges that visual representations are cognitive aids but emphasizes that they often merely bring about a *feeling* and *illusion* of understanding. So, there is the need of epistemological analyses which would make clear in which cases, and how, visual representations can be reliable guides and self-certifying vectors of knowledge, which partly enable their users to determine whether and how much they should trust them.

### 34.4.6 Understanding and Disciplinary Norms

All the above discussion has been based on general arguments about explanations and understanding. However, as already emphasized, explanatory norms sometimes differ from one field to another, economics being, at least in its mainstream branches, a paradigmatic case of a field in which simulation methods are shunned [34.37]. Similarly, the explanatory status of computer simulations and computational models varies across fields like cognitive sciences, artificial intelligence [34.137], artificial life [34.170] or within fields themselves (see, e.g., [34.171] for the case of computational chemistry and [34.79] for that of climate science).

This is not the place to discuss whether these variations regarding explanatory norms are deep, or whether they result from differences in theoretical contexts, in the degrees of complexity of the systems investigated, in the difficulties to collect evidence about them, in the scientific maturity and empirical success of these fields, etc. Such questions cannot be answered on the basis of armchair investigations. Field-specific studies of the explanatoriness of computer simulations, made by scholars who are in the same time acutely aware of present discussions about scientific explanation, are needed.

## 34.5 Comparing: Computer Simulations, Experiments and Thought Experiments

Computer simulations, experiments, and, to a lesser extent, thought experiments share various similarities, which calls for an explanation. Indeed, similarities between experimental activities and computational science are even found in mathematics, where some methods are claimed to be *experimental* (Sect. 34.5.1). Computer simulations, experiments and thought experiments can sometimes be seen as ways of carrying out similar activities, or activities having similar constraints (Sect. 34.5.2). Should an additional step be made, and computer simulations be considered as experiments?

A close scrutiny of the existing arguments in favor of this claim shows that it meets insuperable difficulties, both regarding the analysis of computer simulations and experiments. Further, the claim does not even seem necessary to account for the importance of the material aspects of simulations (Sect. 34.5.3). Finally, even if computer simulations can yield knowledge, which can sometimes be more reliable than that produced by experiments, unless a strong case against empiricism is properly made, computer simulations do not seem to seriously threaten the unique foundational role of exper-

iments as the source of primary evidence upon which science is built (Sect. 34.5.4). In any case, discussions about the relationships between experiments and computer simulations should remain compatible with the actual existence of hybrid (both computational and experimental) methods (Sect. 34.5.5).

When in the 1990s philosophers of science started investigating computer simulations, they soon realized that the object of their inquiry cross-cut traditional categories like those of theories, models, experiments or thought experiments. Similarities with experiments were particularly striking, since, among other things, computer simulations involved the treatment of massive data and statistical reasoning, required robustness analysis, and were claimed to yield new knowledge. As a result, computer simulations were suggestively dubbed by various authors as computer *experiments*, numerical *experimentation* or in-silico *thought experiments*, even though it was not always conceptually clear what these potentially metaphorical characterizations meant exactly.

All such similarities are worth analyzing and potentially call for explanations. They may be the sign of an identical nature between (some of) these activities, of common essential features, or may just be shallow or fortuitous. Clarifying this issue is also a way to analyze these activities more acutely by singling out what is specific to each or common to them and to determine to what extent epistemological insights can be transferred between them.

### 34.5.1 Computational Mathematics and the Experimental Stance

#### Experimental Proofs in Mathematics

Since aspects related to the representation of material systems are absent from mathematics, a comparison with this field can be hoped to be fruitful to *analyze* what exactly is experimental in computational science.

The mathematical legitimacy of computers for the production of proofs has been discussed for several decades. Computational proofs like that of the four-color theorem by Appel et al. [34.172, 173] were rapidly labeled *quasi-empirical* and discussions raged about how they should be interpreted [34.174, p. 244]. Such computational *proofs* can actually be seen as having roots in the older tradition of *quasi-empirical mathematics*, practiced for example by mathematicians like Euler, and philosophically defended by authors like Lakatos [34.175] or Putnam [34.176]. Interestingly, even in these contexts, the labels *empirical* or *experimental* were used to refer to various aspects of the activity of proving results.

Like experiments, computational proofs involve external processes, which are fallible. Their reliability can then be seen as being partly of a probable nature and needs to be assessed a posteriori by running these external processes several times and checking that the apparatus involved worked correctly. By contrast, proofs which can be *actively* and *directly* produced by humans minds, can provide a priori knowledge, the validity of which is assessed by (mentally) inspecting the proof itself, qua mathematical entity. Further, computational proofs, like experiments and empirical methods in mathematics, usually provide particular numerical results: as the computational physicist Keith Roberts writes it, “each individual calculation is [...] analogous to a single experiment or observation and provides only numerical or graphical results” (quoted in [34.70, p. 137]). Therefore, to obtain more general statements (and possibly theories), probabilistic inductive steps are needed. Overall, such debates illustrate the need to clarify the use in this context of labels like *experimental* or *empirical*.

#### The Experimental Stance

The case of computational mathematics also makes clear how scientists can adopt an experimental stance for inquiries where no physical process is investigated, and the nature of the object which is *experimented upon* is completely known.

Experimenting involves being able to trigger changes, or to intervene on material or symbolic dynamical processes, and to record how they vary accordingly. As noted by Dowling [34.136, p. 265] and Jebeile [34.169, II, §7.2], processes for which the dynamics is known can also work as black boxes, since the opacity of the process may stem either from our lack of knowledge about its dynamics, or from the mathematical unpredictability (or epistemic inaccessibility) of its known dynamics. In this perspective, contrarily to *Guala* [34.177], being a black box is not a specific feature of experiments.

Finally, when experimenting on a material or formal object, it is better that interactions with the object be made easy and the results be easily accessible to the experimenters (e.g., by means of visual interfaces) so that tinkering is made possible [34.136] and intuitions, familiarity, and possibly some form of understanding [34.159, 169, III] can be developed.

### 34.5.2 Common Basal Features

Some similarities of computer simulations and experiments (and thought experiments) may be accounted for by highlighting common basal features of these activities, which in turn account for the existence of their common epistemological features, such as the shared

concerns of practitioners of experiments and computer simulations for “error tracking, locality, replicability, and stability” [34.70, p. 142]. In this perspective, one should characterize the nature and status of these common basal features.

### Role or Functional Substitutability

Though computer simulations, thought experiments and experiments are activities of different types, they can sometimes be claimed to play identical roles. Typically, computer simulations are used to gain knowledge about how physical systems behave (hereafter *behavioral knowledge*) when experiments are unreliable, or making them is politically or ethically unacceptable [34.41, p. 107]. Importantly, acknowledging that computer simulations can sometimes be used as substitutes for experiments by no means implies that they can play *all the roles of experiments* (Sect. 34.5.4). Further, one should be aware that, at a high-level of abstraction, all activities may be described as doing similar things; therefore, these shared roles should be shown in addition to have nontrivial epistemological implications. For example, one may argue that *providing knowledge* or *producing data* are roles that are endorsed by computer simulations, thought experiments, or experiments. However, this may be seen as some partially sterile hand-waving. Indeed this points at a too abstract similarity if these activities produce items of knowledge of totally different types, and nothing epistemologically valuable can be inferred from this shared characterization (see [34.81] for a presentation of the different types of knowledge involved in science).

*El Skaf and Imbert* [34.87] make an additional step when they claim that these activities can in certain cases be *functionally substitutable*, that is, that we can sometimes use one instead of the other for the purpose of a common inquiry – which remains compatible with the fact that these activities do not play the roles in question in the same way, that they come with different epistemic credentials, provide different benefits, and therefore, as role holders, are not *epistemologically substitutable*. *El Skaf and Imbert*, in particular, claim that computer simulations, experiments, and thought experiments are sometimes used for the purpose of unfolding scenarios (see also *Hughes’* notion of demonstration in Sect. 34.6.1) and argue that investigations concerning the possibility of a physical Maxwellian demon were indeed pursued by experimental, computational and thought experimental means. The existence of such common roles then provides grounds for analyzing similarities in the epistemological structure of the corresponding inquiries.

*Morrison* [34.178] goes even further since she argues that some computer simulations are used *as mea-*

suring instruments and therefore that they have the same epistemic status as experimental measurements. She first claims that models can serve as measuring instruments, and then shows that this role can be fulfilled in connection with both computer simulations and experiments, which are similarly model shaped. An important part of her strategy is to relax the conditions for something to count as an experiment, by discretely giving primacy, in the definitions of scientific activities, to the roles which are played (here measuring) and by downplaying the importance of physical interactions with the investigated target systems in the definition of experiments (which are simply seen as *a way* to perform this measuring role). *Giere’s* rejoinder denies the acceptability of this strategy, and follows the empiricist tradition, when he claims that “a substitute for a measurement is not a measurement, which traditionally requires causal interaction with the target system” [34.179, p. 60]. Indeed, the potential additional pay-offs of experiments, as primary sources of radically new evidence, come from these causal interactions. Accordingly, their specificity is not due to their roles, qua information sources (since thought experiments, models, or theories are also information sources), but from the type of epistemological credentials that come with the corresponding information, and grounds our ultimate scientific beliefs. A different nonempiricist epistemology might be developed, but the bait must then be swallowed explicitly, and it must be explained why such an epistemology, in which activities are exclusively individuated on the basis of their function and the importance of other differences is downplayed, should be preferred. In any case, an account of how to individuate these functions would be needed, since at a high level of abstraction, various activities can be seen as performing the same function.

### Beyond Anthropocentric Empiricism

To practice science, humans need to collect observations and make inferences. Since human capacities are limited, various instruments have been developed to extend them and these instruments have been partly computational for decades. These parallel developments of observational and inferential capacities come with common epistemological features. In both cases, restricted empiricism, which gives a large and central role to *human* sensorial or inferential capacities in the description of how scientific activities are carried out, is no longer an appropriate paradigm to understand scientific practices. Indeed, the place of human capacities within modern science needs to be reconsidered [34.8, 41, 180]. Further, the externalization of observations and inferences comes at the price of some epistemic opacity and passivity for the practitioner, since, as humans, we

no longer consciously carry out these activities. Instead we simply state the results of experimental or computational apparatus. However, this also comes with gains in objectivity since observational and informational procedures are now carried out by external, transparent and controlled apparatus, which no longer have hidden psychological biases nor commit fallacies.

The development of computational instruments and computer simulations also raises similar epistemological problems. For example, the apparently innocuous notion of data seems to raise new issues in the context of computational science. Computer simulations, like models, have been claimed to be useful to *probe* physical systems and to be used as *measuring instruments* [34.178]. Whatever the interpretation of such statements (Sect. 34.5.3), it is a fact that both computer simulations and computational instruments provide us with *data*, which raises transversal questions.

A datum is simply the value of a variable. It can be taken to describe a property of any object. In this simple sense, data coming from experiments and computer simulations can play a similar role by standing for the properties of some target system within some representational inquiries. Furthermore, in both cases, their interpretation usually involves heavy computational treatments. In particular, mathematical transforms of various types serve to separate information from noise, remove artifacts, or recover information about a system property out of intertwined causal signals, like in computed tomography imaging techniques [34.121]. From this point of view, as emphasized by *Humphreys* [34.181], here one departs from a principle frequently used by traditional empiricists, and according to which “the closer we stay to the raw data, the more likely those data are to be reliable sources of evidence.”

At the same time, there are different types of epistemological data, and the need for their common study should not introduce confusion in their understanding. In science, one seeks to determine how much data reliably stand for their target, and which properties exactly they refer to. *Humphreys*’s remark above the computational treatment of data, reproduced above, highlights the fact that causal information concerning the source is crucial to treat and interpret data and to determine what empirical content they bring about this source (this is the *inverse inference problem*), given that data do not wear on their sleeves details of how they were produced. From this point of view, experimental and computational data have utterly different causal histories – so what gives its sense to the computational treatment is potentially of a different nature [34.91, 121]. Overall, more pointed comparative analyses of data obtained by computer simulations and computational instruments are still to be carried out, to understand their semantics

and epistemology and highlight both their nonaccidental similarities and specific differences (see [34.182] for the case of computational instruments).

Computational science must also face the challenge of data management. While the steps of *traditional* mathematical proofs and arguments, once produced, can be verified by scientists, things are usually different for computer solutions, even if they are merely executions of computational programs [34.91], or arguments [34.72]. Details of computer simulations are in general not stored since this would require too large amounts of memory (even if, in some cases like the Millennium Run, scientists may decide to keep track of the evolution of the computer simulation). In other words, like experimental science, computational science involves choosing which data to keep track of, developing powerful devices to store them, finding appropriate ways to organize them, providing efficient interfaces to visualize, search, and process them, and, more generally, developing new methods to produce knowledge from them. This also raises questions about how these data can or should be accessed by the scientific community, and which economic model is appropriate for them [34.183]. In brief, the epistemology of computer simulations here meets that of big data [34.184, 185], even if it cannot be assumed that on-going debates and analysis about the latter, because they are mostly focused on questions raised by empirically collected data, will naturally apply to, or be insightful for, the corresponding problems raised by computer simulations.

#### Different Activities, Similar Patterns of Reasoning

As noted by *Parker* [34.186], strategies developed to build confidence in experimental results, and described in particular by Allan Franklin, seem to have close analogs for the justification of results generated by computer simulations. Indeed, the interpretation of the results of computer simulations as evidence for hypotheses about physical systems can sometimes be made through an error-statistical perspective [34.187] as in the case of experiments [34.188].

Similar patterns of reasoning are also used to argue in favor of the existence of specific mechanisms or entities on the basis of patterns within data, modes of visualizations of these patterns, or our ability to manipulate the actual or represented systems and find pattern regularities in their behavior (see [34.71] for a description of the *homomorphic tradition*, in which visual forms are given much importance, in contrast to the *homologic tradition*, which is more based on logical relationships). More generally, visualization techniques, aimed to facilitate the reasoning about results present in

large databases, are crucial in the case of both experiments and computer simulations (Sect. 34.4.4).

Importantly, these similarities may have different explanations. For example, they may simply stem from the need to treat massive amount of data by efficient standard procedures, or be a consequence of features shared by experimental and computational data, independently of their quantity, like the presence of noise, or may correspond to the application of general types of evidential or explanatory arguments to data having different natures.

### The Reproducibility of Results

Reproducibility is a typical requirement for experiments, though it is one that is sometimes difficult to achieve because of the tacit knowledge involved in the carrying out of experiments [34.189]. Similar problems may arise with computer simulations. Even if the latter are nothing more than computations and are in principle reproducible, in practice reproducibility may sometimes be difficult, especially in the context of big science. For example, computer simulations may be too big to be reproduced (all the more since scientists have in general little incentive to reproduce results). Numerical codes may not be public (because they are not published or shared), and many of the computational details may be left tacit. Finally, computer simulations involving stochastic processes may not be exactly reproducible because the random numbers came from external physical signals or because the details of the pseudorandom number generator are not made public.

### Experimenters' and Simulationists' Regresses

Good scientific results are usually expected to be robust against various changes [34.190], in particular those related to implementation or material details, and this is why failure of exact reproducibility should not be a worry.

Still, when one faces an inability to reproduce a result, the problem may arise from a lack of robustness or flaw in the original experiment or computer simulation, or from a failure to reproduce it correctly. Accordingly, as emphasized by *Gelfert* [34.191], computer simulations are affected by a problem similar to that of the experimenter's regress [34.192], which is met when to determine whether an experimental apparatus is working properly scientists have no criterion other than the fact that it produces the expected results. As noted by *Godin* and *Gingras* [34.193], regresses like that highlighted by Collins are instances of well-known types of arguments already analyzed in the framework of ancient skepticism (more specifically, regresses or circular relations regarding justification). As such, they are specific neither to experiments nor to computer simulations –

even if solutions to these problems, as those described by *Godin* and *Gingras* or *Franklin* [34.194], may be partly activity specific. In any case, adopting a general comparative perspective provides a way to analyze more acutely what is epistemologically specific or common to scientific activities.

### 34.5.3 Are Computer Simulations Experiments?

Some authors go as far as claiming that, at least in *some* cases, what we call computer simulations *are* in fact experiments. In this perspective, Monte Carlo methods, sometimes labeled *Monte Carlo experiments* or *Monte Carlo simulations*, seem to be a philosophical test case (like analog simulations, Sect. 34.2.2). Such methods are used to compute numbers (e.g., pi), sample target distributions or produce dynamical trajectories with adequate average properties. They rely crucially on the use of randomness [34.8, 72]. They may look closer to experiments because they sometimes use physical systems, like a Geiger counter, to generate random events.

Still, *Beisbart* and *Norton* claim that Monte Carlo methods are not experiments, since randomizers can be replaced by computer codes of pseudorandomizers [34.72, p. 412]. This shows that these computer simulations do not require contact with the randomizer as an external object; therefore no direct empirical discovery about the nature of physical systems can be made by them and they should not be seen as having an experimental nature. In brief, in Monte Carlo simulations, the physical systems involved are simply used as computers to generate mathematically random sequences.

Beyond the analysis of specific cases, some authors have defended the bolder claim that all computer simulations are experiments (what *Winsberg* calls the *identity thesis* [34.195, §5]). While this goes against inherited scientific common sense (computations are not experiments!), the claim should be carefully examined. Indeed, in principle there is no impossibility here: while computations, logically defined, are not experiments, we need physical machines to carry them out. Therefore, in the end, computers, instruments and experimental systems are physical systems that we use for the purpose of doing science – and it all boils down to how we conceptualize in a coherent and fruitful way these external worldly activities. In brief, perhaps, after all, we would be better off revising our epistemological notions so that computer simulations are seen as genuine examples of experiments – a revisionary position with regard to the empiricist tradition since it ignores the specificity of experiments as primary evidential sources of knowledge.

In what follows, I review existing arguments in favor of the claim that computer simulations are experiments, and how these arguments have been criticized. Overall, as we shall see, in contrast to what is claimed in [34.195, §5], it is very dubious that discussions about the *identity thesis* are simply a matter of perspective and where the emphasis is placed. A minute, conceptually rigorous, and sharp treatment of this question can be found in [34.53, 72, 91, 196] and [34.169, Chap. 7].

### Problems with Analyses in Terms of Common Physical Structures

Some authors analyze computer simulations as manipulations of physical systems (the computers), which instantiate or realize models that are also instantiated or realized by the investigated physical systems.

*Norton and Suppe* [34.114] are good representatives of this tradition. They first try to describe formal relations between what they call a *lumped model*, the structure of the target system, and the programmed computer, which is supposed to embed the lumped model. They further argue that these relations account for the experiment-like character of computer simulations: instead of experimenting on real systems, computer simulations are used as physical stand-ins or analogs to probe real-world phenomena, and one thereby learns thing about the represented systems. This suggestive position has charmed various authors. It also has similarities with accounts of scientific representation made in terms of similarity [34.28], isomorphism, or weaker relationships between the representation and the target system [34.197, 198], even if the authors that defend the above view have not adopted so far this line of argument.

However, in the case of computer simulations, this view does not seem to resist close scrutiny, for reasons specific to computational activities. While in the case of analog simulations both the represented system and the analog computer instantiate a common mathematical structure (Sect. 34.2.2), such a claim cannot be made for digital computers. The general idea is that steps of computational processes are multiply realizable and that, conversely, how physical states of computers are to be interpreted is contextual and partly arbitrary [34.4]. It is true that for every step of a computation to be carried out in practice, one needs to use a physical machine that can be seen as instantiating the corresponding transition rule. However, physically different machines can be used to carry out different parts of a computation (for example when the computation is distributed). Furthermore, even if a single machine is used, different runs of the program will correspond to different physical processes, since the computer may process several tasks in the same time and contextually

decide how its memories are organized, and even within the same computation, a single part of the memory may be used at different steps to code for different physical variables [34.91, pp. 564–566], [34.196, pp. 81–84]. Overall, in the general case, the relation between the physical states of the represented target system and the physical states of the computer(s) that may be used to simulate its behavior is a many-many one, and the idea that the phenomenon is recreated in the machine “is fundamentally flawed for it contradicts basic principles of computer architecture” [34.196, p. 84]: in the case of a successful computer simulation, one can simply say that every step of the computation has been carried out by some appropriate physical mechanism, but there is no such thing as a computer instantiating the structure of the model investigated. (Note that the argument based on multiple realizability is in the spirit of those originally developed by *Fodor* [34.126] in his discussion of the reduction of the special sciences).

### Problems with Common Analyses in Terms of Intervention or Observation

Computer simulations have also been claimed to qualify as experiments “in which the system intervened on is a programmed digital computer” [34.199, p. 488], or to involve observations of the computer as a material system [34.114, p. 88]. *Winsberg* even goes as far as to claim that [34.195]

“nothing but a debate about nomenclature [...] would prevent us from saying that the epistemic target of a storm simulation is the computer, and that the storm is merely the epistemic motivation for studying the computer.”

Such claims can be answered along the same lines as the previous argument. There is of course no denying that when one runs a computer simulation one interacts with the interface of the computer, which triggers some physical change in the computer so that the right computation is carried out. Similarly, once the computation is finished, the physical state of the memory in which the result is stored, triggers a causal mechanism that produces changes in the interface so that the result can be read by the user. However, the definition of an intervention at the model level does not determine a specific intervention at the physical level of the computer. The reason is that, as emphasized above, even within the same computational process, the way that the intervened model variable is physically represented in the computer may vary, and how the computer, qua physical system, evolves precisely may depend on various parameters such as the other tasks that it carries out at the same time. In brief, the idea that actual com-

puter simulations, defined at the model level, could be seen as the investigation of the computer, qua physical machine, which is used to carry them out, seems to be riddled with insuperable difficulties.

Finally, one should mention that epistemic access to the physical states of the computer corresponding to the successive steps of a computation is usually not possible in practice [34.196, p. 81].

#### Problems with General Analyses in Terms of Epistemological and Representational Structure

Some authors have also argued that computer simulations and experiments share an epistemological structure, or epistemological aspects, and have used this claim to justify the identity thesis.

For example, it has been claimed that in both cases one interacts with a system to gain knowledge about a target system, and the internal and external validity of the processes needs to be checked. This type of analysis stems from a 2002 paper by *Guala* [34.177] in which he presents a laboratory experiment in economics aimed at investigating behavioral decision making by giving decisional tasks to real human subjects in the laboratory. In this case, a hypothesis about how agents behave in the laboratory is investigated (internal validity hypotheses); then, based on similarities between the experimental situation and the real-life situation, an external hypothesis is made about the behavior of agents in real life situations (external validity hypothesis). The notion of internal validity comes from social science and corresponds to the (approximate) truth of inferences about causal relationships regarding the system that is experimented on. External validity corresponds to the (approximate) truth of the generalization of causal inferences from an initial system, for which internal validity has been demonstrated, to a larger class of systems. *Guala* further claims that both computer simulations and experiments fit this epistemological description in terms of internal and external validity arguments, but cautiously concludes that their “difference must lie elsewhere” [34.177]. According to him, computer simulations and experiments are different, since in the latter case there is a material similarity between the object and the target, whereas, in the former case, there is a formal similarity between the simulating and the simulated systems (a claim which seems to be falling under the above criticism directed at Norton and Suppe and their followers).

*Guala*’s conceptual description is endorsed by most authors who try to picture computer simulations as some sort of experiment. For example, *Winsberg* accepts the description, but claims that the difference between experiments and computer simulations lies

in the type of background knowledge that researchers use to justify the external validity hypothesis [34.113, p. 587], a position which is again revisionary with regard to the empiricist tradition if this is the only specificity ascribed to experiments.

A serious worry is that describing the investigation of the computational model in terms of internal validity is problematic and artificial, since, as can be seen above, computer simulations cannot be considered as investigations of the causal behavior of the computer, qua physical system. For the same reason, the use of the notion of external validity is inappropriate, since for computer simulations inferences about the target system do not involve the generalization of causal relations taking place in the computer to other systems by comparing their material properties but involve the *representational* validity of the computational model.

A final problem is that the characterization of the methodology of experimental studies in terms of internal and external validity, though useful in the social sciences, is not a general one. Using it as an accepted general framework to compare experiments and computer simulations looks like a hasty extrapolation of the case of laboratory experiments in experimental economics, not to mention the fact that economics may be seen as a bold pick to build a general conceptual framework for experimental studies.

It is true that in experiments, the measured properties are often not the ones that we are primarily interested in and the former are used as evidence about these latter target properties. Typically, vorticity in turbulent flows is difficult to measure directly, and is often assessed by measuring velocity, based on imaging techniques. In more complex cases, the properties measured can be seen as a way to *observe* different and potentially remote target systems, as is vividly analyzed by *Shapere* with his case study of the observation of the core of the sun by the counting of  $^{37}\text{Ar}$  atoms in a tank within a mine on Earth [34.180]. Importantly, in all such cases, the measuring apparatus, the directly measured property, and the indirectly probed target system are related by causal processes. The uses of the collected empirical information then vary with the type of inquiry pursued. The evidence may be informational about the physics of a particular system, like the Sun. Or, it may be used to confirm or falsify theories (like in the case of the 1919 experiment by Eddington and the relativity theory). In some cases, *though by no means all*, it may be used to draw inferences about the nature or behavior of a larger class of similar systems – which are not related to the measured system by a causal relationships. If this latter case of reasoning about external validity is taken as paradigmatic for experiments, and the causal processes between the target experimented

systems (the source) and the measuring apparatus (the receptors), which are present in all experiments, are considered as a secondary feature, experimental activities are misrepresented. As Peschard nicely puts it “the idea that the experiments conducted in the laboratory are aimed at understanding some system that is outside the laboratory is a source of confusion” [34.200]. General conceptual frameworks that do not introduce such confusion are however possible. For example, *Peschard* proposes [34.200]

“to make a distinction between the *target system*, that is manipulated in the experiment or represented in the computer simulation, and the *epistemic motivation*, which in both cases may be different from the target system ”

(see also the distinction between the result of the unfolding of a scenario and the final result of the inquiry in [34.87]).

Overall, the common description provided by Guala, and heavily relied upon in [34.113, 199] to support versions of the *identity thesis* can be defended only by squeezing experiments and computer simulations into a straightjacket which misrepresents these activities, is not specifically fruitful, and meets insuperable difficulties.

#### Materiality Matters

Clearly, for both experiments and computer simulations, materiality is crucial. However, it does matter differently, and one does not need to endorse a version of the identity thesis to acknowledge the importance of materiality when claiming for example that, to understand computational science, the emphasis should be on computer simulations which can be in practice, and therefore materially, carried out by actual systems [34.41, 91].

For experiments, material details are relevant throughout the whole inquiry when producing, discussing and interpreting results, their validity and their scope (especially if one tries to extrapolate from the investigated system to a larger class of materially similar ones). By contrast, for computer simulations, material details are important to establish the reliability of the computation, but not beyond: only the mathematical and physical details of the investigation matter when discussing and interpreting the results of the computer simulation and the reliability of the inquiry.

#### 34.5.4 Knowledge Production, Superiority Claims, and Empiricism

The question of the epistemic superiority of experiments over simulations has also been discussed. *Parke*

[34.201] takes it for granted that “experiments are commonly thought to have epistemic privilege over simulations” and claims that this is in fact a context-sensitive issue. As we shall see, if one puts aside the question of the specific role of experiments as the source of primary evidence about nature, it is not clear whether the general version of the superiority claim has actually been defended, or whether a straw man is attacked.

#### Computer Simulations, Experiments and the Production of Radically New Evidence

Let us try to specify what the general superiority claim could be and how it has really been defended.

The obvious sense in which experiments may be *superior* is that they can provide scientists with primary evidence about physical systems, which originate in interactions with these systems, and cannot be the product of our present theoretical beliefs. It is unlikely that computer simulation can endorse this role. As *Simon* pithily puts it, “a simulation is no better than the assumptions built into it, and a computer can do only what it is programmed to do” [34.12, p. 14]. From this perspective, experiments have the potential to surprise us in a unique way, in the sense that they can provide results contradictory to our most entrenched theories, whereas a computer simulation cannot be more fertile than the scientific model used to build it (even if computer simulations can surprise us and bring about novel results, see Sect. 34.3.4). This is what *Morgan* seems to have in mind when she emphasizes that “[N]ew behaviour patterns, ones that surprise and at first confound the profession, are only possible if experimental subjects are given the freedom to behave other than expected,” whereas “however unexpected the model outcomes, they can be traced back to, and re-explained in terms of, the model” [34.202, pp. 324–5]. In brief, experiments are superior in the sense that, in the empirical sciences, they can serve a function which computer simulations cannot.

*Roush* [34.203] has highlighted another aspect regarding which experiments can be superior to simulations. She first insists that we should compare the two methods *other things being equal*, especially in terms of what is known about the target situation. Then, in any case in which there are elements in the experimenter’s study system that affect the results and are unknown, we may still run the experiment and learn how the target system behaves; by contrast, in the same epistemic situation, the simulationist cannot build a reliable computer simulation that yields the same knowledge. However, when all the physical elements that affect the result are known, a simulation may be as good as an experiment, and it is a practical issue to determine which one can in practice be carried out in the most reliable way.



Thus, for a quantitative comparison to be meaningful it should be related to roles which can be shared by experiments and computer simulations, such as the production of *behavioral* knowledge about physical systems, the relevant dynamics of which is known (Sect. 34.5.2).

### Grounds for Comparative Claims

Scientists and philosophers have emphasized over the last decades that computer simulations are often *mere simulations* [34.177], the results of which should be taken carefully. As seen above, economists *shun* simulation; similarly, *Peck* states that evolutionary biologists view simulations with suspicion and even contempt [34.204, p. 530]. Nevertheless, however well advised these judgments may be, they cannot by themselves support a *general* and *comparative* claim of superiority in favor of other methods, but at most the claim that, in fields where other methods are successful and computer simulations have little epistemic warrants or face serious problems, these other methods will usually or on average be more reliable (exceptions remaining possible).

Some authors have discussed the comparative claim by analyzing the power of the types of inferences made to justify knowledge claims in each case. In [34.199], *Parker* adopts *Guala*'s description of experiments (resp. computer simulations) as having material (resp. formal) similarities with their target systems (see the discussion in Sect. 34.5.3) and studies the claim that inferences made on the basis of material similarities would have an epistemic privilege. (*Guala* does not seem to endorse a comparative claim. He argues that material similarities are a specific feature of experiments, implying that the prior knowledge needed to develop simulations is different from that needed to develop experiments.) Again, the common description in terms of internal and external validity regarding the inferences from one physical system to another gives the semblance of a new problem. However, if, as suggested above, the material properties of computers matter only in so far as they enable scientists to make logically sound computations, and no similarity between systems is involved, the grounds and rationale for this discussion between the properties of the computer and those of the target system collapse. A way to save the argument is to claim that the aforementioned formal similarities are simply those between the computational model and the target system, but then the question boils down to the much more familiar comparison between model-based knowledge (here extracted by computational means) and some type of experiment-based knowledge.

On what grounds could the general privilege of experiment-based behavioral knowledge then be de-

fended? Since experiments and computer simulations are different activities, which are faced with specific difficulties, it is hard to see why computer simulations should always fare worse. Why could simulations based on reliable models not sometimes provide more reliable information than hazardous experiments? Indeed, it is commonly agreed that, when experiments cannot be carried out, are unreliable, or ethically unacceptable, computer simulations may be a preferable way to gain information [34.41, p. 107].

### Justified Contextual Superiority Claims

Interestingly, superiority claims can sometimes be made in specific contexts. *Morgan* presents cases in economics in which a precise and contextual version of the superiority claim may be legitimate [34.202].

Like *Guala*, *Morgan* discusses laboratory experiments in economics, that is, purified, controlled, and constrained versions of real world systems, which are studied in artificial laboratory environments (in contrast with field experiments, which “follow economic behavior *in the wild*” [34.202, p. 325]) and are aimed at investigating what is or would be the case in actual (nonsimplified) economic situations. Mathematical models can also be used for such inquiries and, in each case, scientists run the risk of describing artificial behaviors. *Morgan* then makes the following contextual claim that “any comparison with the model experiment is still very much to the real experiment’s advantage *here*” [34.202, p. 321] (my emphasis) on the grounds that, in this case, the problem of making ampliative analog inferences from laboratory system to real-world systems is nothing compared with the problem of the realism of assumptions for models exploring artificial models [34.202, pp. 321–322]. She does not justify this point further, but a plausible interpretation is that, in such cases, mathematical models necessarily abstract away essential parts of the dynamics of decision making, which arguably are preserved in experiments because of the material similarity between the laboratory and real agents. In brief, while material similarity plays a role in her argument she does not make the general claim in the core of her paper that material similarity will *always* provide more reliable grounds for external validity claims than other methods (even if her formulation is less cautious in her conclusion).

Overall, such sound contextual comparative judgments require two premises: first that in some context computer simulations are not reliable (or have reliability  $r$ ) and second that in the same context material similarities provide reasonably reliable inferences (or have reliability  $s > r$ ). (Indeed, analogical reasoning based on material similarities, in which one reasons based on systems that are representative of or for larger

classes of systems [34.127], can sometimes be powerful ways to make sound – though not infallible! – contextual inferences. As emphasized by *Harré* and *Morgan*, “shared ontology [...] has epistemological implications” [34.202, p. 323], since “the apparatus is a version of the naturally occurring phenomenon and the material setup in which it occurs” [34.205, pp. 27–8]. After all, different samples of the same substance obey the same laws, even if contextual influences may change how they behave and any extrapolation is not possible.)

### 34.5.5 The Epistemological Challenge of Hybrid Methods

Whether computer simulations and experiments are ontologically, conceptually, and epistemologically distinct activities or not, it is a fact that jointly experimental and computational *mixed* activities have been developed by scientists. Their study was pioneered by *Morgan*, who presents various types of hybrid cases in economics [34.206] and biomechanics [34.127]. For example, she reports different mixed studies aimed at investigating the strength of bones and carried out by cutting slices of bone samples, photographing them, creating digital 3-D images, and applying the laws of mechanics to these experiment-based representations. *Morgan* further attempts to provide a principled typology of these activities. This proves difficult because “modern science is busy multiplying the number of hybrids on our epistemological map” and because the qualities of hy-

brids “run along several dimensions” [34.127, p. 233]. Overall, sciences illustrate “how difficult it is to cut cleanly, in any practical way, between the philosopher’s categories of theory, experiment and evidence” [34.127, p. 232], and, we may add, computer simulations or thought experiments.

Should these hybrid methods lead philosophers to reconsider the conceptual frontiers between experiments and computer simulations? We can first note that their existence may be seen as a confirmation that the traditional picture of science, in which theoretical, representational or inferential methods on one hand and experimental activities on the other play completely different but complementary roles, is not satisfactory (Sect. 34.5.2). Then, if one grants that activities like experiments, thought experiments and computer simulations can sometimes play identical roles, it is no surprise that they can also be jointly used to fulfill them. Similarly, a group of four online players of queen of spades sometimes involve virtual players – but most people will be reluctant to see this as sufficient grounds for claiming that bots are human creatures.

In any case, these hybrid activities raise epistemological questions. What, if anything, distinguishes a computer simulation that makes heavy use of empirical data from a measurement involving the computational refinement of such data [34.53, 121]? How much should the results of these methods be considered as empirical? Overall, what type of knowledge and data is thereby generated (see [34.53] for incipient answers)?

## 34.6 The Definition of Computational Models and Simulations

The main definitions of computer simulations are critically presented: Humphreys’s 1994 definition in terms of computer-implemented methods, Hartmann’s 1996 definition in terms of imitation of one process by another process, Hughes’s DDI (denotation, demonstration, interpretation) account of theoretical representation, and finally Humphreys’s 2004 definition, with its emphasis on the notion of a computational template (Sect. 34.6.1). The questions that a satisfactory definition should answer are then discussed, in particular which notions should be primitive in the definition, whether computer simulations should be defined as logical or physical entities, whether they correspond to success terms, how the definition should accommodate the possibility of scientific failure and the pursuit of partly open inquiries, or to what extent computer simulations are social, intentional, or natural entities (Sect. 34.6.2).

I come back finally to the issue of the definition of computer simulations. Providing a definition may look at first sight to be easy, since what computers are is well-known and clear cases of computer simulations are well identified. However, a sound definition should also be helpful to analyze less straightforward cases and be fruitful regarding epistemological issues related to computer simulations, not least by forcing philosophers to clarify the various intuitions which are entertained across scientific fields about these methods.

It is not difficult to present definitions that accommodate *some* types of computer simulations or *some* particular (or field specific) uses of computer simulations. Nevertheless, failing to distinguish between what is typical of computer simulations in general and what is specific to particular cases can lead (and has led) to heedless generalizations (Sect. 34.5.3). Things are all the more tricky as the very same types of computer

simulations, qua formal tools (e.g., agent-based, CA models, equation-based simulations, etc.) can be used in different epistemic contexts for different purposes, and require totally different epistemological analyses. The case of CA-based computer simulations exemplifies the risk of too quick *essentialist* characterizations. While it was believed that these models were appropriate for phenomenological simulations only [34.9, 135], their use in fluid dynamics has shown that they could supply theoretical models based on the same underlying physics as traditional methods [34.100].

The following section is organized as follows. Existing definitions and the problems they raise are presented first, and then issues that a good definition of computer simulations should clarify are emphasized.

### 34.6.1 Existing Definitions of Simulations

#### Computer-Implemented Methods

As emphasized by Humphreys, a crucial feature of simulations is that they enable scientists to go beyond what is possible for humans to do with their native inferential abilities and pen-and-paper methods. Accordingly, he offered in 1991 the following working definition [34.7]:

“A computer simulation is any computer-implemented method for exploring the properties of mathematical models where analytic methods are unavailable.”

This definition requires that we possess a clear definition of what counts as an analytic method, which is not a straightforward issue [34.60]. Further, as noted by Hartmann et al. [34.10, pp. 83–84], it is possible to simulate processes for which available models are analytically solvable. Finally, as acknowledged by Humphreys, the definition covers areas of computer-assisted science that one may be reluctant to call computer simulations. Indeed, this distinction does sometimes matter in scientific practice. Typically, economists are not reluctant to use computers to analyze models but shun computer simulations [34.37]. Since both computational methods and computer simulations involve computational processes, their difference must be either in the different types (or uses) of computations involved either at the mathematical and/or the representational level.

#### One Process Imitating Another Process

Hartmann proposes the following characterization, which gives the primacy to the representation of the temporal evolution of systems [34.10, p. 83]:

“A model is called *dynamic*, if it [...] includes assumptions about the time-evolution of the system.

[...] Simulations are closely related to dynamic models. More concretely, a simulation results when the equations of the underlying dynamic model are solved. This model is designed to imitate the time-evolution of a real system. To put it another way, a simulation imitates one process by another process. In this definition, the term process refers solely to some object or system whose state changes in time. If the simulation is run on a computer, it is called a computer simulation.”

This definition has been criticized along the following lines. First, as noted by Hughes [34.13, p. 130], the definition rules out computer simulations that do not represent the time evolution of systems, whereas arguably one can simulate how the properties of models or systems vary in their phase space with other parameters, such as temperature. Accordingly, a justification for the privilege granted to the representation of temporal trajectories should be found, or the definition should be refined, for example, by saying that computer simulations represent successive aspects or states of a well-defined trajectory of a system *along a physical variable* through its state space. Second, the idea that a specific trajectory is meant to be represented may also have to be abandoned. For example, in Monte Carlo simulations, we learn something about *average values* of quantities along sets of target trajectories by generating a potential representative of these trajectories, but the computer simulations are not aimed at representing any trajectory in particular. One may also want a computer simulation to be simply informative about structural aspects of a system. Overall, the temporal dynamics of the simulating computer is a crucial aspect of computer simulations since it “enables us to draw conclusions about the behavior of the model” [34.13, p. 130] by unfolding these conclusions in the temporal dimension of our world, but the temporal dynamics of the target system may not have to be represented for something to count as a computer simulation.

Third, the definition is probably too centered on models and their solutions [34.207], since it equates computer simulations with the solving of a dynamic model that represents the target system. This is tantamount to ignoring the fact that describing computer simulations as mathematical solutions of dynamic models is not completely satisfactory. What is being solved is a computational model (as in Humphreys’s definition [34.41], see below), which can be significantly different from, and somewhat independent of, the initial dynamic model of the system, which usually derives from existing theories. Effectively, different layers of models, often justified empirically, can be needed in-between [34.13, 97, 208]. For this reason, the repre-

sentational relation between the initial dynamic model and the target system, and between the computational system and the target system, are epistemologically distinct.

Finally, the definition may be reproached for entertaining a recurrent confusion about the role of materiality in computer simulations (Sect. 34.5.3), by describing the representational relation as being between two physical processes, and not between the computational model and succession of mathematical states which unfold it (*in whatever way they are physically implemented and computed*) and the target system.

### Computer Simulations as Demonstrations

*Hughes* does not propose a specific definition of computer simulations since he believes that computer simulations naturally fit in the DDI account of scientific representation that he otherwise defends [34.13, p. 132]. According to the DDI, which involves denotation, demonstration, and interpretation as components [34.13, p. 125]:

“Elements of the subject of the model (a physical system evincing a particular kind of behavior, like ferromagnetism) are *denoted* by elements of the model; the internal dynamic of the model then allows conclusions (answers to specific questions) to be *demonstrated* within the model; these conclusions can then be *interpreted* in terms of the subject of the model.”

The demonstration can be carried out by a physical model (in the case of analog simulations) or by a logical or mathematical deduction, such as a traditional mathematical proof, or a computer simulation. Further, according to *Hughes*, in contrast to Hartmann’s account, “the DDI account allows for more than one layer of representation” [34.209, p. 79]. Overall, a virtue of this account is that it emphasizes the common epistemological structures of different activities by pointing at a similar demonstrative step, which excavates the epistemic content and resources of the model (see also [34.210] for refinements, [34.87] for an analysis which extends the idea of demonstration, or *unfolding*, to thought experiments and some types of experiments, and [34.72] for the related idea that computer simulations are arguments). While as a definition of computer simulation, *Hughes*’s sketchy proposal has somewhat been neglected (see however [34.208]) it is a legitimate contender and it remains to be seen how much a more developed version of it would provide a fruitful framework for philosophical discussions about computer simulations.

### Computer Simulations as the Concrete Production of Solutions to Computational Models

In order to answer problems with the previous definitions, *Humphreys* proposed in 2004 another definition of computer simulations, which is built along the following lines [34.41]. He defines the notion of a theoretical template, which is implicitly defined as a general relation between quantities characterizing a physical system, like Newton’s second law, Schrödinger’s equation, or Maxwell’s equations. A theoretical template can be made less general by specifying some of its variables. When the result is *computationally tractable*, we end up with a computational template. (Thus, what qualifies as a computational template seems to depend on our computational capacities at a given time.) When a computational template is given (among other things) an interpretation, construction assumptions, and an initial justification, it becomes a computational model. Finally, *Humphreys* offers the following characterization [34.41, pp. 110–111]:

“System S provides a core simulation of an object or process B just in case S is a concrete computational device that produces, via a temporal process, solutions to a computational model [...] that correctly represents B, either dynamically or statically. If in addition the computational model used by S correctly represents the structure of the real system R, then S provides a core simulation of system R with respect to B.”

Another important distinction lies between the computer simulation of the behavior of a system and that of its dynamics [34.41, p. 111] since, even when the computational model initially represents the structure and dynamics of the system, the way its solutions are computed may not follow the corresponding causal processes. Indeed, in a computer simulation, the purpose is not that the computational procedure exactly mimics the causal processes, but that it efficiently yields the target information from which an appropriate dynamic representation of the target causal processes can finally be built for the user. For reasons of computational efficiency, the representation may be temporally and spatially dismembered at the computational level (e.g., by computing the successive states in a different order), as may happen with the use of parallel processing, or of any procedure aimed at partially short cutting the actual physical dynamics.

The space here is insufficient to analyze all the aspects of the above definition and to do justice to their justification – all the more so since further complications may be required to accommodate even more

complex cases [34.208]. Suffice it to say that this elaborate definition, which is aimed at providing a synthetic answer to the problems raised by previous definitions, is one of the most regularly referred to in the literature.

### 34.6.2 Pending Issues

#### Simulating or Computing

Giving a definition of computer simulations implies choosing which notions should be regarded as primitive and how to order them logically. Some authors first define the notion of simulation and present computer simulations as a specific type of simulations. For example, *Bunge* first defines the notion of analogy, then that of simulation, and finally that of representation, as sub-relation of simulation. For him, an object  $x$  simulates another object  $y$  when (among other things) (1) there is a suitable analogy between  $x$  and  $y$  and (2) the analogy is valuable to  $x$ , or to another party  $z$  that controls  $x$  (see [34.11, p. 20] for more details).

A potential benefit of this strategy is that it becomes possible to unify in the same general framework various different types of analogous relations between systems such as organism versus society, organism versus automaton, scale ship versus its model, computer simulations of both molecular and biological evolution, etc. Similarly, *Winsberg* [34.195, §1.3] suggests that the hydraulic dynamic scale model of the San Francisco Bay model should be viewed as a case of simulation (see [34.211] for a recent presentation and philosophical discussion of this example in the context of modeling). While scale models can obey the same dimensionless equations as their target systems and be used to provide analog simulations of them, *Winsberg's* claim is not uncontroversial and may require an extension of the notion of simulation. Indeed the model and the Bay itself do not exactly obey the same mathematical equations. For example, distortions between the vertical and horizontal scales in the model increase the hydraulic efficiency, which implies adding copper strips and the need for empirical calibration. Therefore, this is not exactly a case of a bona fide analog simulation (Sect. 34.2.2) but of a complex dynamical representation between closely analogous systems. In any case, if one adopts such positions, it is then a small step to describe other cases of analogical reasoning between material systems (and possibly cases of experimental economics, in which the dynamics of the analogous target system is not precisely known and external validity is to be assessed by comparing the material systems involved) as cases of simulations (Sect. 34.5.3).

At the same time, unification is welcome only if it is really fruitful (and is, of course, not misleading). As seen above, the problem with such analyses is that

they tend to describe computer simulations as involving a representational relationship between two material systems and to misconstrue how computers work (see again Sect. 34.5.3). They thereby tend to misrepresent the epistemological role of the physical properties of computers and the fact that computational science involves two distinct steps; one in which computer scientists warrant that the computer is reliable and another in which scientists use computations and do not need to know anything about computers qua physical systems. A way out of this deadlock may be to use a flexible notion of simulation, which can be applied to relations between physical *or* logical–mathematical simulating processes and the target simulated physical processes. Then, the question remains as to what exactly is gained (and lost) from an epistemological point of view by putting in the same category modes of reasoning of such different types – if one puts aside the emphasis on the obvious similarities with analog simulations, which are a *very specific* type of computer simulation (Sect. 34.2.2). Overall, it is currently far from clear whether this unificatory move should be philosophically praised.

#### Abstract Entities or Physical Processes

Arguably, computations are logical entities that can be carried out by physical computers. Then, the question arises should computer simulations also be seen as abstract logical entities, or should they be seen as material processes instantiating abstract computations? *Hartmann's* definitions present computer simulations as processes, whereas *Humphreys's* definition is more careful in the sense that the computing systems simply *produce* the solution or *provide* the computer simulation. Clearly, to analyze computational science, it is paramount to take into account material and practical constraints since a computer simulation is not really a part of *our* science and we have no access to its content unless a material system *carries* it out for us. At the same time, just like the identity of a text is not at the material level, the identity of a computing simulation (and the corresponding equivalence relationship between runs of the same computer simulation) is defined at the logical (if not the mathematical) level and the physical computer simply presents a token of the computer simulation. From this point of view, the material existence of computer simulations and the in principle/in practice distinction emphasized by *Humphreys* [34.41] have epistemological, not ontological, significance, that is, they pertain to what we may learn by interacting with actual tokens of computer simulations [34.91, p. 573] but not to the nature of computer simulations. Similarly the identity of a proof seems to be at the logical level, even if a proof has no existence

nor use for us unless some mathematician provides some token of it.

### Success, Failure, and the Definition of Computer Simulations

A computer simulation is something that reproduces the behavior or dynamics of a target system. The problem with characterizations of this type is that they make computer simulation a success term and if a computer simulation mis-reproduces the target behavior, it is no longer a computer simulation. This problem is a general one for representations, but is specifically acute for scientific representations (*Frigg* and *Nguyen* Chap. 3, this volume). Indeed, while anything in art can be used to represent anything else, scientific representations are meant to be informative about the natural systems they represent. This is part of their essential specificities and, arguably, a definition according to which any process could be described as a scientific computer simulation of any other process is not satisfactory. At the same time, one does not want something to be a computer simulation, or a scientific representation, based on whether it is scientifically successful and exactly mirrors its target (remember that, for some scientific inquiries, representational faithfulness is not a goal and may even impede the success of the investigation [34.212] and [34.23, Chaps. 1 and 3]).

An option is to say that something is a scientific representation if it correctly depicts what its user wants it to represent. However, this may raise a problem for computer simulations that were carried out and had subsequent nonintended uses, like the millennium simulation. It may also raise a problem for fictions, which strictly speaking seem to represent nothing [34.25, p. 770].

Finally, failed representations, which do not represent what their producers believe them to depict, are also a problem. Representational inquiries can fail in many ways, and failures are present on a daily basis in scientific activity, from theories and experiments to models and simulations. For this reason, descriptions of scientific activities should be compatible with failure, especially if they are to account for scientific progress and the development of more successful inquiries. Indeed, it would be weird to claim that many of the computer simulations that scientists perform and publish about are actually not computer simulations. Further, whether a genuine computer simulation is carried out should be in general transparent to the practitioner, and this cannot be the case if computer simulation is defined as a success term and scientific failure is frequent (see also [34.213, pp. 57–58]).

Overall, a question is to determine where the frontier should lie between unsuccessful or failed computer

simulations, and potential cases in which something that was believed to be a computer simulation by scientists actually is not. This in turn requires knowing how computer simulations can fail specifically [34.92] and which failures are specific to them. In brief, one needs to be able to decide on a justified basis which failures disqualify something from being a computer simulation and which ones simply alter its scientific, epistemic, or semantic value. This analysis may also have to be coherent with analyses about how other types of scientific activities such as experiments and thought experiments can fail [34.214], especially when these activities play similar or identical roles.

An option to consider is that something is a computer simulation based on criteria that do not involve empirical success, and that it qualifies as an empirical success depending on additional semantic properties and on whether it correctly represents the relevant aspects of its (real or fictional) target system(s). This option is potentially encompassing enough (the scientifically short-sighted student can be said to perform a computer simulation), but discriminating between good and bad computer simulations is still possible. It is compatible with the fact that research inquiries are often open and scientists need not know in advance what in their results will have representational value in the end. Finally, it is also compatible with a different treatment of representational and implementation failures. Indeed, the possibility of being unsuccessful at the representational level is consubstantial to empirical inquiries and is in this sense *normal*. By contrast, an implementation failure is simply something that should be fixed. It corresponds to a case in which we did not manage to carry out the intended computation, whereas computing is not supposed to be a scientific obstacle, and we learn nothing by fixing the failure.

### Natural, Intentional, or Social Entities?

A similar but distinct issue is to determine which type of objects computer simulations are, *qua token physical processes carried out by computing devices* – a question which is close to that of the nature of physical computers and is also related to that of the ontology of model (*Gelfert* Chap. 1, this volume).

Arguably, they are not simply natural objects which are defined by some set of physical properties and exist independently of the existence of the agents using them. Indeed, because computations can be multirealized and some runs of computations built by patching different bits of computation on physically different machines, it is unlikely that all computations can be described in terms of natural kind predicates (massively disjunctive descriptions not being allowed here) [34.126].

Further, for both computations and computer simulations, pragmatic conditions of use seem to matter. To quote *Guala* commenting on the *anthropomorphism* of Bunge's definition (see above), [34.177, p. 61]

“it makes no sense to say that a natural geyser *simulates* a volcano, as no one controls the simulating process and the process itself is not useful to anyone in particular.”

Indeed, even if any physical system can be seen as computing some (potentially trivial) functions (see below), any physical object cannot be *used* as a (general) computer, and we may have to endorse a position along the lines of Searle's notion of social objects [34.215], or of any analysis doing the same work: a physical object X counts as Y in virtue of certain cognitive acts or states out of which they acquire certain sorts of functions (here computing), given that these objects need to demonstrate appropriate physical properties so that they may serve these functions for us. A specificity of computer simulations is that, unlike entrenched social objects, such as cars or wedding rings, a small group of users may actually be enough for a physical system to be seen as carrying out a computer simulation. Thus, the evolution of a physical system (like a fluid) may count for some users as an analog computer, which performs a computer simulation, and for other users as an experiment, even if experiments and computer simulations are in general objects of different types, and this case is unlikely to be met in practice (Sect. 34.5.3).

In any case, what is needed for something to be used as a computer or a computer simulation is not completely clear. The physical process must clearly be recognized as instantiating a computer model. Control is useful but not necessarily mandatory (e.g., we may use the geyser to simulate a similar physical system, even if the geyser would not count as a controlled *versatile* analog computer). The possibility to extract the computed information is clearly useful – an issue that matters for discussions about analog and quantum computer simulations, and of course cryptography.

An alternative position is not to mention users in the definition and to claim that, pace the peculiar case of man-made computations (which may make use heavily of the possibility offered by multiple realizability, see Sect. 34.5.3), physical processes are the one-piece physical instantiations of running computer models (resp. computer simulations) and, as such, are computations (even if, sometimes, trivial ones). See [34.216] for a sober assessment of this pancomputationalist position. In this perspective, one may say that it is a practical problem to create artificial human-friendly computers which can *in addition* be controlled and the informa-

tion of which can be extracted. While such positions may be palatable for those, like *Konrad Zuse*, *Edward Fredkin* and their followers [34.64, 217, 218], who want to see nature as a computer, it is not clear that such pan-computationalist theses, whatever their intrinsic merits for discussing foundational issues like the computational power of nature or which types of computers are physically possible, serve the purpose of understanding science as it is actually practiced.

An important distinct question is whether intentional or pragmatic analyses should also be endorsed regarding computational models and computer simulations, *qua representational mathematical entities*, that is, how much the intentions of users and conditions detailing how their use by scientists is possible, should be part and parcel of their definitions. Arguably, a scientific model is not simply a piece of syntax or an entity which inherently and by itself represents, completely or partially, a target system in virtue of the mathematical similarities it intrinsically possesses with this system. In order to understand how scientific representations and computer simulations work and actually play their scientific role, their description may have to include captions, legends, argumentative contexts, intentions of users, etc., since these elements are part of what makes them scientifically meaningful units. Indeed, how one and the same mathematical model represents significantly varies depending on the inquiry, subject matter and knowledge of the modelers. This is particularly clear in the case of computational templates, which are used across fields of research for different representational and epistemological purposes [34.41, §3.7], and which are scientific units at the level of which different types of theoretical and conceptual exchanges take place within and across disciplines [34.45]. Overall, this issue is not specific to computer simulations but can be raised for other scientific representations [34.23, 168, 219–221]. Thus, this point shall not be developed further.

### Computer Simulations and Computational Inquiries

How should computer simulations be delineated? Computer simulations do not wear on their sleeves how they were built, contribute to scientific inquiries, should be interpreted and how their results should be analyzed. Accordingly, authors like Frigg and Reiss distinguish between computer simulations in the narrow sense (corresponding to the use of the computer), and in the broad sense (corresponding to the “entire process of constructing, using and justifying a model that involves analytically intractable mathematics” [34.30, p. 596]). See also the distinction between the unfolding

of a scenario and the computational inquiry involving this unfolding at its core [34.87], or the description of how the demonstration activity is encapsulated in other activities in the DDI account of representation [34.13].

Whatever the choice which is made, there is tension here. As underlined above, an analysis of the identity of scientific representations cannot rest on the logical and mathematical properties of scientific models and their similarities with their physical targets, and indications about how these representations are to be interpreted cannot be discarded as irrelevant to the analysis of their nature and uses. At the same time, computer simulation, qua computational process, and the arguments that are developed by humans about it, are activities of different natures and play different roles. Therefore an encompassing definition should not lead to blur the specificities of the different components of computational inquiries (just like a good account of thought experiments should not blur that they crucially involve mental activities at their core and are part of inquiries also involving scientific arguments).

### 34.6.3 When Epistemology Cross-Cuts Ontology

Whatever the exact definition of computer simulations, it is clear that they are of a computational nature, involve representations of their target systems and that their dynamics is aimed at investigating the content of these representations.

Importantly, whereas the investigation of scientific representations is traditionally associated with the production of theoretical knowledge, the nature of com-

puter simulations does not seem to determine the type of knowledge they produce.

Clearly, computer simulations can yield theoretical knowledge when they are used to investigate theoretical models. At the same time, even if computer simulations are not experiments (Sect. 34.5.3), they produce knowledge, which may qualify as empirical in different and important senses. As we have seen, computer simulations provide information about natural systems, the validity of which may be justified by empirical credentials rooted in interactions with physical systems for aspects as various as the origin of their inputs, the flesh of their representations of systems (see in Sect. 34.5.5 the examples by Morgan about the studies of the strength of bones), the calibration or choice of their parameters, or their global validation by comparison with experiments (Sect. 34.3.2). However, information about the dynamics represented cannot completely be of empirical origin, since it involves the description of general relations between physical states, and general relations cannot be observed.

From this point of view, computer simulations may be seen as a mathematical mode of demonstrating the content of scientific representations that is in a sense neutral regarding the type of content that is processed: empirically (resp. theoretically) justified representations in, empirically (resp. theoretically) justified information (or knowledge) out. This suggests that when analyzing and classifying types of scientific data and knowledge, the ways that they are produced and processed (experimentally or computationally) and where their reliability comes from (e.g., theoretical credentials or experimental warrants) are, at least in part, independent questions.

## 34.7 Conclusion: Human-Centered, but no Longer Human-Tailored Science

Computer simulations and computational science keep developing and partly change scientific practices (Sect. 34.7.1). Human capacities no longer play the role they had in traditional science, hence the need to analyze the articulation of computational and mental activities within computational science (Sect. 34.7.2). This requires in particular studying computational science for its own sake, which however should not be seen as implying that computer simulations always correspond to scientific activities of radically new types (Sect. 34.7.3). In any case, whatever the exact relations between computer simulations and traditional activities like theorizing, experimenting or modeling, it is a fact that recent investigations about computer simulations

have shed light on epistemological issues which were de facto not treated in the framework of previous philosophical studies of science (Sect. 34.7.4).

Before the development of computers, humans were involved at every step of scientific inquiries. Various types of devices, tools, or instruments were invented to assist human senses and inferential abilities, and they were tailored to fit human capacities and organs. In brief, science was for the most anthropocentric science, that is to paraphrase *Humphreys* [34.41, §1.1] “science by the people for the people,” and analysts of science, from Locke, Descartes, Kant to Kuhn, or Quine offered a human-centered epistemology [34.123, 124, p. 616]. Similarly, theories and models needed to be couched



in formalisms which made their symbolic manipulation possible for humans (hence the success of differential calculus), problems were selected in such a way that they could be solved by humans, results were retrieved in ways such that humans could survey or browse them, etc.

### 34.7.1 The Partial Mutation of Scientific Practices

The use of computers within representational inquiries has modified, and keeps modifying, scientific practices. Theorizing is easier and therefore less academically risky, even in the absence of well-entrenched backing-up theories; solutions to new problems become tractable and how scientific problems are selected evolves; the models which are investigated no longer need to be easily manipulated by human minds (e.g., CA are well adapted for computations, but ill-suited to carry out mental inferences [34.43]; the exploration of models is primarily done by computers, making mental explorations and traditional activities like thought experiments are somewhat more dispensable [34.117] and [34.41, pp. 115–116]; the treatment of computational results, as well as their verification, is made by computational procedures; the storage of data, but also their exploration by expected or additional inquirers, are also computer based. Finally, the human, material, and social structure of science is also modified by computers, with a different organization of scientific labor, the emergence in the empirical sciences of computer-oriented scientists, like numerical physicists and computational biologists or chemists, or the development of big computational pieces of equipment and centers, the access to which is scientifically controlled by the scientific community (like for big experimental pieces of equipment).

### 34.7.2 The New Place of Humans in Science

Overall, the place and role of humans in science has been modified by computational science. Arguably, human minds are still at the center of (computational) science, like spiders in their webs or pilots in their spacecrafts, since science is still led, controlled, and used by people. Thus, we are in a hybrid scenario in which we face what *Humphreys* calls the anthropocentric predicament of how, we, as humans, can “understand and evaluate computationally based scientific methods that transcend our own abilities” [34.42, p. 134]. In other words, interfaces and interplays between humans and computers are the core loci from

which computational science is controlled and its results skimmed by its human beneficiaries. More concretely, scientific problems still need to be selected; computational models, even if designed for computers, need to be scientifically chosen (e.g., CA-based models of fluids were first demonstrated to yield the right Navier–Stokes-like behavior by means of traditional analytic methods [34.43]; results of computer simulations, even if produced and processed by computers, need to be analyzed relative to the goals of our inquiries; and ultimately scientific human-sized understanding needs to be developed for new fundamental or applied scientific orientations to be taken.

### 34.7.3 Analyzing Computational Practices for Their Own Sake

Over the last three decades, philosophers of science have emphasized that in most cases computer simulations cannot simply be viewed as extensions of our *theoretical* activities. However, as discussed above, the assimilation of computer simulations with experimental studies is still not satisfactory. A temptation has been to describe the situation as one in which computer studies lay in-between theories and experiments. While this description captures the inadequacy of traditional characterizations based on a sharp and exclusive dichotomy between scientific activities, it is at best a metaphor. Further, this one-dimensional picture does little justice to, let alone help one understand, the intricate and multidimensional web of complex and context-sensitive relations between these activities.

An alternative is to analyze computational models, computer simulations, and computational science for their own sake. Indeed, computer simulations clearly provide a variety of new types of scientific practices, the analysis of which is a problem in its own right. Importantly, this by no means implies that these practices require a radically new or autonomous epistemology or methodology. Similarly mathematical and scientific problems can be genuinely independent, even when *in the end* they can be reduced by complex procedures to a set of known or solved problems. Indeed, the epistemology of computer simulations often overlaps piecewise with that of existing activities like theorizing, experimenting, or thought experimenting. Disentangling these threads, clarifying similarities, highlighting specific features of computational methods, and analyzing how the results of computer simulations are justified in actual cases is an independent task for naturalistic philosophers, even if one believes that, in principle, computer simulations boil down to specific mixes of already existing, more basic activities.

### 34.7.4 The Epistemological Treatment of New Issues

In practice, the analysis of computer simulations has raised philosophical issues, which were not treated by philosophers before computational studies were taken as an independent object of inquiry, either because they were ignored or unnoticed in the framework of previous descriptions of science, or because they are genuinely novel [34.96, 124, 207]. This a posteriori justifies making the epistemological analysis of computational models and computer simulations a specific field of the philosophy of science. How much computer simulations will keep modifying scientific practices and how much their philosophical analysis will finally bring about further changes in the treatment of important issues like realism, empiricism, confirmation, explanation, or emergence, to quote just a few, remains an open question.

**Acknowledgments.** I have tried to present a critical survey of the literature with the aim of clarifying discussions. I thank the editors for providing me the op-

portunity to write this article and for being so generous with space. I also thank T. Boyer-Kassem, J.M. Durán, E. Arnold, and specifically P. Humphreys for feedback or help concerning this review article. I am also grateful to A. Barberousse, J. P. Delahaye, J. Dubucs, R. El Skaf, R. Frigg, S. Hartmann, J. Jebeile, M. Morrison, M. Vorms, H. Zwirn for stimulating exchanges over the last couple of years about the issue of models and simulations and related questions. All remaining shortcomings are mine.

Various valuable review articles, such as (J.M. Durán: A brief overview of the philosophical study of computer simulations, *Am. Philos. Assoc. Newslett. Philos. Comput.* **13**(1), 38–46 (2013), W.S. Parker: Computer simulation, In: *The Routledge Companion to Philosophy of Science*, 2nd edn., ed. by S. Psillos, M. Curd (Routledge, London 2013)), have been recently written about the issue of computer simulations. (P. Humphreys: Computational science in Oxford bibliographies online, (2012) doi:10.1093/OBO/9780195396577-0100) presents and discusses important references and may be used as a short but insightful research guide.

### References

- 34.1 M. Mahoney: The histories of computing(s), *Interdiscip. Sci. Rev.* **30**(2), 119–135 (2005)
- 34.2 A.M. Turing: Computing machinery and intelligence, *Mind* **59**, 433–460 (1950)
- 34.3 A. Newell, A.S. Herbert: Computer science as empirical inquiry: Symbols and search, *Commun. ACM* **19**(3), 113–126 (1976)
- 34.4 Z.W. Pylyshyn: *Computation and Cognition: Toward a Foundation for Cognitive Science* (MIT Press, Cambridge 1984)
- 34.5 H. Putnam: Brains and behavior. In: *Analytical Philosophy: Second Series*, ed. by R.J. Butler (Blackwell, Oxford 1963)
- 34.6 J.A. Fodor: *The Language of Thought* (Crowell, New York 1975)
- 34.7 P. Humphreys: Computer simulations, *Proceedings of the Biennial Meeting of the Philosophy of Science Association*, Vol. 2, ed. by A. Fine, M. Forbes, L. Wessels (Univ. Chicago Press, Chicago 1990) pp. 497–506
- 34.8 P. Humphreys: Numerical experimentation. In: *Philosophy of Physics, Theory Structure and Measurement Theory*, Patrick Suppes: Scientific Philosopher, Vol. 2, ed. by P. Humphreys (Kluwer, Dordrecht 1994)
- 34.9 F. Rohrlich: Computer simulations in the physical sciences, *Proceedings of the Biennial Meeting of the Philosophy of Science Association*, ed. by A. Fine, M. Forbes, L. Wessels (Univ. Chicago Press, Chicago 1991) pp. 507–518
- 34.10 S. Hartmann: The world as a process: Simulations in the natural and social sciences. In: *Modelling and Simulation in the Social Sciences from the Philosophy of Science Point of View, Theory and Decision Library*, ed. by R. Hegselmann, U. Mueller, K.G. Troitzsch (Kluwer, Dordrecht 1996) pp. 77–100
- 34.11 M. Bunge: Analogy, simulation, representation, *Rev. Int. Philos.* **87**, 16–33 (1969)
- 34.12 H.A. Simon: *The Sciences of the Artificial* (MIT Press, Boston 1969)
- 34.13 R.I.G. Hughes: The Ising model, computer simulation, and universal physics. In: *Models as Mediators: Perspectives on Natural and Social Science*, ed. by M.S. Morgan, M. Morrison (Cambridge Univ. Press, Cambridge 1999) pp. 97–145
- 34.14 S. Sismondo: Models, simulations, and their objects, *Sci. Context* **12**(2), 247–260 (1999)
- 34.15 E. Winsberg: Sanctioning models: The epistemology of simulation, *Sci. Context* **12**(2), 275–292 (1999)
- 34.16 E. Winsberg: Simulations, models, and theories: Complex physical systems and their representations, *Philos. Sci.* **68**, S442–S454 (2001)
- 34.17 E. Winsberg: Simulated experiments: Methodology for a virtual world, *Philos. Sci.* **70**(1), 105–125 (2003)
- 34.18 M. Black: *Models and Metaphors: Studies in Language and Philosophy* (Cornell Univ. Press, New York 1968)

- 34.19 M. Hesse: *Models and Analogies in Science* (Sheed Ward, London 1963)
- 34.20 M. Redhead: Models in physics, *Br. J. Philos. Sci.* **31**, 145–163 (1980)
- 34.21 N. Cartwright: *How the Laws of Physics Lie* (Clarendon, Oxford 1983)
- 34.22 M. Morgan, M. Morrison: *Models as Mediators* (Cambridge Univ. Press, Cambridge 1999)
- 34.23 B. Van Fraassen: *Scientific Representation: Paradoxes of Perspective* (Clarendon Press, Oxford 2008)
- 34.24 R. Frigg: Scientific representation and the semantic view of theories, *Theoria* **55**, 49–65 (2006)
- 34.25 M. Suárez: An inferential conception of scientific representation, *Philos. Sci.* **71**(5), 767–779 (2004)
- 34.26 R. Laymon: Computer simulations, idealizations and approximations, Proceedings of the Biennial Meeting of the Philosophy of Science Association (Univ. Chicago Press, Chicago 1990) pp. 519–534
- 34.27 R.N. Giere: *Understanding Scientific Reasoning* (Holt Rinehart Winston, New York 1984)
- 34.28 R.N. Giere: *Explaining Science: A Cognitive Approach* (Univ. Chicago Press, Chicago 1988)
- 34.29 J. Kulvicki: Knowing with images: Medium and message, *Philos. Sci.* **77**(2), 295–313 (2010)
- 34.30 R. Frigg, J. Reiss: The philosophy of simulation: Hot new issues or same old stew?, *Synthese* **169**(3), 593–613 (2008)
- 34.31 M. Mahoney: The history of computing in the history of technology, *Ann. Hist. Comput.* **10**(2), 113–125 (1988)
- 34.32 D.A. Grier: Human computers: The first pioneers of the information age, *Endeavour* **25**(1), 28–32 (2001)
- 34.33 L. Daston: Enlightenment calculations, *Crit. Inq.* **21**(1), 182–202 (1994)
- 34.34 I. Grattan-Guinness: Work for the hairdressers: The production of de Prony's logarithmic and trigonometric tables, *Ann. Hist. Comput.* **12**(3), 177–185 (1990)
- 34.35 T. Schelling: Models of segregation, *Am. Econ. Rev.* **59**(2), 488–493 (1969)
- 34.36 A. Johnson, J. Lenhard: Towards a new culture of prediction. Computational modeling in the era of desktop computing. In: *Science Transformed?: Debating Claims of an Epochal Break*, ed. by A. Nordmann, H. Radder, G. Schiemann (Univ. Pittsburgh Press, Pittsburgh 2011)
- 34.37 A. Lehtinen, J. Kuorikoski: Computing the perfect model: Why do economists shun simulation?, *Philos. Sci.* **74**(3), 304–329 (2007)
- 34.38 R. Hegselmann, U. Mueller, K.G. Troitzsch: *Modelling and Simulation in the Social Sciences from the Philosophy of Science Point of View* (Springer, Dordrecht, Pays-Bas 1996)
- 34.39 G.N. Gilbert, K.G. Troitzsch: *Simulation for the Social Scientist* (Open Univ. Press, Berkshire 2005)
- 34.40 J. Reiss: A plea for (good) simulations: Nudging economics toward an experimental science, *Simul. Gaming* **42**(2), 243–264 (2011)
- 34.41 P. Humphreys: *Extending Ourselves. Computational Science, Empiricism, and Scientific Method* (Oxford Univ. Press, Oxford 2004)
- 34.42 P. Humphreys: Computational science and its effects. In: *Science in the Context of Application, Boston Studies in the Philosophy of Science*, Vol. 274, ed. by M. Carrier, A. Nordmann (Springer, New York 2011), pp. 131–142, Chap. 9
- 34.43 A. Barberousse, C. Imbert: Le tournant computationnel et l'innovation théorique. In: *Précis de Philosophie de La Physique*, ed. by S. Le Bihan (Vuibert, Paris 2013), in French
- 34.44 I. Lakatos: Falsification and the methodology of scientific research programmes. In: *Criticism and the Growth of Knowledge*, ed. by I. Lakatos, A. Musgrave (Cambridge Univ. Press, Cambridge 1970) pp. 91–195
- 34.45 T. Knuuttila, A. Loettgers: Magnets, spins, and neurons: The dissemination of model templates across disciplines, *The Monist* **97**(3), 280–300 (2014)
- 34.46 T. Knuuttila, A. Loettgers: The productive tension: Mechanisms vs. templates in modeling the phenomena. In: *Representations, Models, and Simulations*, ed. by P. Humphreys, C. Imbert (Routledge, New York 2012) pp. 3–24
- 34.47 A. Carlson, T. Carey, P. Holsberg (Eds.): *Handbook of Analog Computation*, 2nd edn. (Electronic Associates, Princeton 1967)
- 34.48 M.C. Gilliland: *Handbook of Analog Computation: Including Application of Digital Control Logic* (Systron-Donner Corp, Concord 1967)
- 34.49 V.M. Kendon, K. Nemoto, W.J. Munro: Quantum analogue computing, *Philos. Trans. R. Soc. A* **368**, 3609–3620 (2010), 1924
- 34.50 C. Shannon: The mathematical theory of communication, *Bell Syst. Tech. J.* **27**, 379–423 (1948)
- 34.51 M.B. Pour-el: Abstract computability and its relation to the general purpose analog computer (Some connections between logic, differential equations and analog computers), *Trans. Am. Math. Soc.* **199**, 1–28 (1974)
- 34.52 M. Pour-El, I. Richards: *Computability in Analysis and in Physics. Perspective in Mathematical Logic* (Springer, Berlin, Heidelberg 1988)
- 34.53 E. Arnold: Experiments and simulations: Do they fuse? In: *Computer Simulations and the Changing Face of Scientific Experimentation*, ed. by J.M. Durán, E. Arnold (Cambridge Scholars Publishing, Newcastle upon Tyne 2013)
- 34.54 R. Trenholme: Analog simulation, *Philos. Sci.* **61**(1), 115–131 (1994)
- 34.55 P.K. Kundu, I.M. Cohen, H.H. Hu: *Fluid Mechanics*, 3rd edn. (Elsevier, Amsterdam 2004)
- 34.56 S.G. Sterrett: Models of machines and models of phenomena, *Int. Stud. Philos. Sci.* **20**, 69–80 (2006)
- 34.57 S.G. Sterrett: Similarity and dimensional analysis. In: *Philosophy of Technology and Engineering Sciences*, ed. by A. Meijers (Elsevier, Amsterdam 2009)
- 34.58 G.I. Barenblatt: *Scaling, Self-Similarity, and Intermediate Asymptotics*, Cambridge Texts in Applied Mathematics, Vol. 14 (Cambridge Univ. Press,

- Cambridge 1996)
- 34.59 R.W. Shonkwiler, L. Lefton: *An Introduction to Parallel and Vector Scientific Computing* (Cambridge Univ. Press, Cambridge 2006)
- 34.60 M.J. Borwein, R.E. Crandall: Closed forms: What they are and why we care, *Not. Am. Math. Soc.* **60**(1), 50 (2013)
- 34.61 B. Fillion, S. Bangu: Numerical methods, complexity, and epistemic hierarchies, *Philos. Sci.* **82**(5), 941–955 (2015)
- 34.62 N. Fillion, R.M. Corless: On the epistemological analysis of modeling and computational error in the mathematical sciences, *Synthese* **191**(7), 1451–1467 (2014)
- 34.63 R. Feynman: Simulating physics with computers, *Int. J. Theor. Phys.* **21**(6/7), 467–488 (1982)
- 34.64 T. Toffoli: Cellular automata as an alternative to (rather than an approximation of) differential equations in modeling physics, *Physica D* **10**, 117–127 (1984)
- 34.65 N. Margolus: Crystalline computation. In: *Feynman and Computation: Exploring the Limits of Computers*, ed. by A. Hey (Westview, Boulder 2002)
- 34.66 R. Hegselmann: Understanding social dynamics: The cellular automata approach. In: *Social Science Microsimulation*, ed. by K.G. Troitzsch, U. Mueller, G.N. Gilbert, J. Doran (Springer, London 1996) pp. 282–306
- 34.67 C.G. Langton: Studying artificial life with cellular automata, *Physica D* **22**, 120–149 (1986)
- 34.68 B. Hasslacher: Discrete Fluids, Los Alamos Sci. Special issue **15**, 175–217 (1987)
- 34.69 N. Metropolis, S. Ulam: The Monte Carlo method, *J. Am. Stat. Assoc.* **44**(247), 335–341 (1949)
- 34.70 P. Galison: Computer simulations and the trading zone. In: *The Disunity of Science: Boundaries, Contexts, and Power*, ed. by P. Galison, D. Stump (Stanford Univ. Press, Stanford 1996) pp. 118–157
- 34.71 P. Galison: *Image and Logic: A Material Culture of Microphysics* (Univ. Chicago Press, Chicago 1997)
- 34.72 C. Beisbart, J. Norton: Why Monte Carlo simulations are inferences and not experiments. In: *International Studies in Philosophy of Science*, Vol. 26, ed. by J.W. McAllister (Routledge, Abington 2012) pp. 403–422
- 34.73 S. Succi: *The Lattice Boltzmann Equation for Fluid Dynamics and Beyond* (Clarendon, Oxford 2001)
- 34.74 A.M. Bedau: Weak emergence, *Philos. Perspect.* **11**(11), 375–399 (1997)
- 34.75 T. Grüne-Yanoff: The explanatory potential of artificial societies, *Synthese* **169**(3), 539–555 (2009)
- 34.76 B. Epstein: Agent-based modeling and the fallacies of individualism. In: *Models, Simulations, and Representations*, ed. by P. Humphreys, C. Imbert (Routledge, London 2011) p. 115444
- 34.77 S.B. Pope: *Turbulent Flows* (Cambridge Univ. Press, Cambridge 2000)
- 34.78 P.N. Edwards: *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming* (MIT Press, Cambridge 2010)
- 34.79 M. Heymann: Understanding and misunderstanding computer simulation: The case of atmospheric and climate science – An introduction, *Stud. Hist. Philos. Sci. Part B* **41**(3), 193–200 (2010), Special Issue: Modelling and Simulation in the Atmospheric and Climate Sciences
- 34.80 E. Winsberg: Handshaking your way to the top: Inconsistency and falsification in intertheoretic reduction, *Philos. Sci.* **73**, 582–594 (2006)
- 34.81 P. Humphreys: Scientific knowledge. In: *Handbook of Epistemology*, ed. by I. Niiniluoto, M. Sintonen, J. Woleński (Springer, Dordrecht 2004)
- 34.82 W.S. Parker: Understanding pluralism in climate modeling, *Found. Sci.* **11**(4), 349–368 (2006)
- 34.83 W.S. Parker: Ensemble modeling, uncertainty and robust predictions, *Wiley Interdiscip. Rev.: Clim. Change* **4**(3), 213–223 (2013)
- 34.84 M. Sundberg: Cultures of simulations vs. cultures of calculations? The development of simulation practices in meteorology and astrophysics, *Stud. Hist. Philos. Sci. Part B* **41**, 273–281 (2010), Special Issue: Modelling and simulation in the atmospheric and climate sciences
- 34.85 M. Sundberg: The dynamics of coordinated comparisons: How simulationists in astrophysics, oceanography and meteorology create standards for results, *Soc. Stud. Sci.* **41**(1), 107–125 (2011)
- 34.86 E. Tal: From data to phenomena and back again: Computer-simulated signatures, *Synthese* **182**(1), 117–129 (2011)
- 34.87 R. El Skaf, C. Imbert: Unfolding in the empirical sciences: Experiments, thought experiments and computer simulations, *Synthese* **190**(16), 3451–3474 (2013)
- 34.88 L. Soler, S. Zwart, M. Lynch, V. Israel-Jost: *Science After the Practice Turn in the Philosophy, History, and Social Studies of Science* (Routledge, London 2014)
- 34.89 H. Chang: The philosophical grammar of scientific practice, *Int. Stud. Philos. Sci.* **25**(3), 205–221 (2011)
- 34.90 H. Chang: Epistemic activities and systems of practice: Units of analysis in philosophy of science after the practice turn. In: *Science After the Practice Turn in the Philosophy, History and Social Studies of Science*, ed. by L. Soler, S. Zwart, M. Lynch, V. Israel-Jost (Routledge, London 2014) pp. 67–79
- 34.91 A. Barberousse, S. Franceschelli, C. Imbert: Computer simulations as experiments, *Synthese* **169**(3), 557–574 (2009)
- 34.92 P. Grim, R. Rosenberger, A. Rosenfeld, B. Anderson, R.E. Eason: How simulations fail, *Synthese* **190**(12), 2367–2390 (2013)
- 34.93 J.H. Fetzer: Program verification: The very idea, *Commun. ACM* **31**(9), 1048–1063 (1988)
- 34.94 A. Asperti, H. Geuvers, R. Natarajan: Social processes, program verification and all that, *Math. Struct. Comput. Sci.* **19**(5), 877–896 (2009)
- 34.95 W.L. Oberkampff, C.J. Roy: *Verification and Validation in Scientific Computing* (Cambridge Univ. Press, Cambridge 2010)
- 34.96 W.S. Parker: Computer simulation. In: *The Routledge Companion to Philosophy of Science*, ed. by S. Psillos, M. Curd (Routledge, London 2013)

- 34.97 J. Lenhard: Computer simulation: The cooperation between experimenting and modeling, *Philos. Sci.* **74**(2), 176–194 (2007)
- 34.98 N. Oreskes, K. Shrader-Frechette, K. Belitz: Verification, validation, and confirmation of numerical models in the earth sciences, *Science* **263**(5147), 641–646 (1994)
- 34.99 J. Lenhard, E. Winsberg: Holism, entrenchment, and the future of climate model pluralism, *Stud. Hist. Philos. Sci.* **41**(3), 253–262 (2010)
- 34.100 A. Barberousse, C. Imbert: New mathematics for old physics: The case of lattice fluids, *Stud. Hist. Philos. Sci. Part B* **44**(3), 231–241 (2013)
- 34.101 J.M. Boumans: Understanding in economics: Gray-box models. In: *Scientific Understanding: Philosophical Perspectives*, ed. by H.W. de Regt, S. Leonelli, K. Eigner (Univ. Pittsburgh Press, Pittsburgh 2009)
- 34.102 C. Imbert: L'opacité intrinsèque de la nature: Théories connues, phénomènes difficiles à expliquer et limites de la science, Ph.D. Thesis (Atelier national de Reproduction des Thèses, Lille 2008), <http://www.theses.fr/2008PA010703>.
- 34.103 J. Hardwig: The role of trust in knowledge, *J. Philos.* **88**(12), 693–708 (1991)
- 34.104 H. Reichenbach: On probability and induction, *Philos. Sci.* **5**(1), 21–45 (1938), reprinted in S. Sarkar (Ed.) *Logic, Probability and Induction* (Garland, New York 1996)
- 34.105 A. Barberousse, C. Imbert: Recurring models and sensitivity to computational constraints, *The Monist* **97**(3), 259–279 (2014)
- 34.106 T. Kuhn: *The Structure of Scientific Revolutions*, 3rd edn. (The Univ. Chicago Press, Chicago 1996)
- 34.107 P. Kitcher: Explanatory unification and the causal structure of the world. In: *Scientific Explanation*, ed. by P. Kitcher, W. Salmon (Univ. Minnesota Press, Minneapolis 1989)
- 34.108 R. De Langhe: A unified model of the division of cognitive labor, *Philos. Sci.* **81**(3), 444–459 (2014)
- 34.109 A. Lyon: Why are normal distributions normal?, *Br. J. Philos. Sci.* (2013), doi:10.1093/bjps/axs046
- 34.110 R. Batterman: Why equilibrium statistical mechanics works: Universality and the renormalization group, *Philos. Sci.* **65**, 183–208 (1998)
- 34.111 R. Batterman: Multiple realizability and universality, *Br. J. Philos. Sci.* **51**, 115–145 (2000)
- 34.112 R. Batterman: Asymptotics and the role of minimal models, *Br. J. Philos. Sci.* **53**, 21–38 (2002)
- 34.113 E. Winsberg: A tale of two methods, *Synthese* **169**(3), 575–592 (2009)
- 34.114 S.D. Norton, F. Suppe: Why atmospheric modeling is good science. In: *Changing the Atmosphere: Expert Knowledge and Environmental Governance*, ed. by P. Edwards, C. Miller (MIT Press, Cambridge 2001)
- 34.115 C. Beisbart: How can computer simulations produce new knowledge?, *Eur. J. Philos. Sci.* **2**, 395–434 (2012)
- 34.116 E.A. Di Paolo, J. Noble, S. Bullock: Simulation models as opaque thought experiments, *Proc. 7th Int. Conf. Artif. Life*, ed. by K.A. Bedau, J.S. Mc-Caskill, N. Packard, S. Rasmussen (MIT Press, Cambridge 2000) pp. 497–506
- 34.117 S. Chandrasekharan, N.J. Nersessian, V. Subramanian: Computational modeling: Is this the end of thought experimenting in science? In: *Thought Experiments in Philosophy, Science and the Arts*, ed. by J. Brown, M. Frappier, L. Meynell (Routledge, London 2012) pp. 239–260
- 34.118 J.D. Norton: Are thought experiments just what you thought?, *Can. J. Philos.* **26**, 333–366 (1996)
- 34.119 J.D. Norton: On thought experiments: Is there more to the argument?, *Philos. Sci.* **71**, 1139–1151 (2004)
- 34.120 R. Descartes: Discours de la méthode. In: *Oeuvres de Descartes*, Vol. 6, ed. by C. Adam, P. Tannery (J. Vrin, Paris 1996), first published in 1637
- 34.121 P. Humphreys: What are data about? In: *Computer Simulations and the Changing Face of Experimentation*, ed. by E. Arnold, J. Durán (Cambridge Scholars Publishing, Cambridge 2013)
- 34.122 M. Stöckler: On modeling and simulations as instruments for the study of complex systems. In: *Science at Century's End: Philosophical Questions on the Progress and Limits of Science*, ed. by M. Carrier, G. Massey, L. Ruetsche (Univ. Pittsburgh Press, Pittsburgh 2000) pp. 355–373
- 34.123 P. Humphreys: Computational and conceptual emergence, *Philos. Sci.* **75**(5), 584–594 (2008)
- 34.124 P. Humphreys: The philosophical novelty of computer simulation methods, *Synthese* **169**(3), 615–626 (2008)
- 34.125 A. Barberousse, M. Vorms: Computer simulations and empirical data. In: *Computer Simulations and the Changing Face of Scientific Experimentation*, ed. by J.M. Durán, E. Arnold (Cambridge Scholars Publishing, Newcastle upon Tyne 2013)
- 34.126 J.A. Fodor: Special sciences (or: The disunity of science as a working hypothesis), *Synthese* **28**(2), 97–115 (1974)
- 34.127 M.S. Morgan: Experiments without material intervention: Model experiments, virtual experiments and virtually experiments. In: *The Philosophy of Scientific Experimentation*, ed. by R. Hans (Univ. Pittsburgh Press, Pittsburgh 2003) pp. 216–235
- 34.128 C. Hempel: *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science* (Free Press, New York 1965)
- 34.129 W. Salmon: *Scientific Explanation and the Causal Structure of the World* (Princeton Univ. Press, Princeton 1984)
- 34.130 W. Salmon: Causality without counterfactuals, *Philos. Sci.* **61**, 297–312 (1994)
- 34.131 P. Railton: Probability, explanation, information, *Synthese* **48**, 233–256 (1981)
- 34.132 P. Kitcher: *The Advancement of Science: Science Without Legend, Objectivity Without Illusions* (Oxford Univ. Press, New York 1993)
- 34.133 T. Grüne-Yanoff, P. Weirich: The philosophy and epistemology of simulation: A review, *Simul. Gaming* **41**(1), 20–50 (2010)

- 34.134 A. Ilachinski: *Cellular Automata: A Discrete Universe* (World Scientific, Singapore 2001)
- 34.135 E.F. Keller: Models, simulation and computer experiments. In: *The Philosophy of Scientific Experimentation*, ed. by H. Radder (Univ. Pittsburgh Press, Pittsburgh 2003) pp. 198–215
- 34.136 D. Dowling: Experimenting on theories, *Sci. Context* **12**(2), 261–273 (1999)
- 34.137 G. Piccinini: Computational explanation and mechanistic explanation of mind. In: *Cartographies of the Mind*, ed. by M. Marraffa, M. De Caro, F. Ferretti (Springer, Dordrecht 2007) pp. 23–36
- 34.138 E. Arnold: What's wrong with social simulations?, *The Monist* **97**(3), 359–377 (2014)
- 34.139 S. Ruphy: Limits to modeling: Balancing ambition and outcome in astrophysics and cosmology, *Simul. Gaming* **42**(2), 177–194 (2011)
- 34.140 B. Epstein, P. Forber: The perils of tweaking: How to use macrodata to set parameters in complex simulation models, *Synthese* **190**(2), 203–218 (2012)
- 34.141 W. Bechtel, R.C. Richardson: *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research* (MIT Press, Cambridge 1993)
- 34.142 H. Zwirn: *Les Systèmes complexes* (Odile Jacob, Paris 2006), in French
- 34.143 Y. Bar-Yam: *Dynamics of Complex Systems* (Westview, Boulder 1997)
- 34.144 R. Badii, A. Politi: *Complexity: Hierarchical Structures and Scaling in Physics* (Cambridge Univ. Press, Cambridge 1999)
- 34.145 D. Little: *Varieties of Social Explanation: An Introduction to the Philosophy of Social Science* (Westview, Boulder 1990)
- 34.146 H. Kincaid: *Philosophical Foundations of the Social Sciences: Analyzing Controversies in Social Research* (Cambridge Univ. Press, Cambridge 1996)
- 34.147 C. Hitchcock: Discussion: Salmon on explanatory relevance, *Philos. Sci.* **62**, 304–320 (1995)
- 34.148 C. Imbert: Relevance, not invariance, explanatory power, not manipulability: Discussion of Woodward's views on explanatory relevance, *Philos. Sci.* **80**(5), 625–636 (2013)
- 34.149 W.C. Salmon: *Four Decades of Scientific Explanation* (Univ. Pittsburgh Press, Pittsburgh 2006)
- 34.150 G. Schurz: Relevant deduction, *Erkenntnis* **35**(1–3), 391–437 (1991)
- 34.151 H.E. Kyburg: Comment, *Philos. Sci.* **32**, 147–151 (1965)
- 34.152 M. Scriven: Explanations, predictions, and laws. In: *Scientific Explanation, Space, and Time*, Vol. 3, ed. by H. Feigl, G. Maxwells (Univ. Minnesota Press, Minneapolis 1962) pp. 170–230
- 34.153 J. Woodward: Scientific explanation. In: *The Stanford Encyclopedia of Philosophy*, ed. by E.N. Zalta (Stanford Univ., Stanford 2014), <http://plato.stanford.edu/archives/win2014/entries/scientific-explanation/>
- 34.154 J. Woodward: *Making Things Happen* (Oxford Univ. Press, Oxford 2003)
- 34.155 S. Wolfram: *A New Kind of Science* (Wolfram Media, Champaign 2002)
- 34.156 H.W. de Regt, D. Dieks: A contextual approach to scientific understanding, *Synthese* **144**(1), 137–170 (2005)
- 34.157 R.P. Feynman, R.B. Leighton, M.L. Sands: *The Feynman Lectures on Physics*, Vol. 3 (Addison-Wesley, Reading 1963)
- 34.158 C. Hempel: Reasons and covering laws in historical explanation. In: *The Philosophy of C.G. Hempel: Studies in Science, Explanation, and Rationality*, ed. by J.H. Fetzer (Oxford Univ. Press, Oxford 2000), first published in 1963
- 34.159 J. Lenhard: Surprised by a nanowire: Simulation, control, and understanding, *Philos. Sci.* **73**(5), 605–616 (2006)
- 34.160 M. Bedau: Downward causation and the autonomy of weak emergence, *Principia* **6**, 5–50 (2003)
- 34.161 P. Huneman: Determinism, predictability and open-ended evolution: Lessons from computational emergence, *Synthese* **185**(2), 195–214 (2012)
- 34.162 C. Imbert: Why diachronically emergent properties must also be salient. In: *World Views, Science, and Us: Philosophy and Complexity*, ed. by C. Gershenson, D. Aerts, B. Edmonds (World Scientific, Singapore 2007) pp. 99–116
- 34.163 H. Zwirn, J.P. Delahaye: Unpredictability and computational irreducibility. In: *Irreducibility and Computational Equivalence*, Emergence, Complexity and Computation, Vol. 2, ed. by H. Zenil (Springer, Berlin, Heidelberg 2013) pp. 273–295
- 34.164 J. Kuorikoski: Simulation and the sense of understanding. In: *Models, Simulations, and Representations*, ed. by P. Humphreys, C. Imbert (Routledge, London 2012)
- 34.165 C.R. Shalizi, C. Moore: *What Is a Macrostate? Subjective Observations and Objective Dynamics* (2003) arxiv:cond-mat/0303625
- 34.166 N. Israeli, N. Goldenfeld: Computational irreducibility and the predictability of complex physical systems, *Phys. Rev. Lett.* **92**(7), 074105 (2004)
- 34.167 N. Goodman: *Language of Arts* (Hackett, Indianapolis 1976)
- 34.168 M. Vorms: Formats of representation in scientific theorizing. In: *Models, Simulations, and Representations*, (Routledge, London 2012) pp. 250–273
- 34.169 J. Jebeile: Explication et Compréhension Dans Les Sciences Empiriques. Les modèles Scientifiques et le Tournant Computational, Ph.D. Thesis (Université Paris, Paris 2013)
- 34.170 S. Bullock: Levins and the lure of artificial worlds, *The Monist* **97**(3), 301–320 (2014)
- 34.171 J. Lenhard: Autonomy and automation: Computational modeling, reduction, and explanation in quantum chemistry, *The Monist* **97**(3), 339–358 (2014)
- 34.172 K. Appel, W. Haken: Every planar map is four colorable. I. Discharging, III. *J. Math.* **21**(3), 429–490 (1977)
- 34.173 K. Appel, W. Haken, J. Koch: Every planar map is four colorable. II. Reducibility, III. *J. Math.* **21**(3),

- 491–567 (1977)
- 34.174 T. Tymoczek: *New Directions in the Philosophy of Mathematics: An Anthology* (Princeton Univ. Press, Princeton 1998)
- 34.175 I. Lakatos: *Proofs and Refutations* (Cambridge Univ. Press, Cambridge 1976)
- 34.176 H. Putnam: What is mathematical truth? In: *Mathematics, Matter and Method*, Vol. 1, (Cambridge Univ. Press, Cambridge 1975) pp. 60–78
- 34.177 F. Guala: Models, simulations, and experiments. In: *Model-Based Reasoning*, ed. by L. Magnani, N.J. Nersessian (Springer, New York 2002) pp. 59–74
- 34.178 M. Morrison: Models, measurement and computer simulation: The changing face of experimentation, *Philos. Stud.* **143**(1), 33–57 (2009)
- 34.179 R.N. Giere: Is computer simulation changing the face of experimentation?, *Philos. Stud.* **143**(1), 59–62 (2009)
- 34.180 D. Shapere: The concept of observation in science and philosophy, *Philos. Sci.* **49**(4), 485–525 (1982)
- 34.181 P. Humphreys: X-ray data and empirical content. Logic, methodology and philosophy of science, Proc. 14th Int. Congr. (Nancy), ed. by P. Schroeder-Heister, W. Hodges, G. Heinzmann, P.E. Bour (College Publications, London 2014) pp. 219–234
- 34.182 V. Israel-Jost: The impact of modern imaging techniques on the concept of observation: A philosophical analysis, Ph.D. Thesis (Université de Paris, Panthéon-Sorbonne 2011)
- 34.183 D. Resnik: Some recent challenges to openness and freedom in scientific publication. In: *Ethics for Life Scientists*, Vol. 5, (Springer, Dordrecht 2005) pp. 85–99
- 34.184 M. Frické: Big data and its epistemology, *J. Assoc. Inf. Sci. Technol.* **66**(4), 651–661 (2014)
- 34.185 S. Leonelli: What difference does quantity make? On the epistemology of big data in biology, *Big Data Soc.* (2014), doi:[10.1177/2053951714534395](https://doi.org/10.1177/2053951714534395)
- 34.186 W.S. Parker: Franklin, Holmes, and the epistemology of computer simulation, *Int. Stud. Philos. Sci.* **22**(2), 165–183 (2008)
- 34.187 W.S. Parker: Computer simulation through an error–statistical lens, *Synthese* **163**, 371–384 (2008)
- 34.188 D.G. Mayo: *Error and the Growth of Experimental Knowledge* (Univ. Chicago Press, Chicago 1996)
- 34.189 H.M. Collins: *Tacit and Explicit Knowledge* (Univ. Chicago Press, Chicago 2010)
- 34.190 L. Soler, E. Trizio, T. Nickles, W.C. Wimsatt: *Characterizing the Robustness of Science: After the Practice Turn in Philosophy of Science* (Springer, Dordrecht 2012)
- 34.191 A. Gelfert: Scientific models, simulation, and the experimenter's regress. In: *Representation, Models and Simulations*, ed. by P. Humphreys, C. Imbert (Routledge, London 2011) pp. 145–167
- 34.192 H.M. Collins: *Changing Order: Replication and Induction in Scientific Practice* (Sage, London 1985)
- 34.193 B. Godin, Y. Gingras: The experimenters' regress: From skepticism to argumentation, *Stud. Hist. Philos. Sci. Part A* **33**(1), 133–148 (2002)
- 34.194 A. Franklin: How to avoid the experimenters regress, *Stud. Hist. Philos. Sci.* **25**, 97–121 (1994)
- 34.195 E. Winsberg: Computer simulations in science. In: *The Stanford Encyclopedia of Philosophy*, ed. by E.N. Zalta (Stanford Univ., Stanford 2014), <http://plato.stanford.edu/archives/fall2014/entries/simulations-science/>
- 34.196 J.M. Durán: The use of the materiality argument in the literature for computer simulations. In: *Computer Simulations and the Changing Face of Scientific Experimentation*, ed. by J.M. Durán, E. Arnold (Cambridge Scholars, Newcastle upon Tyne 2013)
- 34.197 B. Mundy: On the general theory of meaningful representation, *Synthese* **67**, 391–437 (1986)
- 34.198 O. Bueno: Empirical adequacy: A partial structures approach, *Stud. Hist. Philos. Sci.* **28**, 585–610 (1997)
- 34.199 W.S. Parker: Does matter really matter? Computer simulations, experiments, and materiality, *Synthese* **169**(3), 483–496 (2009)
- 34.200 I. Peschard: Computer simulation as substitute for experimentation?. In: *Simulations and Networks*, ed. by S. Vaienti (Hermann, Paris) forthcoming [http://philsci-archiv.pitt.edu/9010/1/ls\\_simulation\\_an\\_epistemic\\_substitute.pdf](http://philsci-archiv.pitt.edu/9010/1/ls_simulation_an_epistemic_substitute.pdf)
- 34.201 E.C. Parke: Experiments, simulations, and epistemic privilege, *Philos. Sci.* **81**(4), 516–536 (2014)
- 34.202 M.S. Morgan: Experiments versus models: New phenomena, inference and surprise, *J. Econ. Methodol.* **12**(2), 317–329 (2005)
- 34.203 S. Roush: The epistemic superiority of experiment to simulation, Proc. PSA 2014 Conf., Chicago, to be published
- 34.204 S.L. Peck: Simulation as experiment: A philosophical reassessment for biological modeling, *Trends in Ecol. Evol.* **19**(10), 530–534 (2004)
- 34.205 R. Harré: The materiality of instruments in a metaphysics for experiments. In: *The Philosophy of Scientific Experimentation*, ed. by H. Radder (Pittsburg Univ. Press, Pittsburg 2003) pp. 19–38
- 34.206 M.S. Morgan: Model experiments and models in experiments. In: *Model-Based Reasoning: Science, Technology, Values*, ed. by M. Lorenzo, N.J. Nersessian (Springer, New York 2001)
- 34.207 J.M. Durán: A brief overview of the philosophical study of computer simulations, *Am. Philos. Assoc. Newsl. Philos. Comput.* **13**(1), 38–46 (2013)
- 34.208 T. Boyer-Kassem: Layers of models in computer simulations, *Int. Stud. Philos. Sci.* **28**(4), 417–436 (2014)
- 34.209 R.I.G. Hughes: *The Theoretical Practices of Physics: Philosophical Essays* (Oxford Univ. Press, Oxford 2010)
- 34.210 O. Bueno: Computer simulations: An inferential conception, *The Monist* **97**(3), 378–398 (2014)
- 34.211 M. Weisberg: *Simulation and Similarity: Using Models to Understand the World* (Oxford Univ. Press, Oxford 2013)
- 34.212 R. Batterman: *The Devil in the Details, Asymptotic Reasoning in Explanation, Reduction, and Emer-*

- gence (Oxford Univ. Press, Oxford 2002)
- 34.213 E. Winsberg: *Science in the Age of Computer Simulation* (Univ. Chicago Press, Chicago 2010)
- 34.214 A.I. Janis: Can thought experiments fail? In: *Thought Experiments in Science and Philosophy*, ed. by T. Horowitz, G. Massey (Rowman Littlefield, Lanham 1991) pp. 113–118
- 34.215 J.R. Searle: *The Construction of Social Reality* (Free Press, London 1996)
- 34.216 G. Piccinini: Computation in physical systems. In: *The Stanford Encyclopedia of Philosophy*, ed. by E.N. Zalta (Fall 2012 Edition) <http://plato.stanford.edu/archives/fall2012/entries/computation-physicssystem/>
- 34.217 K. Zuse: The computing universe, *Int. J. Theor. Phys.* **21**, 589–600 (1982)
- 34.218 E. Fredkin: Digital mechanics: An informational process based on reversible universal cellular automata, *Physica D* **45**, 1–3 (1990)
- 34.219 R.N. Giere: How models are used to represent reality, *Philos. Sci.* **71**, 742–752 (2004)
- 34.220 U. Mäki: Models and the locus of their truth, *Synthese* **180**(1), 47–63 (2011)
- 34.221 R. Giere: An agent-based conception of models and scientific representation, *Synthese* **172**(2), 269–281 (2010)



# Simulation

## 35. Simulation of Complex Systems

Paul Davidsson, Franziska Klügl, Harko Verhagen

Understanding and managing complex systems has become one of the biggest challenges for research, policy and industry. Modeling and simulation of complex systems promises to enable us to understand how a human nervous system and brain not just maintain the activities of a metabolism, but enable the production of intelligent behavior, how huge ecosystems adapt to changes, or what actually influences climatic changes. Also man-made systems are getting more complex and difficult, or even impossible, to grasp. Therefore we need methods and tools that can help us in, for example, estimating how different infrastructure investments will affect the transport system and understanding the behavior of large Internet-based systems in different situations. This type of system is becoming the focus of research and sustainable management as there are now techniques, tools and the computational resources available. This chapter discusses modeling and simulation of such complex systems. We will start by discussing what characterizes complex systems.

35.1	<b>Complex Systems</b> .....	783
35.1.1	Features Associated with Complex Systems.....	784
35.1.2	Summing Up.....	785
35.2	<b>Modeling Complex Systems</b> .....	785
35.2.1	Macro-Level Versus Micro-Level Simulation.....	786
35.2.2	Purpose of Modeling Complex Systems.....	788
35.3	<b>Agent-Based Simulation of Complex Systems</b> .....	789
35.3.1	Elements of Agent-Based Simulation Models.....	790
35.3.2	Engineering Agent-Based Simulations.....	792
35.4	<b>Summing Up and Future Trends</b> .....	795
	<b>References</b> .....	796

### 35.1 Complex Systems

What makes a system *complex* can be formulated from various points of view. In daily speech, *complex* may be primarily a relative notion depending on what a human observer may experience as not so easy to grasp or control. A formal definition capturing what a complex system is, with the necessary general validity, turns out to be difficult to formulate. There is currently no generally accepted definition of the term *complex system* [35.1]. Since there are many sources of complexity, very different systems can be described as complex ones. There are many different notions of *complexity*. In computer science complexity has two major facets: information complexity (how large is the minimal description of the information that fully capture it?) or computation complexity (how much time [space] does an algorithm need for solving a problem in relation to the problem size?). Yet, these technical terms

do not match the intuition of a complex system being difficult to understand, verify or control. Ladyman et al. [35.1] collect various definitions from different domains. Those definitions focus on structures, on the effect of initial conditions, self organization and informal *complicatedness*. Based on the analysis of these definitions, they discuss a list of features that we will tackle in the next section as they form the core properties of complex systems. None of them alone forms a definitory aspect, but combinations give an important impression of what a complex system might be.

Over the years there have been a number of approaches aimed at capturing complex systems. Originating in physics, the idea of self-organized criticality [35.2] was central for the analysis of complexity. *Critical attractor* and *nonequilibrium system* are the

central terms in the related vocabulary. A second related approach is termed *complex adaptive systems*, originating more from biology, based on the idea that complexity arises from many entities adapting to their local environment.

### 35.1.1 Features Associated with Complex Systems

Following the discussion above, an intuition of what is a *complex system* can be best imparted by describing the features of complex systems. None of the features alone capture a complex system and a complex system does not need to exhibit all of them. A complex system in physics may have a completely different shape compared to a complex system in biology or social science. Nevertheless, a combination of those features can be identified at any complex system that we will consider in the following.

#### Nonlinearity

Nonlinearity is one of the most frequently identified properties of a complex system. It basically describes that when combining results/solutions/behavior from multiple elements into one system, these do not add in a linear way, but rather in a nonlinear way due to interactions between elements. Examples of nonlinear systems are systems in which saturation occurs. Prominent are also systems in which a small change causes a big effect somewhere else. Weather, economies, social systems form examples for this kind of *chaotic* system.

Nonlinear phenomena are hard to model, capture and analyze mathematically. Often simulation is a last resort for handling the complexity, yet not each simulation paradigm is suitable.

#### Distributedness, Scale and Interaction

Many systems coined as complex ones, especially in biology or social science, are large and distributed. That means they contain a huge number of entities that are distributed in some way. This may not just refer to geographic distribution, but also to an entity taking a position in a network of relations. The important idea is that there is a form of local interaction, an entity or component interacts with a (selected) number of others distributed over some form of more or less abstract environment. Scale also plays an important role simply because a small system can be overlooked, in a huge one perceiving what happens where in that distributed system obscures the overall dynamics, or as *Auyang* [35.3, p. 11] coins it:

“The relational network is chiefly responsible for the great variety of phenomena in the world. It is

also responsible for the difficulty in studying large composite systems.”

Clearly, the origins of complexity are not merely in the number of entities capable of participating in an interaction, but there is some form of trade-off between the complexity of interaction and the number of interacting entities. Clearly the interaction of hundreds of millions of humans creates a complex system of epidemic spread [35.4], but there are also examples for systems with only a few entities, but complex interactions; for example the model describing the *emergence of political actors* [35.5] contains only 10 entities deciding about whether to pay tribute or not based on decisions taken before. Although there are only a few entities, their decision making and interaction is based on several feedback loops resulting in a complex system.

#### Multiple Levels of Observation, Self Organization, Emergence

*Auyang* [35.3] assumes that complexity arises from large-scale composition. The idea of a system composed of many interacting entities may lead to complexity even if the interaction does not lead to nonlinearity. In such a system, there are at least two levels of observation: the individual entity and the overall system level. The behavior and pattern observable on the latter originate from actions and interactions among the lower level entities, between the entities and their environment as well as if there is some manifestation on a high level (some form of *organization*) this may also impact on the lower level entities.

Two additional concepts are relevant in relation to the multilevel feature of complex systems: self organization and emergence. Self organization denotes some process in which local interactions of lower-level entities produce some form of sustainable regularity from an initially unordered situation. Often random fluctuations are responsible for locating the produced regularity. In his famous book *Kauffman* [35.6] discusses self organization in biological systems based on (evolutionary) adaptation, clearly distinguishing true self organization from organization following some predefined scheme expressing some intention to build structure. A key to self organization is that local decision making is adaptive to a changed environment; that means the local behavior is not fully predefined but at least conditional to environmental conditions. Prominent examples for self-organizing systems can be found in many natural systems, percolation processes in physics, genesis of structures in biology, bird flocking, etc. The principles are also used in technology, mainly in systems such as swarm robotics or biologically inspired optimization.

A related concept, capturing phenomena or patterns originating from local interaction on a lower level is emergence. In addition to self organization, it contains some associated aspect of *astonishment*. As this is often very subjective, the actual definition of the term *emergence* is partially controversial. *The whole is more than the sum of its parts* forms the intuitive characterization. This idea fascinated researchers from Aristotle to Holland [35.7]. More formal approaches for defining emergence are based on the idea that the description of the overall phenomenon uses different vocabulary than the description of the description of the lower-level entities producing the phenomenon. Well-known examples are the Mexican wave in cheering audiences in which the individual spectators stand up depending on their neighbors' actions so that a wave forms, or traffic jams originating from too high density of vehicles. The congestion travels into the opposite direction than the individual vehicles. *Darley* [35.8] discusses the dilemma of relating the definition of emergence to some limited understanding by an observer and defines emergence as a phenomenon that is best predicted by simulation.

#### Adaptivity, Flexible Decision Making and Feedback Loops

The features tackled so far mostly address the system level, but there is an important aspect also in the individual entities' behavior making up the complex system: For producing a complex system, decision making and interaction of the lower-level entities requires some degree of freedom. The entities must be able to adapt

their decision making to their local context. Only if the entities can adapt, feedback loops can be formulated and something complex (unexpected) can be produced. Flexible decision making refers to the next action the entity chooses to perform possibly being related to the extent with which an entity does something or referring to the selection of the interaction partner. Adaptivity hereby means a change in the behavior in reaction to some immediate environmental context, for example governed by some rules that control an entity to change its movement direction before bumping into a suddenly occurring obstacle. True learning is different from adaptive behavior: it changes the behavior program itself – for example, the entity learns a new rule about how to deal with obstacles. Learning can also happen on the system (population) level in the form of evolution combining the production of new entities with a fitness-based selection and/or survival.

### 35.1.2 Summing Up

The last section attempted to capture the idea of a *complex system* based on particular features. The main assumption hereby is a complex system originates from multilevel systems in which phenomena and patterns on the overall system level are produced by entities that are capable of flexible behavior and interaction.

Complex systems – as characterized here – are best analyzed and their overall behavior is best predicted using simulation. In the following, we will focus on different approaches and different motivations to modeling and simulating complex systems.

## 35.2 Modeling Complex Systems

Already in the early days of computer development, modeling and simulation were used in different research areas to predict the behavior of complex systems. Such models were typically based on differential equations and focused on describing phenomena on the overall system level. For instance, models of predator-prey populations could fairly accurately reproduce empirical data, but were limited in the sense that the models did not capture the actual low-level entity behavior and decision making, as well as the interaction between entities, but were based on the assumption that all low-level units were homogeneous. The development of individual-based modeling offers a possible solution to this problem with its (seemingly) natural mapping onto interacting, individual entities with incomplete information and capabilities, leading to models without global control, with decentralized data, asynchronous

computing, and inclusion of heterogeneities. These models also offer the possibility of studying the dynamics of the interaction processes instead of focusing on the (static) results of these processes [35.9, 10].

The main task of computer simulation is the creation and execution of a formal model of the behavior of the system being simulated. *Formal* means here that the model is represented using a formal language with so clearly defined syntax and semantics that the model can be executed using a computer. In scientific research, computer simulation forms a research methodology that can be contrasted to empirically driven research. As such, simulation belongs to the same family of research as analytical models. One way of formally modeling a system is to use a mathematical model, and then attempt to find analytical solutions enabling the prediction of the behavior of the system from a set of

parameters and initial conditions. Computer simulation, on the other hand, is often used when simple closed-form analytic solutions are not possible, which is typically the case for complex systems.

Computer simulation consists of three main steps:

1. Designing a model of an actual or theoretical system
2. Executing the model on a computer, and
3. Analyzing the execution output.

Although there are many different types of computer simulation, they typically attempt to generate a sample of representative scenarios for a model in which a complete enumeration of all possible states would be prohibitive or impossible. In this chapter we will describe simulation as a tool for understanding the behavior of complex systems. For a deeper philosophical treatment of the concept of simulation, we refer to Chap. 34 of this book by *Imbert*.

### 35.2.1 Macro-Level Versus Micro-Level Simulation

As indicated above, we can identify two main approaches to modeling of complex systems:

- Macro-level (or equation-based) simulation, which is typically based on mathematical differential or difference models. It views the overall system with its set of (the population) individual entities as one structure that can be characterized by a number of variables.
- Micro-level simulation, in which the behaviors and decision making of the specific individual entities are explicitly modeled. In contrast to macro-level simulation, it views the overall system behavior as emerging from the interactions between individuals. Thus only with micro-level simulation, the idea that complex effects need not have complex causes, can be explored. Micro-level simulation is sometimes referred to as multi-agent-based simulation (MABS), or agent-based simulation.

As argued by *Parunak* et al. [35.11], micro-level simulation is most appropriate for domains characterized by a high degree of localization and distribution and dominated by discrete decisions. Equation-based modeling, on the other hand, is most naturally applied to systems that can be modeled centrally and in which the dynamics are dominated by physical laws rather than information processing. Clearly, in many cases an abstract view onto aggregate dynamics as provided by macro-level simulation may be sufficient. However, for simulating complex systems micro-level approaches are more relevant. In the following sections we will

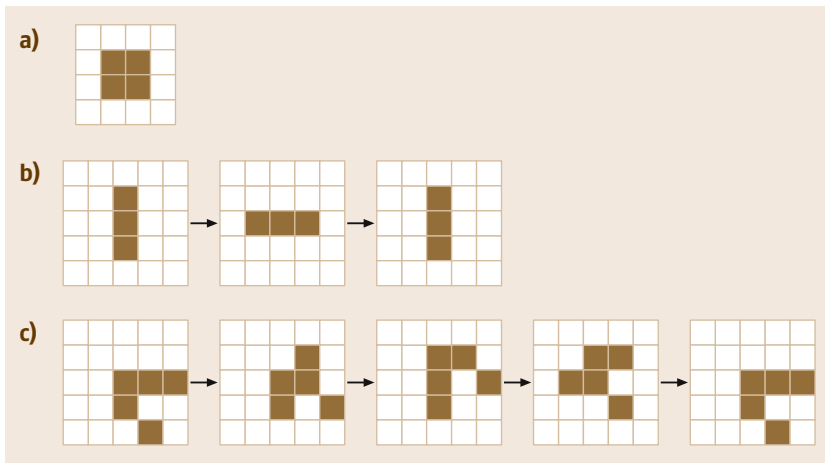
describe common ways of implementing micro-level simulations.

#### Cellular Automata

Individual-based modeling can be traced back to von Neumann, who in the 1950s invented what was later termed *cellular automata* [35.12]. The core definition of a cellular automaton refers to a model of a system that is discrete in time, space and state, yet the term is often used in a broader way denoting to models with discrete space (cells). In principle, such a simulation model consists of a grid of cells, i.e., entities, each in one of a finite number of states. The state of a cell at time  $t$  is a function of the states of a finite number of cells (called its neighborhood) at time  $t - 1$ . These neighbors are a selection of cells relative to the specified cell, and do not change. Every cell has the same rule for updating, based on the values in this neighborhood. Each time the rules are applied to the whole grid a new generation is created. These were used by *Conway* [35.13] in the 1970s when he constructed the well-known *Game of Life*. It is based on very simple rules determining the life and death of the cells in a virtual world in the form of a two-dimensional (2-D) grid just based on the state of a few neighbors, but was able to produce astonishing macro-level dynamics as shown in Fig. 35.1.

Inspired by this work, researchers developed more refined models, often modeling the social behavior of groups of animals or artificial creatures based on local interactions. With respect to human societies, *Epstein* and *Axtell* [35.14] developed in the 1990s one of the first agent-based models, called *Sugarscape*, to explore the role of social phenomena such as seasonal migrations, pollution, sexual reproduction, combat, and transmission of disease. This work is in spirit closely related to one of the best-known and earliest examples of the use of simulation in social science, namely, the *Schelling* model [35.15], in which cellular automata were used to simulate the emergence of segregation patterns in neighborhoods based on a few simple rules expressing the preferences of the agents. Another pioneer from the 1950s worth mentioning is *Barricelli* [35.16], who to some extent used agent-based modeling for simulating biological systems.

The cellular automata models closely resemble the models used in statistical physics, which has inspired physicists to include the simulation of social phenomena in large-scale social systems in their research agenda. In this area, sometimes referred to as sociophysics, phenomena such as opinion spreading in a society and competition between languages have been studied. These models originally described the behavior of atoms and molecules, which are quite simple objects,



**Fig. 35.1a–c** Macro-level phenomena produced by purely local rules determining the state (white = dead, alive = dark) of a cell based on the neighbors' state in the Game of Life

and the macro-level phenomena caused by their interaction (rather than by complex behavior of the individual as in the case of humans). Thus, in these models little attention is paid to individual variation and the individual decision making is rather primitively modeled. A prominent example of sociophysics is the work of *Galam* [35.17].

Another prominent example of a model in which local interactions lead to interesting macro-level behavior is the Boid model by *Reynolds* [35.18], which simulates coordinated animal motion such as bird flocks and fish schools. The underlying spatial representation and consequently the state of the entities in form of direction of movement are continuous values. So it is not a cellular automata in the narrow sense, yet belongs to the same category of models that generate complex system behavior from locally interacting, simple entities.

### Dynamic Micro-Simulation

One of the first, and most simple, ways of performing micro-level simulation in social science is often called *dynamic micro-simulation* [35.19, 20]. It is used to simulate the effect of the passing of time on individuals. Data from a (preferably large) random sample from the population to be simulated is used to initially characterize the simulated individuals. Some examples of sampled features are: age, sex, employment status, income, and health status. A set of transition probabilities are used to describe how these features will change over a given time period, e.g., there is a probability that an employed person will become unemployed over the course of a year. The transition probabilities are applied to the population for each individual in turn and then repeatedly reapplied for a number of simulated time periods. Sometimes it is necessary to also model changes in the population, for example birth, death, and

marriage. This type of simulation can be used to, for example, predict the outcome of different social policies. However, the quality of such simulations depends on the quality of:

- The random sample, which must be representative, and
- The transition probabilities, which must be valid and complete.

### Micro-Level Simulation in Technical Domains

Also for socio-technical systems, that means for systems in which people and technology interact, a number of micro-level modeling and simulation techniques exist. Examples are complex production lines in which human workers cooperate with machines executing more or less automated process steps. Formulating a complex system using for example a Petri nets and queuing system is based on a process-oriented point of view of locally interacting entities and is particularly apt for entities traveling through a complex system in a more or less automated way not involving individual reasoning and on-the-fly adaptation.

Petri nets hereby have been accepted as a powerful formal modeling tool for dealing with performance and functionality issues in systems with distributed concurrent processes based on local behavior – for example complex software systems (see as an introduction [35.21]). According to its basic definition, the core of a Petri net (of the type condition-event net) model is a bipartite graph consisting of a finite set of places connected with elements from a finite set of transitions. Places can hold tokens (one or more tokens with or without colors) that *travel* from place to place when a transition *fires*. The colors of the token represent its internal state and allow formulating behavior depending on internal state.

Queuing systems are used for performance analysis in distributed systems in which entities are traveling through a system, for example patients through a hospital or jobs through a production system (see for introduction: [35.22]). Also queuing systems models are graphs with two different types of nodes: servers and queues. The servers represent resources that the jobs have to be processed by. If the server is busy, the job has to wait in the queue in front of that resource. Every queue may have its special queuing discipline. Connections between the different elements are either deterministic (without branching) or probabilistic (branching or merging).

These micro-level models and simulations are used for complex socio-technical systems – with a focus on technology – that benefit from a distributed, process-oriented point of view. In contrast to cellular automata, space is abstracted to times that a token or job needs for traveling between places or servers. Simulated time is usually handled event-based, that means simulated time is advanced to times in which change happens that triggers other changes. A general formalism for object-oriented modeling and simulation is Discrete Event System Specification (DEVS) [35.23], which is possible to use for micro-level modeling and simulation for complex systems. Based on the concept of eventually coupled *systems*, the underlying abstractions are quite generic and thus can be used without doubt, but they are not specifically supportive for modeling complex systems. Below, we will introduce agent-based simulation – a micro-level modeling approach for complex systems that is more general than these more formalized modeling and simulation paradigms.

### 35.2.2 Purpose of Modeling Complex Systems

Modeling and simulation of complex systems can be done for different purposes, such as:

- Supporting theory building and evaluation
- Supporting the engineering of systems, e.g., validation, testing, etc.
- Supporting planning, policy making, and other decision making and
- Training, in order to improve a person's skills in a certain domain.

It is possible to distinguish between four types of end users: *scientists*, who use the models in the research process to gain new knowledge or verify hypotheses; *policymakers*, who use it for making strategic decisions; *managers* (of systems), who use it to make operational decisions; and *other professionals*, such as architects, who use it in their daily work. Below we describe how

these types of end users may use modeling and simulation of complex systems for different purposes.

#### Supporting Theory Building and Evaluation

In the context of theory building, a simulation model can be seen as an experimental method or as a theory in itself [35.24]. In the former case, simulations are run to test the predictions of theories, whereas in the latter case simulations in themselves are formal models of theories. Using a formal language for describing ambiguous, natural language-based theories helps to find inconsistencies and other problems, and thus contributes to theory building.

Simulation may also be used to evaluate a particular theory, model, hypothesis, or system, or compare two or more of these. Moreover, simulation can be used to verify whether a theory, model, hypothesis, system, or software is correct. Using simulation studies as an experimental tool offers great possibilities. For instance, many experiments with human societies are either unethical or even impossible to conduct. Experiments in silico, on the other hand, are fully possible. These can also breathe new life into the ever-present debate in sociology on the micro-macro link [35.25].

#### Supporting the Engineering of Systems

Many large-scale technical systems are distributed and involve complex interactions between humans and machines. The idea is to model the behavior of human users in terms of software entities (see next section). In particular, this seems useful in situations where it is too expensive, difficult, inconvenient, tiresome, or even impossible for real human users to test out a new technical system. Of course, also the technical system, or parts thereof, may be simulated. For instance, if the technical system includes hardware that is expensive and/or special purpose, it is natural to simulate also this part of the system when testing out the control software. An example of such a case is the testing of control systems for *intelligent buildings*, where software entities simulate the behavior of the people in the building [35.26].

#### Supporting Planning, Policy Making, and Other Decision Making

In simulation for decision making the focus is on exploring different possible future scenarios in order to choose between alternative actions. Besides this type of prediction, modeling of complex systems may be used for analysis; to gain deeper knowledge and understanding of a certain phenomenon.

An area in which several studies of this kind have been carried out is disaster management, such as experiments concerning different roles and the efficiency of reactions to emergencies [35.27]. Here also software

entities are models of humans. Based on individuals' observations, personal characteristics and skills, past experience and role characteristics, and social network, the entities create a plan to execute. The effect of adding a role (floor warden) in a fire alarm scenario upon the evacuation efficiency in an abstract environment is analyzed. Sometimes environmental information is based on GIS (geographical information system) data, thereby tying the simulation closer to the physical reality [35.24]. In yet another study, real-world data were used for both the environment and the software entities' internal decision-making model to analyze the effect of different insurance policies on the willingness of modeled humans to pay for a disaster insurance policy [35.28].

Another application area for this type of simulation study is disease spreading. Again, software entities are used to represent human beings and the simulation model is linked to real-world geographical data. One study [35.29] also included software entities that represent towns acting as the epicenter of disease outbreak. The town entity's behavior repertoire consisted of dif-

ferent containment strategies. The simulation model can be quickly adapted to local circumstances via the geographical data (given that there is data on the population as well) and is used to determine the effects of different containment strategies.

A third area where social simulation has been used to support planning and policy making is traffic and transport. An example of this is the simulation of all car travel in Switzerland during morning peak traffic [35.30].

### Training

The main advantage of using modeling and simulation for training purposes is to be part of a real-world-like situation without real-world consequences. Especially in the military the use of simulation for training purposes is widespread. Also in medicine, where mistakes can be very expensive in terms of money and lives, the use of simulation in education is on the rise.

An early product in this area was a tool to help train police officers to manage large public gatherings such as crowds, demonstrations, and marches [35.31].

## 35.3 Agent-Based Simulation of Complex Systems

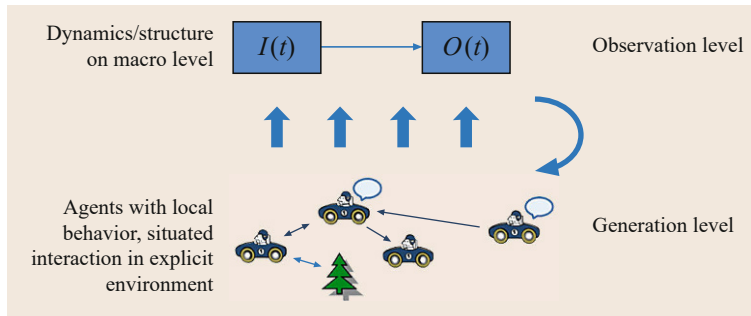
A special case of micro-level modeling is agent-based modeling. Multi-agent systems [35.32] consist of interacting entities embedded into a shared environment are coined as *agents*. Agents are characterized by some kind of agency and autonomy. There are many entities that can be referred to as *agents*, see [35.33] for a famous example describing a thermostat as an agent. Less controversial examples of agents are autonomous robots and human beings. Agency contains hereby notions of situatedness – being embedded into a (local) environment that is accessible to the agent by sensors and manipulatable using sensors. A second important aspect of agency relates to the agents' social abilities. This means that the agent is capable of interacting with other agents, taking part in conversations, coordinating activities, etc. Autonomy is the most difficult aspect as it relates to different aspects ranging from the agent is capable of executing a particular sequence of actions without human interventions to the agent learns and adapts its behavior without external control.

Agent-based models mostly focus on the emergence of macro-level properties from the local interaction of adaptive agents that influence one another [35.9, 34]. However, simulations in computational organization theory [35.35, 36], for example, often try to analyze the influence of macro-level phenomena on individuals. Given that complex systems are characterized by com-

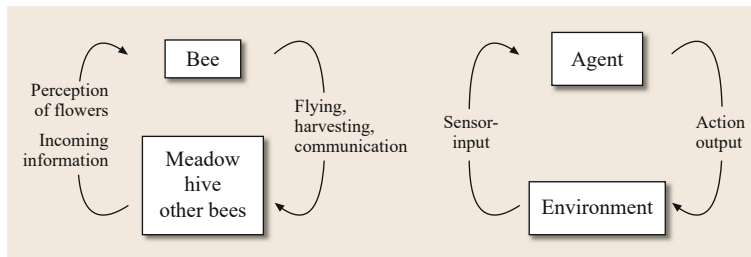
plexity in the interaction of entities that in themselves are and behave in an understandable way and that agents are the computational implementation of such entities, it is clear that agent-based modeling is a suitable paradigm for implementing and studying complex systems. It is supported by their generative nature [35.36]. Figure 35.2 illustrates the general principle.

There are mainly two reasons for the increasing popularity of agent-based simulation: the intuitiveness and the flexibility of the paradigm. The ontological correspondence between agents and original actors facilitates understanding of the model. The unit of description is the active entity in the model: real pedestrians are mapped to agents in crowd simulation or households in demographic models. Figure 35.3 illustrates this using the example of a bee and the general concept of an agent.

The second major advantage is the freedom of design. Agent-based models can contain heterogeneous entities, heterogeneous spatial environments or arbitrary complex agent decision making. This flexibility basically allows remodeling of any other micro-level model using an agent-based simulation. In other simulation paradigms, heterogeneity of entities or space, structural variation (entities leaving or entering the system) or context-dependent and flexible individual decision making are hard to achieve in a direct way.



**Fig. 35.2** The principle of agent-based simulation



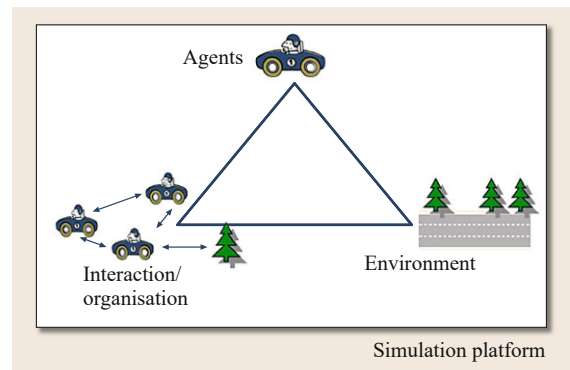
**Fig. 35.3** Correspondence of concepts between real-world actors (such as bees) and agents

### 35.3.1 Elements of Agent-Based Simulation Models

Let us now focus on how the system to be simulated is modeled. This depends on the type of system that should be simulated and for what purpose the simulation is done. A model of a system consists of a set of entities and an environment (in which the entities are situated). The entities are either individuals (agents) that have some decision-making capabilities, or objects (resources) that have no agency and are purely physical. There are a number of characteristics that can be used to differentiate between different types of models. We will first look at how individuals are being modeled, then on the interaction between the individuals, and finally how the environment is being modeled. Figure 35.4 summarizes these three basic aspects of agent-based simulation. In the following, the different elements of an agent-based simulation model are discussed. Section 35.3.2 deals with the engineering aspects also manifested in the simulation infrastructure used.

#### Model of an Individual

A model of an individual could range from being very simple, such a one binary variable (e.g., alive or dead) that is changed using only a single rule, to being very complex. The complexity of the model for a given simulation should be determined by the complexity of the individuals being simulated. Note that in complex systems very complex collective behavior could be achieved from very simple individual models, if the number is sufficiently large.



**Fig. 35.4** Illustration of the three elements of an agent-based model. A fourth relevant part is the simulation platform that will be discussed in Sect. 35.3.2, *More Issues in Engineering Agent-Based Simulations*

We can distinguish between modeling of the state of an individual and the behavior of the individual, i.e., the decisions and actions it takes. The state of an individual, in turn, can be divided into the physical and the mental state. The description of the physical state may include, for example, the position of the individual, and if it is a person being simulated features such as age, sex, and health status. The physical state is typically modeled as a feature vector, i.e., a list of attribute and value pairs. However, this is not always the case as in some domain the physical state of individual is not modeled at all, for example the *Principles of Synthetic Intelligence* (PSI) agent by Künzel and Hämmer [35.37] that was used to give stu-



dents theoretical insights in the area of psychological theory.

Whereas the physical state often is simple to model, the modeling of the mental state is typically much more complex, especially if the individuals modeled are human beings. A common approach is to model the beliefs, desires, and intentions of the individual, for instance by using the belief–desire–intention (BDI) model [35.38, 39]. Such a model may include the social state of the individual, i. e., which norms it adheres to, which coalitions it belongs to, etc. Although the BDI model is not based on any experimental evidence of human cognition it has proved to be quite useful in many applications. There has also been some work on incorporating emotions in models of the mental state of individuals [35.40] as well as obligations (cf. the beliefs, obligations, intentions and desires (BOID) model [35.41] which extends the BDI with obligations).

The modeling of the behaviors (and decisions) of the individuals can be done in a variety of ways, from simple probabilities to sophisticated reasoning and planning mechanisms. As an example where the former is used, we can mention dynamic micro-simulation as described in Sect. 35.2.1, *Dynamic Micro-Simulation*, which was one of the first ways of performing individual-based simulation and is still frequently used. In traditional micro-simulation, the behavior of each individual is regarded as a *black box*. The behavior is modeled in terms of probabilities and no attempt is made to justify these in terms of individual preferences, decisions, plans, etc. Thus, better results may be gained if also the cognitive processes of the individuals were simulated. Opening the black box of individual decision making can be done in several ways. A basic and common approach is to use decision rules, for instance, in the form of a set of situation-action rules. That is, if an individual and/or the environment is in state X then the individual will perform action Y. By combining decision rules and the BDI model quite sophisticated behavior can be modeled. Other models of individual cognition used in agent-based social simulation include the use of Soar (a computer implementation of Allen Newell’s unified theory of cognition [35.42]), which was for example used in Steve for generating a believable simulated tutor [35.43]. Another unified theory of individual cognition for which a computer implementation exists is Adaptive Control of Thought-Rational (ACT-R) [35.44], which is realized as a production system. A less general example is the Consumat model [35.45], a meta-model combining several psychological theories on decision making in a consumer situation. Also, nonsymbolic approaches such as neural networks have been used to model the agents’ decision making [35.27].

As we have seen, the behavior of individual could be either deterministic or stochastic. Also the basis for the behavior of the individuals may vary. We can identify the following categories:

- The state of the individual itself: In most social simulation models the physical and/or mental state of an individual plays an important role in determining its behavior.
- The state of the environment: Also the state of the environment surrounding the individual often influences the behavior of an individual. Thus, an individual may act differently in different contexts although its physical and mental state is the same.
- The state of other individuals: One popular type of simulation where the behaviors of individuals are based on the state of other individuals is those using cellular automata. As introduced above, the state of each cell is updated as a function of the states of a particular set of neighbors. In this case, information about the state of other individuals can be seen as gained through observations. Another possibility is to gain the knowledge through communication and in this case the individuals do not have to be limited to the neighbors.
- Social states (norms etc.) as viewed by the agent: For simulation of social behavior the agents need to be equipped with mechanisms for reasoning at the social level (unless the social level is regarded as emergent from individual behavior and decision making). Several models have been based on theories from economy, social psychology, sociology etc. An example of this is provided by *Guye-Vuillème* [35.46] who has developed an agent-based model for simulating human interaction in a virtual reality environment. The model is based on sociological concepts such as roles, values, and norms and motivational theories from social psychology to simulate persons with social identities and relationships.

In most simulation studies, the behavior of the individuals is static, i. e., the decision rules or reasoning mechanisms do not change during the simulation. However, human beings and most animals do have an ability to adapt and learn. To model dynamic behavior of individuals through learning or adaptation can be done in many ways. For instance, both ACT-R and Soar have learning built in. Other types of learning include the internal modeling of individuals (or the environment) where the models are updated more or less continuously.

Finally, there are some more general aspects to consider the modeling of individuals. One such aspect is whether all the agents share the same behavior or they behave differently, i. e., representation of behavior is ei-

ther individual or uniform. Another general aspect is the number of individuals modeled, i. e., the size of the model, which may vary from a few individuals to billions of individuals. Moreover, the population of individuals could be either static or dynamic. In dynamic populations, changes in the population are modeled, typically births and deaths.

#### Model of the Interaction Between Individuals

In dynamic micro-simulation each simulated individual is considered in isolation without regard to their interaction with others. However, in many situations the interaction between the individuals is crucial for the behavior at system level. Thus, in such cases better results will be achieved if the interactions between individuals were simulated. Two important aspects of interactions are who is interacting with whom, i. e., the interaction topology, and the form of interaction.

A basic form of interaction is physical interaction (or interaction based on spatial proximity). As we have seen, this is used in simulations based on cellular automata, e.g., in the Game of Life, introduced in Sect. 35.2.1, *Dynamic Micro-Simulation*. Another example is the Boids model [35.18], which simulates coordinated animal motion such as bird flocks and fish schools in order to study emergent phenomena. In these examples, the interaction topology is limited to the individuals immediately surrounding an individual. In other cases, as we will see below, the interaction topology is defined more generally in terms of a (social) network. Such a network can be either static, i. e., the topology does not change during a simulation, or dynamic. In these networks the interaction is typically language based. An example of this is the work by *Verhagen* [35.47] where agents that are part of a group use direct communication between the group members to form shared group preferences regarding the decisions they make. Communication is steered by the structure of the social network regardless of the physical location of the agents within the simulated world.

#### Model of the Environment

The state of the environment is usually represented by a set of (global) parameters, for example temperature. In addition, there are a number of important aspects of the environment model, such as:

- **Spatial explicitness:** In some models, there is actually no notion of physical space at all. An example of a scenario where location is of less importance is, for example, *innovation networks* [35.48] in which the individuals are high tech firms that each have a knowledge base that they use to develop artifacts to launch on a simulated market. The firms are able

to improve their innovations through research or by exchanging knowledge with other firms. However, in many scenarios location is very important, and in those each individual (and sometimes each object) is assigned a specific location at each time step of the simulation. In this case, the individuals may be either static (the entity does not change location during the simulation) or mobile. The location could either be specified as absolute positions in the environment, or in terms of relative positions between entities. In some areas the simulation software is integrated with a geographical information system (GIS) in order to achieve a closer match to reality, see [35.49].

- **Time:** There are in principle two ways to address time, and one is to ignore it. In static simulation time is not explicitly modeled; there is only a *before* and an *after* state. However, most simulations are dynamic, where time is modeled as a sequence of time steps. Typically, each individual may change state between each time step.
- **Exogenous events:** This is the case when the state of the environment, for example the temperature, changes without any influence or action from the individuals. Exogenous events, in case they are modeled, may also change the state of entities, for example, decay of resources, or cause new entities to appear. This is a way to make the environment stochastic rather than deterministic.

### 35.3.2 Engineering Agent-Based Simulations

#### Factors to Consider When Choosing a Model

In contrast to some of the more traditional approaches, such as system dynamics modeling, agent-based modeling does not yet have any standard procedures that can support the model development. During the last decade some attempts in this direction have been proposed. For example, *Grimm* et al. [35.50] proposed a structure for documenting an agent-based simulation model originally in the area of ecological systems. However, it is often the case that the only formal description of the model is the actual program code. However, it may be useful to use the unified modeling language (UML) to specify the model [35.51].

Some of the modeling decisions are determined by the features of the system to be simulated, in particular those regarding the interaction model and the environment model. The hardest design decision is often how the mental state and the behaviors of individuals should be modeled, in particular in the case when the individuals are human beings. For simpler animals or machines, a feature vector together with a set of transitions rules

is often sufficient. Depending on the phenomena being studied, this may be sufficient also when modeling human beings. *Gilbert* [35.52] provides some guidelines whether a more sophisticated cognitive model is necessary or not. He states that the most common reason for ignoring other levels is that the properties of these other levels can be assumed to be constant, and exemplifies this by studies of markets in equilibrium where the preferences of individual actors is assumed to remain constant (note, however, that this may not always be true). Another reason for ignoring other levels, according to Gilbert, is when there are many alternative processes at the lower level that could give rise to the same phenomenon. He exemplifies this by the famous study by *Schelling* [35.14] regarding residential segregation. Although Schelling used a very crude model of the mental state and behavior of the individuals, i. e., ignoring the underlying motivations for household migration, the simulation results were valid (as the underlying motivations were not relevant for the purpose of Schelling's study). On the other hand, there are many situations where a more sophisticated cognitive model is useful, in particular when the mental state or behavior of the individual provides constraints on, or in other ways influences the behavior at the system level. However, as Gilbert concludes, the current research is not sufficiently mature in order to give advice on which cognitive model to use (BDI, Soar, ACT-R, or other). Rather, he suggests that more pragmatic considerations should guide the selection.

The model of the environment is mostly dictated by the system to be simulated, where the modeler has to decide on the granularity of the values of the attributes of the environment. The interaction model is often chosen based on the theory or practical situation that is the basis for the simulation, but sometimes the limitations of the formal framework used restricts the possibilities. Also here the modeler has to decide upon the granularity of the values of the attributes. In general *Edmonds* and *Moss* [35.53] give the advice that a modeler shall not optimize for simplicity of the model, but more for understandability and believability.

#### More Issues in Engineering Agent-Based Simulations

Decision making about the granularity of agent decision is not the only issue when developing an agent-based simulation model for a complex system – yet is the most important one. In the following we will discuss more issues. A more elaborate discussion of issues can be found in [35.54].

**Generative Micro-Macro Link.** A basic reason for attractiveness of agent-based simulation for complex sys-

tem modeling comes from its generative nature [35.36]. The structure and dynamics of the overall system are not directly described, but generated from behavior and interactions of simulated, individual agents. So, there are at least two levels of modeling and observation: the low-level agents and the aggregate system level. *Running* the low level produces the structure and behavior on the aggregate level; in general, a formal a priori analysis before simulating the system is hardly possible, only by running the simulation, the what, where and when of a social phenomenon emerging from the low-level agents, can be fully determined. Formal analyses (or even prediction) of overall simulation outcomes from low-level agent behavior are difficult or impossible.

In many applications, a certain macro level phenomenon in the original system is to be reproduced or optimized. Thus, the micro-level rules determining the behavior have to be adapted in a way that the intended aggregate phenomenon is produced. For agent-based simulation, exploratory, experience-based, less informal methodologies appear to be more appropriate [35.55]. The basic question on how to come up with the appropriate low-level behavior is left to individual creativity and experience. This issue, together with the general level of granularity of the model has also been discussed in the last subsection.

#### Critical Parameter Structures and Calibration.

A simple model is preferable as it contains fewer assumptions and thus less parameters. Parameters may be factors in formulas or thresholds for decision making. Also values for initial values for state variables of all entities are parameters. A model with too many parameters can be tuned to produce anything. That means, a constellation of parameter values can be found, so that the given agent actions and interactions produce any intended overall outcome. So, structural falsification becomes impossible. This limits the analytical value of the model if not accompanied with rigorous processes for quality assurance.

Many parameters also cause practical problems: they need to be set to appropriate values. The necessary effort for calibration becomes an issue in developing agent-based simulations. Hereby, one has to pay attention as the sheer number of parameters may be critical. This can be remedied by putting parameters in relation to each other. Another important problem relates to the nature of the parameters themselves. A single parameter can have an enormous effect on the overall aggregated behavior when shared by many agents or if there are nonlinear feedback loops amplifying its effect. *Izquierdo* and *Polhill* [35.56] denoted decision threshold parameters as *knife-edge parameters*: if the

behavior of the agent changes depending on this parameter. Setting such a parameter homogeneously for all agents can cause chaotic behavior: small changes in the parameter values result in completely different phenomena. This issue can be addressed by allowing individual values for the individual agents or by *smoothing* the effects of a threshold with some stochastic transition. Finally, the test whether the simulation leads to the intended system and individual behavior may not be automatized – especially in abstract models without sufficient underlying empirical data – but human intelligence is needed to identify whether the generated pattern of the complex system is the searched one.

**Size and Scalability.** There is a variety of complex system models with respect to the number of agents, ranging from one-agent systems capable of complex interaction behavior to large-scale simulations with several millions of agents. For many complex system models, a minimum agent number is necessary – for example the effect of pheromone-based ant recruitment cannot be shown by only a small number of simulated ants, but the number of agents has to be synchronized with the environmental configuration and the evaporation rate of the pheromone used to establish a pheromone trail. Scalability with respect to agent numbers is only half of the story: scalability depends on the complexity of the agent behavior and architecture as discussed above. Complex system models contain agents that are capable of flexible decision making, which is clearly more costly than fully scripted behavior programs. With an appropriate tool, as one of the platforms discussed in Sect. 35.3.2, *More Issues in Engineering Agent-Based Simulations*, at least the simpler scalability issues are addressable.

**Other Technical Issues.** Besides those principled problems, there one can identify engineering issues at technical design and implementation – often supported by tools and platforms. If an analysis of the dynamics of the complex system model is needed, the model must be implemented. This is challenging in a way similar to multi-agent systems. Simulated multi-agent systems are also consisting of distributed intelligent decision makers, each with its own thread of control, its local beliefs and interacting and acting in parallel. In addition to the challenges developing a multi-agent system, one can identify:

- Issues about extended design choices on the environmental model. For facilitating the design of the agents, the environmental model can be augmented: A prominent example is crowd simulations using

floor fields capturing gradient data for path finding. Information is explicitly stored in the environment without any correspondence to the original system, but for making agent implementation more efficient. Again, it is a matter of modelers' experience to know how far one can go with these additions.

- During simulation, virtual time is advanced to express the dynamics of the model. As environment and time are artificial, the modeler needs a way for explicitly handling artificial parallelism of the agent's update. In principle, every agent could run in its own software process, but for simple agents explicitly handling virtual parallelism is more efficient. Depending on used infrastructure, the modeler has to take care about these low-level aspects of simulation implementation.

### Tools for Agent-Based Simulation of Complex Systems

Despite the many available tools, implementation of an agent-based simulation model is still not trivial. The currently most prominent tools suitable for complex systems are Swarm, Repast, MASON and NetLogo. Platforms such as SeSAM support modeling better, yet simulation runs tend to be slow. In the following, we will shortly discuss these tools. More elaborate comparisons can be found in [35.57, 58] or [35.59]

**Swarm.** Swarm [35.60] is one of the earliest tools for implementation of agent-based simulations (ABSs) and complex systems. Practically, it provides libraries (in Objective-C or newly also JAVA) that developers can use when building their simulations. Agents are hierarchically organized in Swarms.

**Repast.** Repast (Recursive Porous Agent Simulation Toolkit [35.61]) is also a Java-based platform. Following the hierarchical structure of Swarm, Repast provides a library of classes for the most common tasks associated with the implementation of an ABS. Besides, since the initial focus of Repast was social science, it includes some tools that are useful in this domain such as network analysis. The Repast Symphony forms a visual modeling extension based on state charts.

**MASON.** MASON (Multi-Agent Simulator Of Neighborhoods [35.62]) forms a library based on Java with the goal to particularly support large-scale simulations.

**NetLogo.** NetLogo [35.63] is currently probably the most used platform. It was particularly designed for complex system modeling and simulation with the end user in mind. A NetLogo model has basically three elements. The first is the actual implementation of model

behavior. The used modeling language resembles StarLogo, which is easy to understand and learn. The second and third element of a NetLogo model is the simulation interface for visualization and parameter settings and a third explicit element is a structured documentation.

NetLogo is becoming increasingly popular due to its extensive documentation, the existence of good tutorials, and a large library of preexisting models. Introductory books such as [35.64] are based on NetLogo.

**SeSAM.** The Shell for Simulated Agent Systems [35.65] provides a fully visual interface for the model development. A proprietary model representation language forms the basis for visual programming, etc. The kernel of a SeSAM simulation consists of the behavior models of agents and the world, which is represented as a special, global agent that may manage different kinds of maps.

**Other Tools.** In repositories for agent-based simulation platforms many more systems are listed [35.66]

and [35.67]. An overview is practically impossible. Many other tools are specifically designed for particular purposes. For example, MadKit relies upon an organizational model of agents' societies. Therefore its particular strength is in models focusing on intra- and inter-organizational processes. Similarly, CORMAS (Common-pool Resources and Multi-Agent Systems) is a programming environment that targets natural resources management.

In principle, an agent-based simulation platform shall not just support the implementation of agent behavior, but provide basic infrastructure for, for example, integration of input data, handling virtual time, model instrumentation, data collection, and others. Depending on the nature of the tool, the expressiveness of the language for capturing the agent behavior might be limited. Whether it is sufficient or not, is dependent on the actual objective behind simulating the complex system. Nevertheless, a good development platform amends some of the issues discussed above and thus enables the modeler to concentrate on the core aspects of the model.

## 35.4 Summing Up and Future Trends

The ability to understand and manage different types of complex systems is becoming more and more important, both for research, businesses and government. As we have seen, agent-based modeling and simulation seems a promising approach to many problems involving the simulation of complex systems of interacting entities. Although a large number of different methods and tools for agent-based modeling and simulation have been developed, it seems that the full potential of the agent concept often is not realized. In particular, this is the case when modeling complex systems that include human actors. For instance, most models use a very primitive model of agent cognition yet as argued in [35.52] cognitive layers of agent architectures should be intertwined with social layer.

The question of how to balance complexity of the agents' reasoning and transparency and comprehensiveness of the overall system behavior forms one of the many open issues from a methodological point of view. Although agent-based modeling of complex system forms a highly attractive approach for the full variety of possible objectives, there is not yet any established method for developing a model in a sys-

tematic way, comparable to system dynamics. This may be due to the fact that modeling per se often contributes to understanding a complex system. That is, the complex system to be modeled is not fully understood by the stakeholders or even by the modelers themselves before the modeling and simulation endeavor starts – independent of which objective the model is developed for. Thus, model development and model analysis also contain elements of original system exploration that, especially in the case of complex systems, may lead to new questions and insights influencing the ongoing model development process. Even if the specific open issues that we discussed in Sect. 35.3.2, *More Issues in Engineering Agent-Based Simulations*, are addressed by new methodologies to be developed, this will probably remain as a profound issue as long as there are systems that appear to be *complex*. More generally, we argue that the art and practice of engineering agent-based models is an important area of future research; as for handling, predicting and especially for understanding complex systems, modeling and simulation form the centerpiece of any activity.

## References

- 35.1 J. Ladyman, J. Lambert, K. Wiesner: What is a complex system?, *Eur. J. Philos. Sci.* **3**(1), 33–67 (2013)
- 35.2 P. Bak, M. Paczusi: Complexity, contingency, and criticality, *Proc. Natl. Acad. Sci. USA* **92**(15), 6689–6696 (1995)
- 35.3 S. Auyang: *Foundations of Complex Systems Theories in Economics, Evolutionary Biology and Statistical Physics* (Cambridge Univ. Press, Cambridge 1999)
- 35.4 J. Parker: A flexible, large-scale, distributed agent based epidemic model, *Proc. 39th Winter Simul. Conf.* (2007) pp. 1543–1547
- 35.5 R. Axelrod: A Model of the emergence of new political actors. In: *Artificial Societies: The Computer Simulation of Social Life*, ed. by N. Gilbert, R. Conte (Univ. College Press, London 1995)
- 35.6 S. Kauffman: *The Origins of Order: Self-Organization and Selection in Evolution* (Oxford Univ. Press, Oxford 1993)
- 35.7 J.H. Holland: *Emergence: From Chaos to Order* (Addison-Wesley, Redwood City 1998)
- 35.8 V. Darley: Emergent phenomena and complexity, *Artif. Life IV Proc. Fourth Int. Workshop Synth. Simul. Living Syst.*, ed. by R.A. Brooks, P. Maes (MIT Press, Cambridge 1994) pp. 411–416
- 35.9 J.S. Lansing: “Artificial societies” and the social sciences, *Artif. Life* **8**, 279–292 (2002)
- 35.10 R.K. Sawyer: Artificial societies – Multi-agent systems and the micro-macro link in sociological theory, *Sociol. Meth. Res.* **31**(3), 325–363 (2003)
- 35.11 H.V.D. Parunak, R. Savit, R.L. Riolo: Agent-based modeling vs. equation-based modeling: A case study and users’ guide, *Lect. Notes Comput. Sci.* **1534**, 10–25 (1998)
- 35.12 J.L. Schiff: *Cellular Automata: A Discrete View of the World* (Wiley, Hoboken 2008)
- 35.13 M. Gardner: The fantastic combinations of John Conway’s new solitaire game “life”, *Sci. Am.* **223**, 12–123 (1970)
- 35.14 J.M. Epstein, R.L. Axtell: *Growing Artificial Societies: Social Science from the Bottom Up* (MIT Press, Cambridge 1996)
- 35.15 T.C. Schelling: Dynamic models of segregation, *J. Math. Sociol.* **1**(1), 143–186 (1971)
- 35.16 N.A. Barricelli: Symbiogenetic evolution processes realized by artificial methods, *Methodos* **9**(35–36), 143–182 (1957)
- 35.17 S. Galam: Sociophysics: A review of Galam models, *Int. J. Mod. Phys. C* **19**(3), 409–440 (2008), doi:10.1142/S0129183108012297
- 35.18 C.W. Reynolds: Flocks, herds, and schools: A distributed behavioral model, *Comput. Graph* **21**(4), 25–34 (1987)
- 35.19 N. Gilbert: Computer simulation of social processes. Social research update, Issue 6, Department of Sociology, University of Surrey, UK (1994), <http://sru.soc.surrey.ac.uk/SRU6.html>, Accessed 15 Feb 2015
- 35.20 N. Gilbert, K.G. Troitzsch: *Simulation for the Social Scientist*, 2nd edn. (Open Univ. Press, Maidenhead 2005)
- 35.21 J. Wang: Petri nets for dynamic event-driven system model. In: *Handbook of Dynamic System Modeling*, ed. by P. Fishwick (CRC, Boca Raton 2007), Chap. 24
- 35.22 C.G. Cassandras: Queuing system models. In: *Handbook of Dynamic System Modeling*, ed. by P. Fishwick (CRC, Boca Raton 2007), Chap. 25
- 35.23 B.P. Zeigler: *Object Oriented Simulation with Hierarchical Modular Models: Intelligent Agents and Endomorphic Systems* (Academic, London 1990)
- 35.24 T. Takahashi: Agent based disaster simulation evaluation and its probability model interpretation, *Proc. 4th Int. Conf. Intell. Hum.-Comput. Syst. Crisis Resp. Manag.*, Delft (2007) pp. 369–376
- 35.25 J.C. Alexander, B. Giesen, R. Münch, N.J. Smelser (Eds.): *The Micro-Macro Link* (Univ. California Press, Berkeley 1987)
- 35.26 P. Davidsson: Multi-agent-based simulation: Beyond social simulation, *Lect. Notes Comput. Sci.* **1979**, 98–107 (2000)
- 35.27 D. Massaguer, V. Balasubramanian, S. Mehrotra, N. Venkatasubramanian: Multi-agent simulation of disaster response, *Proc. 1st Int. Workshop Agent Technol. Disaster Manag.*, Hakodate (2006)
- 35.28 L. Brouwers, H. Verhagen: Applying the Consumat model to flood management policies, *4th Workshop Agent-Based Simul.* (2004) pp. 29–34
- 35.29 D. Yergens, J. Hiner, J. Denzinger, T. Noseworthy: IDESS – A multi-agent-based simulation system for rapid development of infectious disease models, *Int. Trans. Syst. Sci. Appl.* **1**(1), 51–58 (2006)
- 35.30 B. Raney, N. Cetin, A. Völlmy, M. Vrtic, K. Axhausen, K. Nagel: An agent-based microsimulation model of swiss travel: First results, *J. Netw. Spatial Econ.* **3**(1), 23–41 (2003)
- 35.31 R. Williams: An agent based simulation environment for public order management training, *West. Simul. Multiconf., Object-Oriented Simul. Conf.*, San Diego (1993) pp. 151–156
- 35.32 M. Wooldridge: *An Introduction to Multiagent Systems* (Wiley, Hoboken 2009)
- 35.33 Y. Shoham: Agent-oriented programming, *Artif. Intell.* **60**, 51–92 (1992)
- 35.34 M.W. Macy, R. Willer: From factors to actors: Computational sociology and agent-based modeling, *Annu. Rev. Sociol.* **28**, 143–166 (2002)
- 35.35 M.J. Prietula, K.M. Carley, L. Gasser (Eds.): *Simulating Organizations: Computational Models of Institutions and Groups* (MIT Press, Cambridge 1998)
- 35.36 J.M. Epstein: *Generative Social Science: Studies in Agent-Based Computational Modeling* (Princeton Univ. Press, Princeton 2007)
- 35.37 J. Künzel, V. Hämmer: Simulation in university education: The artificial agent PSI as a teaching tool, *Simulation* **82**(11), 761–768 (2006)
- 35.38 M.E. Bratman: *Intentions, Plans, and Practical Reason* (Harvard Univ. Press, Cambridge 1987)

- 35.39 M. Georgeff, B. Pell, M. Pollack, M. Tambe, M. Wooldridge: The Belief-Desire-Intention model of agency, *Lect. Notes Comput. Sci.* **1555**, 1–10 (1998)
- 35.40 A.L.C. Bazzan, R.H. Bordini: A framework for the simulation of agents with emotions: Report on experiments with the iterated prisoners dilemma, 5th Int. Conf. Auton. Agents (2001) pp. 292–299
- 35.41 J. Broersen, M. Dastani, Z. Huang, J. Hulstijn, L. Van der Torre: The BOID architecture: Conflicts between beliefs, obligations, intentions and desires, 5th Int. Conf. Auton. Agents (2001) pp. 9–16
- 35.42 A. Newell: *Unified Theories of Cognition* (Harvard Univ. Press, Cambridge 1994)
- 35.43 G. Méndez, J. Rickel, A. de Antonio: Steve meets Jack: The integration of an intelligent tutor and a virtual environment with planning capabilities, *Lect. Notes Comput. Sci.* **2792**, 325–332 (2003)
- 35.44 J.R. Anderson, D. Bothell, M.D. Byrne, S. Douglass, C. Lebiere, Y. Qin: An integrated theory of mind, *Psychol. Rev.* **111**(4), 1036–1060 (2004)
- 35.45 M. A. Janssen, W. Jager: An integrated approach to simulating behavioural processes: A case study of the lock-in of consumption patterns, *J. Artif. Soc. Soc. Simul.* **2**(2) (1999)
- 35.46 A. Guye-Vuillème: Simulation of Nonverbal Social Interaction and Small Groups Dynamics in Virtual Environments, Ph.D. Thesis (École Polytechnique Fédérale de Lausanne, Lausanne 2004)
- 35.47 H. Verhagen: Simulation of the learning of norms, *Soc. Sci. Comput. Rev.* **19**(3), 296–306 (2001)
- 35.48 N. Gilbert, A. Pyka, P. Ahrweiler: Innovation networks – A simulation approach, *J. Artif. Soc. Soc. Simul.* **4**(3) (2001)
- 35.49 M. Schüle, R. Herrler, F. Klügl: Coupling GIS and multi-agent simulation – Towards infrastructure for realistic simulation, *Lect. Notes Comput. Sci.* **3187**, 228–242 (2004)
- 35.50 V. Grimm, U. Berger, F. Bastiansen, S. Eliassen, V. Ginot, J. Giske, J. Goss-Custard, T. Grand, S.K. Heinz, G. Huse, A. Huth, J.U. Jepsen, C. Jørgensen, W.M. Mooij, B. Müller, G. Pe'er, C. Piou, S.F. Railsback, A.M. Robbins, M.M. Robbins, E. Rossmanith, N. Rüger, E. Strand, S. Souissi, R.A. Stillman, R. Vabø, U. Visser, D.L. DeAngelis: A standard protocol for describing individual-based and agent-based models, *Ecol. Model.* **198**, 115–126 (2006)
- 35.51 P. Bommel, J.-P. Müller: An introduction to UML for modelling in the human and social sciences. In: *Agent-Based Modelling and Simulation in the Social and Human Sciences*, ed. by D. Phan, F. Amblard (Bardwell, Oxford 2007) pp. 273–294
- 35.52 N. Gilbert: When does social simulation need cognitive models? In: *Cognition and Multi-Agent Interaction From Cognitive Modeling to Social Simulation*, ed. by R. Sun (Cambridge Univ. Press, Cambridge 2006) pp. 428–432
- 35.53 B. Edmonds, S. Moss: From KISS to KIDS – An ‘anti-simplistic’ modelling approach, *Lect. Notes Comput. Sci.* **3415**, 130–144 (2004)
- 35.54 F. Klügl: “Engineering” agent-based simulation models?, *Lect. Notes Comput. Sci.* **7852**, 179–196 (2012)
- 35.55 E. Norling, B. Edmonds, R. Meyer: Informal approaches to developing simulation models. In: *Simulating Social Complexity, Understanding Complex Systems*, ed. by B. Edmonds, R. Meyer (Springer, Berlin, Heidelberg 2013) pp. 39–55
- 35.56 L.R. Izquierdo, J.G. Polhill: Is your model susceptible to floating-point errors?, *J. Artif. Soc. Soc. Simul.* **9**(4), 4 (2006)
- 35.57 S.F. Railsback, S.L. Lytinen, S.K. Jackson: Agent-based simulation platforms: Review and development recommendations, *Simulation* **82**(9), 609–623 (2006)
- 35.58 C. Nikolai, G. Madey: Tools of the trade: A survey of various agent based modeling platforms, *J. Artif. Soc. Soc. Simul.* **12**(2), 2 (2009)
- 35.59 F. Klügl, A.L.C. Bazzan: Agent-based modelling and simulation, *AI Mag.* **33**, 29–40 (2012)
- 35.60 Swarm Development Group: <http://www.swarm.org>
- 35.61 Argonne National Laboratory: <http://www.repast.sourceforge.net>
- 35.62 George Mason University: <http://cs.gmu.edu/~eclab/projects/mason/>
- 35.63 NetLogo Team: <http://ccl.northwestern.edu/netlogo>
- 35.64 S. Railsback, V. Grimm: *Agent-Based and Individual-Based Simulation – A Practical Introduction* (Princeton Univ. Press, Princeton 2012)
- 35.65 SeSAM Team: <http://www.simsesam.org>
- 35.66 J. Schank: Simulators, <http://www.agent-based-models.com/blog/resources/simulators>
- 35.67 L. Tesfatsion: General software and toolkits, <http://www2.econ.iastate.edu/tesfatsi/acecode.htm>

# 36. Models and Experiments in Robotics

Francesco Amigoni, Viola Schiaffonati

This chapter surveys the practices that are being employed in experimentally assessing the special class of computational models embedded in robots. The assessment of these models is particularly challenging mainly due to the difficulty of accurately estimating and modeling the interactions between the robots and their environments, especially in the case of autonomous robots, which make decisions without continuous human supervision. The field of autonomous robotics has recognized this difficulty and launched a number of initiatives to deal with it. This chapter, after a conceptual premise and a broad introduction to the experimental issues of robotics, critically reviews these initiatives that range from taking inspiration from traditional experimental practices, to simulations, benchmarking, standards, and competitions.

36.1	<b>A Conceptual Premise</b> .....	799
36.2	<b>Experimental Issues in Robotics</b> .....	801
36.3	<b>From Experimental Computer Science to Good Experimental Methodologies in Autonomous Robotics</b> .....	802
36.4	<b>Simulation</b> .....	804
36.5	<b>Benchmarking and Standards</b> .....	807
36.6	<b>Competitions and Challenges</b> .....	809
36.7	<b>Conclusions</b> .....	812
	<b>References</b> .....	812

## 36.1 A Conceptual Premise

The issues involved in the experimental assessment of computational models are several and diverse. In this chapter, we focus on a challenging class of computational models that present complex interactions with the real world, namely those of robotics, and in particular of *autonomous robotics* (intended as the discipline aiming at developing robots that operate in unpredictable environments without a continuous human supervision), and on how experiments have recently been conceptualized, discussed, and performed in this field. Taking the perspective that robots are a way to implement models of human-designed processes, these models need to be validated and experiments are usually the way to carry out this validation. This section, therefore, is dedicated to the clarification of the key concepts – robot system, computational model, technical artifact, and experiment – that constitute the starting points of what will be discussed in the rest of the chapter.

In general, a robot system can be conceived as a computational model, provided with a specific ar-

chitecture, capable to interact with the surrounding physical environment by means of sensors and actuators, and to fulfill a function which is the reason why the system has been designed and implemented. By a computational model, we mean a representation that is formulated in terms of computable functions [36.1], where a function  $X()$  is said to be computable if for each argument  $n$  of  $X$ , the corresponding value  $X(n)$  can be obtained by adopting a mechanical calculus (algorithm). In other words, a model is computational if it can be defined in purely mechanical terms: only mechanical procedures of calculus are required to give reason of it. But what is modeled in the case of a robot system, given that the associated model is computational? First, the phenomena or the behavior that the robot system should represent; then, (part of) the environment in which the robot system is inserted and with which it should interact, where the interaction between the system and the environment strongly depends on the way the system itself is modeled.



Let us focus our attention now on the reasons why a robot system is built; these are generally related to responding to some needs, such as to perform some actions in the environment or, more generally, to obtain a desired behavior. The conceptualization of robots as *technical artifacts* [36.2], namely as material objects that have been deliberately produced by humans in order to fulfill a practical function, may help in better clarifying these reasons. The term *artifact* emphasizes the fact that these objects are not naturally occurring, but they are the result of purposeful human actions. More generally, technical artifacts can be characterized as possessing three key features:

- The *technical function* that is related to the question *What is the technical artifact for?* In robotics, the technical function can be defined by referring to the task the robots have to accomplish and to the environment in which the task is performed. Note that a task includes both an activity and a way to quantify a performance in executing this activity.
- The *physical composition* that is related to the question *What does the technical artifact consist of?* In particular, the physical components of robots are selected in order to cope with the intended task and environment. For example, using wheels for locomotion can be afforded only if the environment in which robots are expected to move is rather smooth, planar, and relatively empty. If this is not the case, locomotion needs to be based on legs or crawlers, or the use of aerial robots could be considered. Similarly, equipping a robot with a thermal camera could be useful if the task is to detect victims, while if the task is to assess the topological structure of a building, other sensors could be more appropriate.
- The *instructions for use* that are related to the question *How must the technical artifact be used?* For example, the use of autonomous robots in search and rescue applications is a rather complex job and usually requires the presence of human operators to supervise operations and to actively intervene on the system in case of unexpected problems. Hence, the instructions for using these robots are significantly complex and usually require human operators to undergo some training. Timing in issuing commands and attention in guaranteeing the safety of all involved people (and, to a lesser degree, of all robots) are fundamental and should be part of the training of human operators.

These three features are not independent of each other: to fulfill the technical function that the artifact is for, the artifact itself has to be physically composed in a certain way and the user has to carry out certain actions, specified by the instructions for use. In sum-

mary, a technical artifact can be said to be a “physical object with a technical function and use plan designed and made by human beings” [36.2, p. 7].

Adopting the view that a robot system is a technical artifact embedding a computational model of an intended behavior, its assessment is usually conducted on the field, and experiments are a way to achieve this assessment in a rigorous way. In this context, experiments can be seen as a way to evaluate technical artifacts according to whether and to what amount the function for which they have been built is fulfilled. Apparently, a contradiction seems to arise when considering the necessity of evaluating an artifact, that has been designed and implemented by a human being, and whose complete knowledge, thus, should be under the control of that designer. However, not only these types of artifacts can be so complex that it is difficult to fully manage their knowledge, but also their interaction with the environment is largely unpredictable and, therefore, a rigorous evaluation is needed, which is mainly based on experimentation. Obviously, experimentation can take many forms and to help clarify the kind of experimentation embraced in autonomous robotics, we deem useful to adopt the distinction between epistemic experiments and directly action-guiding experiments recently pointed out by *Hansson* [36.3].

*Epistemic experiments* provide us with information about the workings of the world we live in. The outcome looked for is precisely the one providing such information and does not need to coincide with anything that any person would wish to happen (except as part of the experiment itself). This is the traditional view of experimentation that both historical and philosophical accounts of experimental methods have been almost exclusively devoted to analyze. But if we look at the practice of experimentation in prescientific times, as *Hansson* suggests, we can recognize that there are also so-called *directly action-guiding experiments*, namely experiments that satisfy the following two criteria:

- The outcome looked for should consist in the attainment of some desired goal of human action.
- The interventions studied should be potential candidates for being performed in a nonexperimental setting in order to achieve that goal.

These criteria are satisfied, for example, in a clinical trial of an analgesic, where the outcome looked for is pain reduction, and the intervention is a treatment to be administered to achieve this outcome in ordinary patients. A systematic test on an autonomous robot employed to assist an elderly person in her home is also an example of a directly action-guiding experiment: the outcome looked for is the proper interaction of the robot with the person and the experimental intervention con-

sists in the careful tuning of the abilities that the robot must possess to positively achieve this goal.

The notion of directly action-guiding experiments can be applied to a number of different situations, and indeed to many technological tests performed in the engineering disciplines, including robotics. In the rest of the chapter, we will base our analysis and discussion

referring mostly to this notion of experiment, even if it will be evident that a clear cut distinction between episodic and directly action-guiding experiments is not always possible being the experimental validation of robot systems at the intersection between science and engineering under many respects, as we will try to evidence in the next sections.

## 36.2 Experimental Issues in Robotics

The experimental validation of robots has been playing a decisive role in many areas of robotics, such as in industrial robotics, by addressing both the measures of robot performance and the assessment of their safety. These experiments usually take the form of tests, where standardized procedures have been developed. To assess, for example, the performance of commercialized industrial manipulators (in terms of accuracy, repeatability, and spatial resolution) rigorous testing protocols have been provided (see later). In other areas of robotics research, such as in the field of autonomous robotics for service applications, the term *experiment* has been intended more vaguely and experimental activities are often carried out with lower standards of methodological rigor.

In the last years, however, the autonomous robotics community has experienced a growing interest for the development of good experimental methodologies. This interest can be traced back to different reasons: from a scientific perspective, it concerns the desire of this rather novel community to adopt the same methodological standards of other scientific disciplines; from a more practical and commercial perspective, it deals with the possibility of measuring some parameters in a standard way (e.g., safety of a home assistant robot) or of having objective benchmarks to compare and evaluate different products.

In an effort of improving the quality of experimental activities, some attempts [36.4] have been made to take inspirations from how experiments are performed in traditional sciences, such as physics and biology, by trying to translate in the practice of autonomous robotics the general experimental principles of natural sciences (comparison, repeatability, reproducibility, justification, explanation, . . .). However, from a recent analysis [36.5], it emerges that these principles are not yet full part of the current research practice. Let us consider for example reproducibility, namely the possibility to independently verify the results of a given experiment so that different experimenters should be able to achieve the same results, by starting from the same initial conditions, by using the same type of instruments, and by adopting the same experimental

techniques. Notwithstanding emphasis is put on the importance of reproducibility (usually called replicability in this context) as a way to increase the experimental level of the field, good practices to promote it, such as the availability of shared data and code, are still not very common and few attempts have been made to critically analyze how reproducibility should be attained in experiments with autonomous robots.

It is true that, dealing with technical artifacts, robotics cannot be fully assimilated to a natural science, where experiments are generally conducted for hypotheses testing purposes and with a strong theoretical background. Robotics possess an engineering component that makes it plausible to adopt the notion of directly action-guiding experiment to give reason of its experimental practice: experiments in engineering fields have other objects (technical artifacts rather than natural phenomena) and other purposes (testing rather than understanding) – to put it simple – with respect to experiments in the sciences. Robot systems are human-made artifacts; accordingly, experiments have the goal of demonstrating that a given artifact is working with respect to a reference model (e.g., its requirements or its expected behavior) and, possibly, that it works better than other similar artifacts with respect to some metrics, thus making experiments closer to tests typical of engineering disciplines. At the same time, the most advanced robot systems are extremely complex, and their behavior is hardly predictable, even by their own designers, especially when considering their interactions with the natural world, which are difficult, if not impossible, to model in a fully satisfactory way. In this sense, experiments in autonomous robotics have also the goal of understanding how these complex systems work and interact with the environment and, therefore, are somehow similar to experiments in the natural sciences. We will see how this peculiar position at the intersection between science and engineering is reflected also in the experimental practice characterizing the field.

In the following sections, we attempt to depict the picture of how the topic of experiments in autonomous robotics has been recently addressed. However, as this

debate is still very open, we will not present any systematic and complete survey, but rather a number of issues, results, initiatives – also partially overlapped – able to give back the sense of this intriguing and interdisciplinary research field, devoted not only to the investigation and development of methodological approaches, but also to rethink of the traditional notion of experiment as developed so far. The rest of the chap-

ter, thus, will cover various topics stemming from the attempt to develop good experimental methodologies in autonomous robotics, to simulations, benchmarks, standards, and competitions. Being the field not yet fully systematized, we will present the state of the art intertwined with some considerations in order to offer a more structured, even if still partial, view of this field and its progress.

### 36.3 From Experimental Computer Science to Good Experimental Methodologies in Autonomous Robotics

The reflection on experiments and the application of the experimental method to the broad field of computer science has traversed the history of this discipline from its beginning. Probably one of the first and most famous definitions of computer science as an experimental science goes back to the 1976 paper by *Newell* and *Simon* published in the occasion of their acceptance of the Turing award [36.6, 114]:

“Computer science is an empirical discipline. We would have called it an experimental science, but like astronomy, economics, and geology, some of its unique forms of observation and experience do not fit a narrow stereotype of the experimental method. None the less, they are experiments. Each new machine that is built is an experiment. Actually constructing the machine poses a question to nature; and we listen for the answer by observing the machine in operation and analyzing it by all analytical and measurement means available.”

This conception of machines and programs as experiments has been influential for many years, promoting the idea that the appeal to experience is fundamental in contrast with the view of computer science as a pure mathematical and deductive discipline.

The quest for experiments in computing began to be treated systematically at the beginning of the 1980s, following a crisis in what was then called experimental computer science. In an Association for Computing Machinery (ACM) report published in 1979 [36.7], experimental research in computer science is strongly related to the measurement and testing of computing algorithms and systems. At the same time, a *rejuvenation* of experimental computer science is advocated from very concrete perspectives: for example, by promoting experimental facilities for computer systems research. Experimental computer science is to be rejuvenated also according to *Denning*, who proposed in a short article that the experimental work produced in computer science should be judged by traditional stan-

dards [36.8]. In a way, this approach tries to go beyond the *construct and test* paradigm of Newell and Simon, by proposing that experimental computer science has to deal with the process of supporting and testing hypotheses, thus making computing closer to the standards of rigor and the practice of traditional sciences.

More recently, a trend has once again emerged toward making the experimental scientific method take center stage in computing. These recent efforts have shown a renewed need for an experimental methodology in this discipline [36.9–12]. Experiments are deemed to have an impact on several aspects of computing: their importance is recognized for assessing computing systems’ performance and for triggering new developments. Despite the increasing interest in a more rigorous methodological approach to computing, many lament that the current methodology is inadequate and that, in comparison with other fields (e.g., natural sciences), computer scientists should experiment more [36.13]. Indeed, several articles describe demonstrations rather than real experiments [36.14], and their sections on experimental results present just weak examples to show the superiority of the proposed solution over a limited amount of alternatives [36.15, 16]. Many of these recommendations [36.17–20] present common traits: they stem from the acknowledgment of a crisis in computing that is meant to be overcome with a greater maturity of the discipline, in terms of a more rigorous experimental method and a more scientifically grounded approach to the search for solutions. Taking inspiration from experimental principles adopted in traditional scientific disciplines has become a leitmotif in many analyses, which recognize, for example, the benefits of replicable results [36.21] or the importance of negative results [36.22], two of the cornerstones of experimental scientific method. However, some authors warn about the acritical adoption of principles coming from traditional science [36.23].

The same concerns can be individuated also in the debate about the nature and role of experiments in

autonomous robotics. A seminal paper in this respect is [36.24] that, despite its emphasis on artificial intelligence (AI) and artificial agents, presents an insightful analysis on the topic of experimentation that constitutes the basics of the current debate in autonomous robotics. This paper, in particular, analyzes the possible relationships between controlled experimentation and the rise of benchmarks and testbeds as current research tools. At the same time, it warns against the idea that their use is sufficient to achieve scientific progress. According to the authors, unless generalization and explanation are provided, benchmarks and testbeds, even when appropriately providing metrics for comparing competing systems, do not constitute real scientific progress.

The interest in solidly based experimental research in autonomous robotics has progressively increased in the last 20 years. An interesting example is given by the symposium series on Experimental Robotics, which have been held starting from 1989 on the topics of experimental robotics research [36.25]. However, it is only in the last years that this interest has been coupled with a careful analysis on how the concept of experimentation should be translated in the practice of autonomous robotics, also giving rise to a debate about the status of the discipline itself. To this end both the creation of the EURON Special Interest Group on Good Experimental Methodology in Robotics Research (GEM) and the series of workshops about replicable experiments in robotics [36.26] have been playing a decisive role. The workshop series has contributed to raise several issues and to increase the sensibility of the community on these topics. If we consider the article [36.27] as representative of the kind of approach advocated we can see that, besides the discussion of practical issues, several more theoretical aspects are discussed, ranging from the scientific status of robotics to the definition of a robotics experiment. If, from the one side, much emphasis is put on the definition of a replicable robotic experiment, taking inspiration from the traditional approach of experimental sciences, from the other side, the guidelines proposed within the GEM Special Interest Group represent an attempt to translate in practice the same definition. In particular, the paper [36.26] presents a structured set of questions intended to help reviewers recognize, and authors write, high quality papers reporting of replicable experimental work, addressing questions such as *Are the system assumptions/hypotheses clear? Are the evaluation criteria spelled out explicitly? Is there enough information to reproduce the work? Are the drawn conclusions precise and valid?* The questions are further debated in the special issue on Replicable and Measurable Robotics Research of the IEEE Robotics and Automation Magazine [36.28].

Within the context set by the above efforts, a particular interesting proposal in the direction of promoting rigorous experimental research in robotics is that described in [36.29] in which, besides adopting the standards by which scientific experiments are usually designed, conducted, and presented, particular attention is devoted to the other sciences of the artificial, e.g., human–computer interaction and human–robot interaction, as a way for shaping good experimental research practices in the broader robotics community. It is worth noticing that this work is extremely mature in the acknowledgment and in the analysis of the many open critical issues toward the construction of a science of robotics, putting a right emphasis in particular on two key points: the humility to give oneself the chance to be wrong and the importance of generalizable results that, even if difficult, are not impossible.

In the same direction [36.4] identifies some basic issues of experiments in mobile robot localization and mapping, starting from a representative sample of the current state of the art. These issues, when viewed in the context of some general principles about experiments in science (comparison, reproducibility and repeatability, justification and explanation), permit to derive some insightful considerations on the role of experiments in mobile robotics, ranging from questions about the purpose of experiments to the current publication policies that should be revised accordingly. In particular, the analysis is conducted in the light of some principles that have been proposed for the development of good experimental methodologies in autonomous robotics, and in particular:

- *Comparison*, as the knowledge of what has been already done within a field and the possibility for researchers to accurately compare new results with the old ones.
- *Reproducibility*, as the possibility for independent scientists to verify the results of a given experiment by repeating it with the same initial conditions, instruments, and techniques.
- *Repeatability*, as the property of an experiment that yields the same outcome from a number of trials, possibly performed at different times and in different places.
- *Justification and explanation*, as the capacities to derive well-justified conclusions and to look for an explanation.

Moreover, [36.5] arguably provides one of the few systematic analyses of the current experimental trends in the autonomous robotics papers presented over 11 years at the *International Conference on Autonomous Agents and Multiagent Systems* (AAMAS).

The experimental trends identified from the sample of papers considered for the analysis range from those that fit with the above principles (more experiments in recent years, more use of simulations than real robots, increased use of standard platforms) to those that point out some issues that are still critical at the moment (weakness of experimental comparison of systems, little attention to statistical analysis, low availability of data and code) and show that some of the general experimental principles listed above are currently poorly attained.

It is interesting to observe that, from the general discussions on the ways to promote good experimental methodologies, in the very last years also more concrete proposals have been advanced. From the one side, attempts have been made to practically translate the general guidelines in different specific robotic contexts (see [36.30] for an example of how experimental methodologies and suitable metrics for performance evaluation can be applied to carry out repeatable experiments for unmanned marine vehicles). From the other side, different frameworks to practically promote reproducibility of robot systems experiments have been put at work (see [36.31], for an example of a novel process facilitating the reproduction of robotic experiments by

means of different software tools and [36.32] for a facility for experimenting with robots from remote).

An important issue in evaluation of experimental results in robotics is relative to *ground truth*, namely to the availability of the *optimal behavior* or of the *perfect performance* of a robot in a task. For example, for a robot intended to build the map of an initially unknown building, the ground truth is the blueprint of the building; for a robot intended to estimate its own location within a known environment using the data it perceives, the ground truth is the actual position of the robot (for instance, measured by hand). What sometimes makes the evaluation of robot systems difficult is the impossibility of knowing the ground truth, both before and after the completion of the tasks.

In what follows, we analyze the various concrete forms taken by the debate for the development of good experimental methodologies in autonomous robotics: we start by discussing computer simulations and how they are used as experiments in this field; then we continue with benchmarking as a way to objectively evaluate robots systems and the role played in it by standards, to conclude with an analysis on how competitions and challenges are developing toward a more experimental attitude.

## 36.4 Simulation

Although it is out of the scope of the present section to provide a systematic definition of simulations and to discuss the epistemological problems arising from their growing use in scientific practice (for a complete discussion on these issues see Chap. 34 this handbook), we start with a general terminological clarification to focus later on the way in which computer simulations are used as experiments in the field of autonomous robotics.

The ways a simulation can be defined are different and have been the object of attention of many scholars, particularly in the last 20 years (Chap. 34 this handbook). Without the pretense of taking into account the whole debate, we can say that in general a simulation can be considered a way to reproduce the behavior of a system using another system (in line with [36.33]), by providing a dynamic representation of a portion of reality. Simulations are closely related to dynamic models that include assumptions about the time-evolution of a system, as they are based on models that represent a dynamic portion of reality. However, simulations represent reality in a different way than models do. Without entering the discussion on the various meanings of the term model [36.34], it is sufficient for us to say that a model is a representation, where what is represented

in the model depends on the purposes for which the model has been conceived. Consider for instance a scale physical model of a bridge, which is a representation replicating some features of the real object (the bridge that is going to be built) by abstracting from the full details and concentrating on the aspects relevant for the purpose. For example, if the scale model has been constructed in order to show to its purchasers the final shape of the bridge, then it will not be important its material, color, or dimension. If, instead, the scale model has been constructed to test the resistance of some materials, used in construction, to some atmospheric agents, the sole model is not sufficient, but it has to be put in a (controlled) environment where it can be subjected to the actions of the atmospheric conditions. In this case the model is *executed* in the reality by means of the actions performed by the environment. We call this execution of a model a *simulation*. When the model is computational, namely a formal mechanism able to compute functions (Sect. 36.1) and executed by a computer, we call it a *computer simulation*. Therefore, not every execution of a computational model is a simulation. To call it a simulation, a computational

model must represent a system whose state changes in time.

It is undoubtedly evident that science today has entered what has been called the *age of computer simulation* [36.35]: the massive use of computer simulations in virtually every domain of science has drawn attention to their epistemological justification. Recently the experimental properties of computer simulations have been examined, and philosophers have begun to consider in what sense, if any, computer simulations are experiments (see [36.36] for a detailed analysis of this debate). Positions range from a full acceptance of the identity thesis (computer simulation studies are literally instances of experiments) to its rejection in different degrees. We believe that focusing on autonomous robotics, simulations *can be used* as experiments in the case in which the purposes of simulation and those of experiment coincide. This happens when experiments and simulations are performed for discovering new explanatory hypotheses, for confirming or refusing theories, for choosing among competing hypotheses, and so on. Indeed, besides exploiting similar techniques, experiments and simulations share the ability and necessity of controlling the features under investigation. The set-up of a simulation presents several communalities with that of an experiment: hypotheses to be tested, circumstances to be controlled, and parameters to be set. Moreover, simulations can be used as experiments for a number of practical reasons. They can be used to perform several accelerated experiments, as simulations can be repeated exactly with small efforts, with guarantee of a precision degree not always possible in empirical cases. Simulations can be used to perform experiments that are difficult to make in reality, being free from many of the practical limitations of real experiments, such as the possibility to change boundaries and initial conditions. Simulations can also be used to perform experiments that are impossible to make in reality, such as studying realms of reality which are not physically accessible. However, it is perfectly plausible to have simulations that are made without any experimental purpose in mind (for example, think of simulations adopted for teaching purposes).

The use of computer simulations in science calls for the epistemological justification of their results. Usually the reasons for trusting simulations are two-sided: either the models used in simulations are strongly grounded in well-founded theories or there are experimental data against which simulation results can be checked. This is not always possible when simulations are used as forms of *explorative experiments*, where neither well-founded theories nor experimental data are present, and where the main sources of credibility for such simulations seem to be: the prior successes of the

model-building techniques adopted, the production of outcomes fitting well with previously accepted data, observations, and intuitions, and the capability of making successful predictions and of producing practical accomplishments.

In autonomous robotics, the use of simulations for experimental purposes has grown along the years (see the most recent proceedings of the conference series on *Simulation, Modeling, and Programming for Autonomous Robots* [36.37] for an overview of the different applications). Several papers today present only experiments performed in simulation to validate a proposed robot system. Simulations provide a convenient way to explore different robotic scenarios at lower cost with respect to using real robots. They reduce the effort needed for writing software programs, debugging them, and displaying results. They also often run much faster than their real counterparts. As we have seen, a simulation needs a dynamic model of the system it reproduces. In the case of autonomous mobile robotics, the system that is reproduced is a robot that acts in an environment. The dynamic model must therefore include a representation of the robot and a representation of its interaction with the environment. Let us detail the elements involved in these two representations in the case of mobile robots. Roughly speaking, a mobile robot is modeled by representing its locomotion, sensing, and control subsystems. The interaction of a mobile robot with an environment is a complex issue. For example, it involves a model describing the behavior of the robot in the environment after the control subsystem issued a command. If the command is *go forward* 50 cm, the actual movement of a wheeled robot in a real environment could be more or less than half a meter because of slipping wheels, of rough terrain, of errors in the motors moving the wheels, and of several other reasons. Indeed, it is not always easy to capture this variability in a computational model. Similar problems emerge in modeling the perception of the robot in the environment. Current robotic simulations model in different ways the uncertainties on the effects of actions and on the perceptions. A first approach to uncertainty is usually based on use of physical engines (see below), while a second approach is based on artificially adding uncertainty to the data, according to different probability distributions.

Autonomy makes modeling a robot's interaction with the environment even more complicated, because these interactions are hardly predictable. This is probably one of the reasons for the late adoption of simulations in autonomous robotics. Until few years ago, the models of interaction between robots and the world were not sufficiently accurate and using simulations based on these models was simply not convenient for

the community. If a simulation is based on inaccurate models of the interaction with the world, it is not representative of the behavior of real autonomous robots and, as such, cannot be used to validate the behavior of the simulated robots and to generalize it to real robots. Nowadays, one of the most used *simulator* for autonomous mobile robots, USARSim [36.38], models the interaction of robots with the environment using a software, called Unreal engine, initially developed for a multiplayer combat-oriented first-person shooter computer game, Unreal Tournament 2004. Unreal engine contains a physical engine that accurately simulates the interaction of three-dimensional physical objects and that allows to obtain highly realistic simulations, which have been validated against real data [36.39, 40]. Resorting to components developed in the extremely competitive field of computer games is an interesting way to have state-of-the-art models of the physical interaction between objects.

Another interesting trend that is emerging is the use of publicly available data sets composed of data collected in the real world by some researchers ([36.41, 42] can be considered among the first examples of this tendency). These data sets consist in sensor data collected in real-world experimental tests of robot systems and made available to test other robot systems. From the one hand, we can think of these data sets as models of the interaction between the robots and the real environment. Using these data sets appears very similar to perform a simulation, in which the underlying model is very precise, because it exactly records the interaction of real robots with real world. According to this view, the difficulty of building a model for a simulation of how a robot perceives the environment is addressed by letting the data collected during real operations be the model. From the other hand, using publicly available data sets can be considered as a real (not simulated) experiment, in which activities of collecting and processing real-world data are performed in different places at different times. Another initiative, OpenSLAM [36.43], takes a step further and aims at sharing not only data sets, but also the code of the software systems of robots, in this case relative to the simultaneous localization and mapping problem. What is emerging here is a sort of continuum, ranging from performing completely simulated experiments, to using data sets like to performing real-world experiments.

If we consider the systematic, even if limited, analysis of the trends in using simulations for experimental purposes in autonomous robotics that we have already discussed [36.5], we can observe an increasing use of simulations over the years. Results show that the fraction of papers with experiments that use simulated

robots has increased and, in general, simulation dominates over real robots, which can be explained by the lower costs and the relatively easier operational aspects of simulation (Fig. 36.1).

It is interesting to notice that the fraction of papers presenting experiments with real robots is somehow constant over the years. This could be related to the fact that papers addressing some topics, like target tracking, are more frequently presenting experiments with real robots. A common situation is that in which extensive experiments are performed in simulation and simpler demonstrations are performed with real robots. By considering the simulation tools, with a particular focus on standard simulators as opposed to custom ones (where the former ones refer to commercial or publicly available systems, while the latter ones refer to systems that are usually not available outside the laboratory that developed them), the use of standard simulators seems to be increasing over the years, which could be related to the fact that more and more reliable simulation platforms have recently become available (Fig. 36.2).

Looking at the standard simulators used in the last years, it emerges that most of them are used in competitions like RoboCup (Table 36.1).

Despite their increasing use in this community, simulations are often criticized for not being realistic enough, and for producing results that do not transfer easily to the real world, such that until few years ago simulations were considered as *fake robotics* by most researchers. Such problems arise when simulation models are too naive, because they embed insufficient knowledge on the real system, sensors and actuators models are not carefully calibrated, real-world noise is not accounted for, and, most importantly, the correspondence between the simulated performance and that in the real world is not validated. Particularly interesting

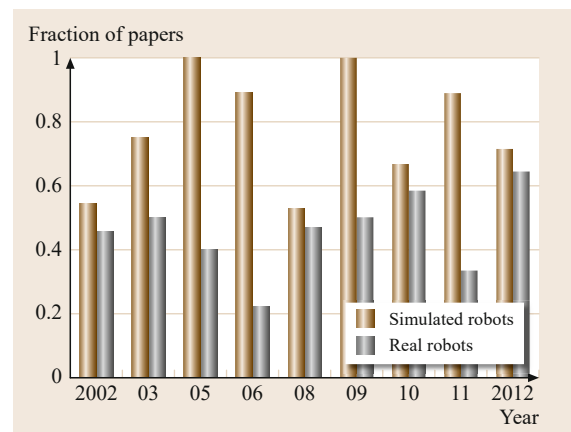
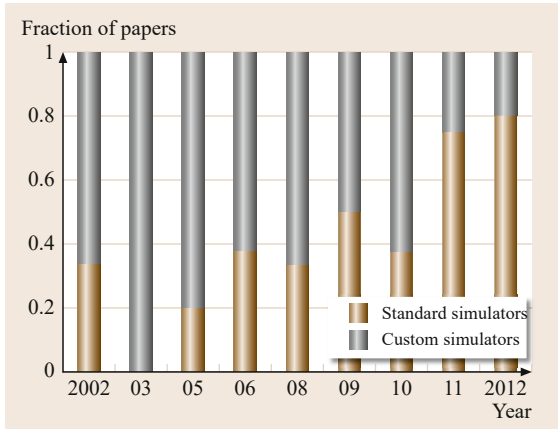


Fig. 36.1 Fraction of robotics papers at AAMAS that use simulated and real robots in experiments (after [36.5])

**Table 36.1** Standard real robot systems used in robotics papers at AAMAS (after [36.5])

Simulator	Number of papers	2002	2003	2005	2006	2008	2009	2010	2011	2012
(Player/) Stage or Gazebo	4				1				1	2
RoboCup simulators (e.g., USARSim)	6				1	1		1	2	1
Cyberbotics' Webots	6	2			1	1		1	1	

**Fig. 36.2** Fraction of robotics papers at AAMAS that employ standard and custom simulators (after [36.5])

for this discussion, as it is representative of many open issues, is a post about simulations as a research tool in robotics recently appeared on *Winfield's* blog [36.44]. Besides the provocative opening (“If you write papers with results based on simulation and submit them for peer-review, then be warned: if I should review your

paper then I will probably recommend it is rejected.”), a list of insightful criticisms is discussed. It is claimed that papers presenting results based on simulation have the following problems, not always related to the very nature of simulations but sometimes to the presentation of results: the lack of details on the simulation tools used, the lack of validation in the case of custom built simulators, the fact that the robots used to test the algorithms are not specified (real robots or abstractions of robots?), the lack of information on how the robot is modeled in the simulator, the lack of a serious analysis of the limits of these simulation tools (the so-called reality gap), the necessity to provide access to all of code in order to make others repeat the same results. The post concludes by equating an engineering simulation to any scientific instrument [36.45] and with the invitation to treat it accordingly, namely to fit for purpose, to be set-up and calibrated for the task at hand, and to be understood – especially in its limitations. These discussions, far from providing definite answers, show how the simulation topic is under attention today and give the feeling of the evolution not only of the tools, but also of the methodological analysis required to use them in an appropriate way.

## 36.5 Benchmarking and Standards

Devising rigorous evaluation procedures to assess the capabilities, reliability, dependability, and performance of robot systems in precisely defined settings is often referred to as *benchmarking*. Objective evaluation of robot systems is a need for their scientific and technical evolution, for their industrial employment, and for their market positioning. Not surprisingly, benchmarking is strictly related to the development of *standards* for robot systems. This section surveys the efforts in these two areas, keeping the discussion at a general level without attempting to cover all the details of the several benchmarking and standard activities relative to specific areas of robotics.

As already discussed, an early attempt to set up a methodological framework that accounts for the relationships between controlled experiments and benchmarks for artificial agents (of which robots constitute

a significant class) is reported in [36.24]. In this framework, benchmarks are intended to involve “precisely defined, standard tasks” [36.24, p. 17] and, ideally, “problems that are both amenable to precise analysis and representative of a more complex and sophisticated reality” [36.24, p. 19] and that “cannot depend on any system-specific details, nor can the scoring system” [36.24, p. 19]. The goal of benchmarks is then to tell “us something we want to know about the behavior” [36.24, p. 18] of an agent, for which a model of the task is required, especially “when we design benchmarks to be failed, but in this case, we need a model of the factors that make the benchmark difficult” [36.24, p. 19]. On the other hand, controlled experimentation is seen as a different activity, in which [36.24, p. 17] “a researcher varies the features of a system or the environment in which it is embedded and measures the effects of these



variations on aspects of system performance” arguably resulting in more generalizable significant results with respect to benchmarks.

From 1999, significant efforts to promote benchmarks in robotics have been put in place within the EU-funded European Robotics Network (EURON). EURON has performed a number of actions with the goal not only to define specific benchmarks, but also to propose and encourage the acceptance of benchmarks in the community, which culminated in the First European Workshop on Benchmarks in Robotics Research and in the IROS Workshop on Benchmarks in Robotics Research, both held in 2006 [36.46]. Starting from these seminal events, a series of similar workshops have blossomed with the organization and supervision of the EURON GEM Special Interest Group [36.26]. The activities performed in EURON include the proposal of a list of features that a benchmark should have [36.47]: it has to be defined for a *task* valuable in the real world, it has to be *standard* and *precisely defined*, it has to be associated to some *performance metrics*, it has to be *repeatable*, *independent*, and *unambiguous*, and it has to be widely *accepted*. Another contribution from EURON is the advancement of the idea that competitions could serve as benchmarks for “comparing the performance of competing systems by means of very well-defined rules and metrics” [36.46, p. 4]. *del Po-bil* [36.46] also presents an exhaustive survey of the state of the art of the efforts in comparative research, including competitions, benchmarks, challenges, and relevant conferences.

Along the direction traced by EURON, some other projects have been funded by the European Union and have aimed at defining benchmarks for specific robotics domains. For example, EUROP (EUropean RObotic Platform) supported activities on standards, RoSta [36.48] defined a standard benchmark for mobile service robots, and Rawseeds [36.42] produced and made publicly available high-quality data sets collected by robots in real environments, corresponding ground truth data, and benchmark problems on which different algorithms for robot self-localization and mapping can be compared.

In the United States, the National Institute of Standards and Technology (NIST) has actively promoted competitions and field exercises as “two different yet effective ways of systematically evaluating the performance of robotic systems” [36.49, p. 2]. For example, NIST has worked toward the development of performance metrics and standards for robots employed in urban search and rescue applications, taking care “not to explicitly test for particular technological solutions; rather, the tests measure how effectively or efficiently a robot can complete certain tasks, without assuming

a particular approach” [36.49, p. 3]. In this way, robot capabilities can be evaluated without inhibiting innovations of robot developers. Such standard test methods [36.50] include detailed specifications of tasks that users expect the robot to perform reliably, of scripts for the test administrator and the robot operator to follow, and of quantitative ways to measure the performance of the robot (Fig. 36.3). In parallel to the definition of such standard benchmarking procedures, the PerMIS (Performance Metrics for Intelligent Systems) workshop series has been established as one of the main venues for discussing and disseminating the definition of methodologies to measure the performance of robots in different settings, from industrial to military applications [36.51].

As a matter of fact, practical use of benchmarks (developed by independent organizations) is still limited for most robot applications, notwithstanding several proposals have been advanced for benchmarking robot-specific capabilities, for instance path planning [36.52], navigation [36.53], simultaneous localization and mapping [36.54], and object manipulation [36.55].

Benchmarks have also been proposed to comparatively evaluate robotics software and architectures, as for example in [36.56]. How to move from benchmarking basic robot capabilities to benchmarking more complex cognitive activities is not yet well understood paralleling, in a sense, the difficulty in choosing human intelligence benchmarks. A possibility [36.57] seems to be evaluating a whole robot system (as opposite to evaluating some of its components) while executing a task

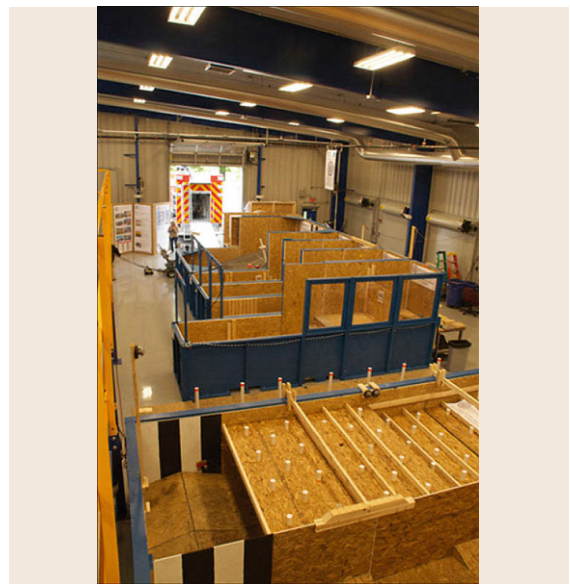


Fig. 36.3 NIST Robot Test Facility (photo by NIST)

which is intended to be difficult and to require the integration of different components (e.g., foraging and survival require navigation, image processing, learning, ...). Other two proposals for developing benchmarks for cognitive abilities of robots consist in a collection of individual tasks that, although individually limited in breadth, collectively cover a broad spectrum [36.58] and in setting up a Total Turing Test in which a human interrogator should distinguish the performance of an autonomous robot from that of a tele-operated robot in performing a given task [36.59]. In this context, tasks intended to test the cognitive abilities of robots are sometimes embedded in competitions (as discussed later).

Some negative effects of benchmarks have been reported in other technology domains. For example, to overcome some limitations of computer processors early benchmarks, designed to measure millions of instructions per second (MIPS) and millions of floating point operations per second (MFLOPS), benchmarks evaluating performance of processors using real-world applications were later introduced, but producers then developed specific optimizations in the attempt to perform well on benchmarks [36.52]. Knowing these shortcomings and building on experiences from other domains [36.60] should be considered to avoid the same errors in developing robot benchmarks.

Standardization activities have been performed with the aim of defining methods for the quantitative measurement of the performance of robot systems in order to keep a global quality culture, to set minimum require-

ment levels, to reduce confusion of consumers when developing innovative robots for specific domains (like manufacturing, floor cleaning, ...), and to help scientific research. Unsurprisingly, standard activities have been mostly related to *security* of industrial robots. For example, ISO (International Organization for Standardization) and IEC (International Electrotechnical Commission) have issued standards for safety of industrial and personal care robots ([36.61] for a survey of these and other standard activities within ISO). Other safety standards have been promoted by RIA (Robotic Industries Association) and ANSI (American National Standards Institute). Another focus for standards is the vocabulary. For example, IEEE (Institute of Electrical and Electronics Engineers) has developed a formal reference vocabulary for communicating knowledge about robotics and automation both among robots and between robots and humans [36.62]. Finally, standards for interoperability in robot software development [36.63] and in robot operations, like navigation [36.62], have been proposed.

Overall, benchmarking and standardizing complex systems like robots pose a number of problems [36.49]. For example, developing standards for robot capabilities (like navigation and self-localization) appear to be of a different, and more difficult, nature than developing standards for other devices and technologies (like the Ethernet and other network technologies). Moreover, standards that regulate the interaction between different subsystems usually result in being too *thick*, becoming hard to apply.

## 36.6 Competitions and Challenges

Robot competitions and challenges have flourished since the 1970s, now counting dozens of events per year. From the beginning, it has been recognized that competitions can play several roles in robotics, for example to promote education and research, to push the field forward, to entertain general audience, and to build community [36.64–67]. These roles are often conflicting and balancing them in devising a robot competition has been proven difficult [36.68]. After some early recommendations about being “careful not to confuse a competition with research” [36.69, p. 39], a more recent trend advocates for recasting robotics challenges and competitions as experiments, recognizing that [36.70, p. 10]:

“[c]hallenge and competition events in robotics provide an excellent vehicle for advancing the state of the art and evaluating new algorithms and tech-

niques in the context of a common problem domain.”

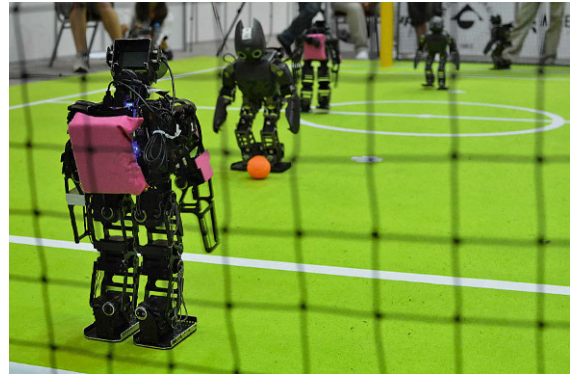
Along the same line, it has also been proposed to use competitions as benchmarks, since they provide standardized test beds that are largely independent of the settings roboticists usually experience in their laboratories and allow for direct comparison of approaches for solving a task [36.71]. However, in defining good benchmarks, designers of competitions should avoid to encourage ad hoc solutions instead of more adaptable and flexible approaches [36.68]. This section discusses in some depth the advancements and the results obtained adopting the perspective that views competitions as sort of experiments.

From a broad perspective, competitions and challenges share some similarities with experiments. Robots usually compete in precisely definite settings

and are scored according to precise performance measures, which parallel, at least to some degree, the controlled conditions and measures of experiments. However, competitions and challenges often evaluate whole robot systems, while experiments presented in the literature mainly evaluate a single robot ability or subsystem. Moreover, some challenges are intended to be performed just once, while experiments aims at being reproducible (Table 36.2). More generally, robotics competitions and challenges usually evaluate general abilities of robot systems and push to development of solutions, while experiments evaluate specific hypotheses, explore phenomena, and share result.

A robotics competition (challenge) usually involves some robots, a dynamic, but rather controlled, environment, a clear measure of success, and rules for calculating scores [36.72–74]. Without attempting at providing any exhaustive survey and detailed account of how specific events are organized and run, in the following we discuss some robotics challenges and competitions in order to trying to assess their potential role as experiments.

One of the best known examples is RoboCup [36.75], which is taking place since 1996 and aims at providing “a standard problem so that various theories, algorithms, and architectures can be evaluated. Computer chess is a typical example of such a standard problem.” RoboCup features robotic soccer (Fig. 36.4), rescue, and home competitions, in which robots compete in dynamic unpredictable environments with real-time constraints. Competitions take place both in the real world and in simulation. The environments are precisely defined and can be easily reproduced in different places. However, this is not true for other elements that characterize the competition, like the opponent team and the light and noise conditions. In the soccer competitions, the measures and the criteria according to which two robot systems (teams) are compared are clearly defined only for the purposes of the game. For example, it is difficult to draw any conclusion about the general behavior of robots and their components from the fact that one team won, say, 2 : 0 against another team. In other leagues, like in RoboCup@Home, some attempts have been recently made toward more solid procedures to benchmark and track the progress of robots in per-



**Fig. 36.4** Humanoid Kid Size League match at RoboCup 2011 (photo by Viktor Orekhov, <http://www.robocup2013.org/press/>)

forming tasks (that are possibly changed over time to keep the competitions challenging) [36.76].

The DARPA Robotic Challenge (DRC) [36.77] consists of tasks related to human assistance in responding to disasters, “[i]t was designed to be extremely difficult.” Tasks are related to the development of autonomous humanoid-like robots able to operate in hazardous settings. During the DRC trials, in December 2013 (Fig. 36.5), the main scoring mechanism has been task completion (e.g., number of open valves), while time has not been a factor (taking 30 s and 30 min to complete a task is worth the same amount of points), but is used as tiebreaker. The goal of the DRC trials, beyond selecting teams that will advance to the finals, has been to set up “a baseline on the current state of robotics.” In the DRC finals, that took place in June 2015, speed weighted more in the score.

The interest in designing competitions as experiments has also resulted in two projects funded by the European Union under the FP7 Challenge 2 *Cognitive Systems and Robotics*. euRathlon [36.78] is an outdoor competition for robots involved in emergency-response scenarios. In the words of their organizers:

“euRathlon aims to be an important milestone for robotics research. Not only will it provide opportunities to put the latest robotics technologies to test under realistic disaster scenario conditions, we also recognise that there is currently an urgent need to develop useful benchmarks that will advance the field of robotics.”

The 2013 competition has been for ground robots, the 2014 competition for underwater robots (Fig. 36.6), and the 2015 competition required a “team of terrestrial, marine, and aerial robots to work collaboratively to survey the scene, collect environmental data and identify

**Table 36.2** Competitions versus experiments

Competitions	Experiments	
Precisely defined settings	Controlled conditions	✓
Performance measure	Measures	✓
Involve whole robot systems	Evaluate single robot abilities	✗
Sometimes performed just one time	Tend to be reproducible	✗



**Fig. 36.5** A test trial of the DARPA Robotic Challenge (DARPA)

critical hazards.” The settings in which the competitions take place are precisely defined to represent mock emergency-response scenarios, including conditions like limited visibility and salty water. Scores of the competitions have been a mix of measured quantities and subjective judgments given by a human Judging Team. Data sets recorded during the competition are made publicly available to the community as “a valuable pool for benchmarking, testing and comparison”.

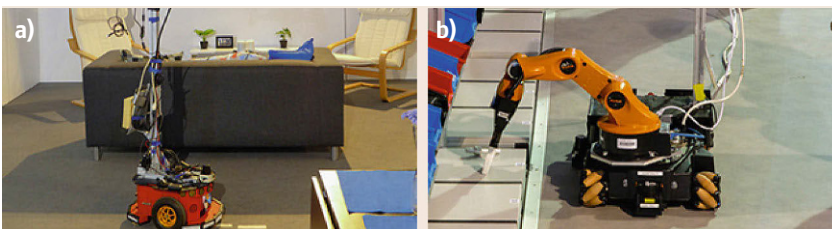
The other EU-funded project is RoCKIn [36.79] that addresses domestic (RoCKIn@Home) and industrial (RoCKIn@Work) environments, focusing on autonomous service and industrial robots, respectively



**Fig. 36.6** euRathlon 2014 competition (<http://www.eurathlon.eu/>)

(Fig. 36.7). Explicit emphasis is put on “assess, compare, and evaluate competing approaches” exploiting “benchmarking procedures and good experimental methods” making the “integration of benchmarking technology with the competition concept [...] one of the main objectives of RoCKIn.” In both @Home and @Work competitions, rules precisely define the settings for the competitions. One of the distinctive features of RoCKIn is in its scoring system, which is based on the presence of two classes of benchmarks, called task benchmarks and functionality benchmarks, somehow following the approaches proposed in [36.47] and in [36.80]. The first ones are devoted to evaluating the performance of integrated robot systems, while the second ones focus on the performance of specific subsystems (like object recognition and localization). A *task benchmark* deals with complete robot systems, implying that a large set of interacting robot elements are examined together at the same time. *Functionality benchmarks* define a precise setup in which a single robot functionality can be evaluated. Such evaluation is performed according to well-specified quantitative measures and criteria, specific for the functionality under test. Also RoCKIn makes data collected during competitions available to the community.

From the above picture, it emerges that the question whether robotics competitions and challenges are



**Fig. 36.7a,b** RoCKIn@Home (a) and RoCKIn@Work (b) competitions (copyright by RoCKIn project)

experiments is far from being satisfactorily answered. However, an interesting attempt to provide a partial answer could come from considering different competitions and challenges as different forms of experimentation, for example starting from the taxonomies of experiments performed in computing and in AI as presented in [36.81] and [36.80]. For instance, referring to the classification of [36.81], RoboCup soccer competitions seem to be between *feasibility experiments*, that are basically a form of empirical demonstration, intended as an existence of proof of the ability to build a tool or a system, and *trial experiments*, which take a step further and evaluate various aspects of systems using some predefined variables which are often measured in laboratories, but can occur also in real contexts of use (with some limitations). The DARPA Robotics Challenge and euRathlon support the idea of *field ex-*

*periments*, which are similar to trial experiments, but take place outside the laboratory in complex sociotechnical contexts of use. Finally, the RoCKIn competitions seem to aim at moving toward *comparison experiments*, which refer to comparing different solutions in some setup and based on some precisely defined measures and criteria to assess the performance. Overall, no current competition fulfills the requirements for the *controlled experiments*, namely the golden standard of experimentation in the natural sciences that refers to the original idea of experiment as controlled experience, where the activity of rigorously controlling (by implementing experimental principles such as reproducibility or repeatability) the factors that are under investigation is central, while eliminating the confounding factors, and allowing for generalization and prediction.

## 36.7 Conclusions

In this chapter, we have discussed the issues involved in the experimental assessment of computational models embedded in autonomous robots. These are particularly challenging because of their interaction with the real world. The current debate on the development of good experimental methodologies in this field has been our starting point, together with the acknowledgment that robotics, at least from a methodological point of view, lies in between science and engineering. This is also the reason why, besides the traditional notion of epistemic experiment, also that of directly action-guiding experiment, as recently conceptualized in the philosophy of science and technology [36.3], is relevant. This field is still very open and, thus, it is impossible to present a coherent and systematic view. For this reason, we have chosen to present a number of topics and initiatives that are currently under discus-

sion and various and profitable interconnections among them. In doing so, we have tried to give an overview of the main trends, but at the same time we have pointed out many criticalities and some promising directions.

Besides the undeniable role for the development of the disciplinary status of autonomous robotics and its methodological maturity, we believe that this debate has the potential to constitute a possible starting point to revise and enlarge traditional model-based science [36.82], intended as a way to representing and understanding the world, toward model-based engineering, where specific scientific and technological practices are taken into account and where modeling practices are justified mostly from a pragmatic point of view in a methodological context that privileges contexts and purposes over representations [36.83].

## References

- 36.1 M. Scheutz: Computation, philosophical issues about. In: *The Encyclopedia of Cognitive Science*, ed. by L. Nadel (Wiley, Hoboken 2002)
- 36.2 P. Vermaas, P. Kroes, I. van de Poel, M. Franssen, W. Houkes: A philosophy of technology. In: *From Technical Artefacts to Sociotechnical Systems*, (Morgan Claypool, San Rafael 2011)
- 36.3 S.O. Hansson: Experiments before science? – What science learned from technological experiments. In: *The Role of Technology in Science. Philosophical Perspectives*, ed. by S.O. Hansson (Springer, Dordrecht 2015)
- 36.4 F. Amigoni, M. Reggiani, V. Schiaffonati: An insightful comparison between experiments in mobile robotics and in science, *Auton. Robot.* **27**(4), 313–325 (2009)
- 36.5 F. Amigoni, V. Schiaffonati, M. Verdicchio: Good experimental methodologies for autonomous robotics: From theory to practice. In: *Methods and Experimental Techniques in Computer Engineering*, Springer Briefs in Applied Sciences and Technology, ed. by F. Amigoni, V. Schiaffonati (Springer, Cham 2014) pp. 37–53

- 36.6 A. Newell, H. Simon: Computer science as empirical inquiry: Symbols and search, *Commun. ACM* **19**(3), 113–126 (1976)
- 36.7 J.A. Feldman, W.R. Sutherland: Rejuvenating experimental computer science, *Commun. ACM* **22**(9), 497–502 (1979)
- 36.8 P.J. Denning: What is experimental computer science?, *Commun. ACM* **23**(10), 543–544 (1980)
- 36.9 W. Tichy: Should computer scientists experiment more?, *Computer* **31**(5), 32–40 (1998)
- 36.10 P. Freeman: Back to experimentation, *Commun. ACM* **51**(1), 21–22 (2008)
- 36.11 P. Denning, P. Freeman: Computing’s paradigm, *Commun. ACM* **52**(12), 28–30 (2005)
- 36.12 C. Morrison, R. Snodgrass: Computer science can use more science, *Commun. ACM* **54**(6), 38–43 (2011)
- 36.13 P.J. Denning: Is computer science science?, *Commun. ACM* **48**(4), 27–31 (2005)
- 36.14 D.G. Feitelson: Experimental computer science: The need for a cultural change (2006), manuscript, <http://www.cs.huji.ac.il/~feit/papers/exp05.pdf>, last accessed August 2016
- 36.15 M.V. Zelkowitz, D.R. Wallace: Experimental validation in software engineering, *Inf. Softw. Technol.* **39**(11), 735–743 (1997)
- 36.16 M.V. Zelkowitz, D.R. Wallace: Experimental models for validating technology, *Computer* **31**(5), 23–31 (1998)
- 36.17 W. Harrison, V.R. Basili: Editorial, *Empir. Softw. Eng.* **1**(1), 5–10 (1996)
- 36.18 M. Barni, F. Perez-Gonzalez, P. Comesana, G. Bartoli: Putting reproducible signal processing into practice: A case study in watermarking, *Proc. IEEE ICASSP* **4**, 1261–1264 (2007)
- 36.19 P. Vandewalle, J. Kovacevic, M. Vetterli: Reproducible research in signal processing, *IEEE Signal Process. Mag.* **26**(3), 37–47 (2009)
- 36.20 B. Mayer, M. Nordio: *Empirical Software Engineering and Verification*, Programming and Software Engineering, Vol. 7007 (Springer, Berlin, Heidelberg 2010)
- 36.21 N. Juristo, O. Gomez: Replication of software engineering experiments, *Lect. Notes Comput. Sci.* **7007**, 60–88 (2012)
- 36.22 M. Barni, F. Perez-Gonzalez: Pushing science into signal processing, *IEEE Signal Process. Mag.* **120**, 119–120 (2005)
- 36.23 C. Drummond: Replicability is not reproducibility: Nor is it good science, *Proc. Eval. Methods Mach. Learn. Workshop 26th Int. Conf. Mach. Learn.* (2009)
- 36.24 S. Hanks, M. Pollack, P. Cohen: Benchmarks, test beds, controlled experimentation, and the design of agent architectures, *AI Magazine* **14**(4), 17–42 (1993)
- 36.25 International Foundation of Robotics Research: International Symposium on Experimental Robotics, <http://www.ifrr.org/iser.php>, last accessed August 2016
- 36.26 F. Bonsignorio, J. Hallam, A. del Pobil, Special Interest Group on Good Experimental Methodology: GEM Guidelines, <http://www.heronrobots.com/EuronGEMSig/downloads/GemSigGuidelinesBeta.pdf>, last accessed August 2016
- 36.27 F. Bonsignorio, J. Hallam, del Defining the requirements of a replicable robotics experiment, *RSS2009 Workshop Good Exp. Methodol. Robot.* 2009 (2009)
- 36.28 IEEE Robotics and Automation Magazine special issue on Replicable and Measurable Robotics Research (2015)
- 36.29 L. Takayama: Toward a science of robotics: Goals and standards for experimental research, *Robot.: Sci. Syst.* (RSS) Workshop Good Exp. Methodol. Robot., Seattle (2009)
- 36.30 M. Caccia, E. Saggini, M. Bibuli, G. Bruzzone, E. Zereik, E. Riccomagno: Towards good experimental methodologies for unmanned marine vehicles, *Lect. Notes Comput. Sci.* **8112**, 365–372 (2013)
- 36.31 F. Lier, J. Wienke, A. Nordmann, S. Wachsmuth, S. Wrede: The cognitive interaction toolkit – Improving reproducibility of robotic systems experiments, *Lect. Notes Artif. Intell.* **8810**, 400–411 (2014)
- 36.32 A. Tanoto, J. Gomez, N. Mavridis, L. Hanyu, U. Rückert, S. Garrido: Teletesting: Path planning experimentation and benchmarking in the Teleworkbench, *Proc. Eur. Conf. Mob. Robot. (ECMR)* (2013) pp. 343–348
- 36.33 S. Hartmann: The world as a process: Simulations in the natural and social sciences. In: *Simulation and Modeling in the Social Sciences from the Philosophy of Science Point of View*, ed. by R. Hegselmann, U. Mueller, K.G. Troitzsch (Kluwer, Dordrecht 1996) pp. 77–100
- 36.34 R. Frigg, S. Hartmann: Models in science. In: *The Stanford Encyclopedia of Philosophy*, ed. by E.N. Zalta (Stanford Univ., Stanford 2012), <http://plato.stanford.edu/entries/models-science/>
- 36.35 E. Winsberg: *Science in the Age of Computer Simulations* (Univ. Chicago Press, Chicago, London 2010)
- 36.36 E. Winsberg: Computer simulations in science. In: *The Stanford Encyclopedia of Philosophy*, ed. by E.N. Zalta (Stanford Univ., Stanford 2015), <http://plato.stanford.edu/entries/simulations-science/>
- 36.37 D. Brugali, J. Broenink, T. Kroeger, B. MacDonald (Eds.): *Simulation, Modeling, and Programming for Autonomous Robots*, Lecture Notes in Artificial Intelligence, Vol. 8810 (Springer, Berlin, Heidelberg 2014)
- 36.38 USARSim (2010) <http://usarsim.sourceforge.net>, last accessed April 2015
- 36.39 S. Carpin, T. Stoyanov, Y. Nevatia, M. Lewis, J. Wang: Quantitative assessments of USARSim accuracy, *Proc. PerMIS* (2006) pp. 111–118
- 36.40 B. Balaguer, S. Balakirsky, S. Carpin, M. Lewis, C. Scrapper: USARSim: A validated simulator for research in robotics and automation, *IEEE/RSJ IROS 2008 Workshop Robot Simulators* (2008), <http://www.robot.uji.es/research/events/iros08/contributions/carpin.pdf>, last accessed August 2016
- 36.41 A. Howard, N. Roy: The robotics data set repository, radish (2003), <http://radish.sourceforge.net/>, last accessed February 2015
- 36.42 Rawseeds: The Rawseeds Project, <http://www.rawseeds.org/home/>, last accessed March 2015

- 36.43 OpenSLAM: OpenSLAM, <http://openslam.org/>, last accessed April 2015
- 36.44 A. Winfield: Robot simulators and why I will probably reject your paper, Alan Winfield's Web Log, November 30th 2014, <http://alanwinfield.blogspot.it/>, last accessed August 2016
- 36.45 J. Bown, P.S. Andrews, Y. Deeni, A. Goltsov, M. Idowu, F.A. Polack, A.T. Sampson, M. Shovman, S. Stepney: Engineering simulations for cancer systems biology, *Curr. Drug Targets* **13**(12), 1560–1574 (2012)
- 36.46 A. del Pobil: Research benchmarks V2, EURON European Robotics Network (2006) <http://www.euron.org/miscdocs/docs/euron2/year2/dr-2-3-benchmarks.pdf>, last accessed August 2016
- 36.47 R. Dillmann: KA 1.10, Benchmarks for robotics research, EURON European Robotics Network (2004) <http://www.cas.kth.se/euron/euron-deliverables/ka1-10-benchmarking.pdf>, last accessed August 2016
- 36.48 RoSta: Robot Standards and Reference Architectures, <http://www.robot-standards.eu/>, last accessed August 2016
- 36.49 R. Madhavan, R. Lakaemper, T. Kalmar-Nagy: Benchmarking and standardization of intelligent robotic systems, *Proc. ICAR* (2009) pp. 1–7
- 36.50 NIST: NIST Standard Test Methods for Response Robots, <http://www.nist.gov/el/isd/ms/robottestmethods.cfm>, last accessed August 2016
- 36.51 PerMIS, Performance Metrics for Intelligent Systems Workshop, <http://www.nist.gov/el/isd/ks/permis.cfm>, last accessed August 2016
- 36.52 J. Baltes: A benchmark suite for mobile robots, *Proc. IROS* **12**, 1101–1106 (2000)
- 36.53 C. Sprunk, J. Roewekaemper, G. Parent, L. Spinello, G.D. Tipaldi, W. Burgard, M. Jalobeanu: An experimental protocol for benchmarking robotic indoor navigation, *Proc. Int. Symp. Exp. Robot. (ISER)* (2014)
- 36.54 R. Madhavan, C. Scrapper, A. Kleiner: guest editorial: Special issue on characterizing mobile robot localization and mapping, *Auton. Robot.* **27**(4), 309–311 (2009)
- 36.55 B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, A. Dollar: The YCB object and model set, <http://rll.eecs.berkeley.edu/ycb/>, last accessed August 2016
- 36.56 A. Shakhimardanov, E. Prassler: Comparative evaluation of robotic software integration systems: A case study, *Proc. IROS* (2007) pp. 3031–3037
- 36.57 O. Michel, F. Rohrer, Y. Bourquin: Rat's life A cognitive robotics benchmark, *Proc. Eur. Robot. Symp.* (2008) pp. 223–232
- 36.58 B. Rohrer: Accelerating progress in artificial general intelligence: Choosing a benchmark for natural world interaction, *J. Artif. Gen. Intell.* **2**(1), 1–28 (2011)
- 36.59 M. Zillich: My robot is smarter than your robot – On the need for a total Turing test for robots, *Proc. AISB/IACAP Symp. Revisiting Turing his Test Compr., Qualia, Real World* (2012) pp. 12–15
- 36.60 A. Torralba, A. Efros: Unbiased look at dataset bias, *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)* (2011) pp. 1521–1528
- 36.61 G. Virk, S. Cameron: ISO-IEC standardization efforts in robotics, *Proc. IROS Workshop Stand. Knowl. Represent. Ontol. Robot. Autom.* (2014) pp. 5–6
- 36.62 R. Madhavan, W. Yu, C. Schlenoff, E. Prestes, F. Amigoni: Draft standards development of two working groups, *IEEE Robot. Autom. Mag.* **21**(3), 20–23 (2014)
- 36.63 OMG: Robotics Domain Task Force, <http://robotics.omg.org/>, last accessed August 2016
- 36.64 P. Bonasso, T. Dean: A retrospective of the AAAI robot competitions, *AI Magazine* **18**(1), 11–23 (1997)
- 36.65 R. Murphy: Using robot competitions to promote intellectual development, *AI Magazine* **21**(1), 77–90 (2000)
- 36.66 J. Casper, M. Micire, J. Hyams, R. Murphy: A case study of how mobile robot competitions promote future research, *Lect. Notes Artif. Intell.* **2377**, 123–132 (2002), *RoboCup* 2001
- 36.67 M. Chew, S. Demidenko, C. Messom, G. Sen Gupta: Robotics competitions in engineering education, *Proc. 4th Int. Conf. Auton. Robot. Agents* (2009) pp. 624–627
- 36.68 J. Anderson, J. Baltes, C.T. Cheng: Robotics competitions as benchmarks for AI research, *Knowl. Eng. Rev.* **26**(1), 11–17 (2011)
- 36.69 T. Bräunl: Research relevance of mobile robot competitions, *IEEE Robot. Autom. Mag.* **6**(4), 32–37 (1999)
- 36.70 M. Anderson, O. Jenkins, S. Osentoski: Recasting robotics challenges as experiments, *IEEE Robot. Autom. Mag.* **18**(2), 10–11 (2011), doi:10.1109/MRA.2011.941627
- 36.71 S. Behnke: Robot competitions – Ideal benchmarks for robotics research, *Proc. IROS Workshop Benchmarks Robot. Res.* (2006), [https://www.ais.uni-bonn.de/nimbropapers/IROS06WS\\_Benchmarks\\_Behnke.pdf](https://www.ais.uni-bonn.de/nimbropapers/IROS06WS_Benchmarks_Behnke.pdf), last accessed August 2016
- 36.72 H. Yanco: Designing metrics for comparing the performance of robotic systems in robot competitions, *Workshop Meas. Perform. Intel. Intel. Syst. (PERMIS)* (2001)
- 36.73 E. Messina, R. Madhavan, S. Balakirsky: The role of competitions in advancing intelligent systems: A practitioner's perspective, *Proc. PerMIS* (2009) pp. 105–108
- 36.74 J. Parker, J. Godoy, W. Groves, M. Gini: Issues with methods for scoring competitors in RoboCup rescue, *AAMAS Workshop Auton. Robot. Multirobot Syst.* (2014), <http://www-users.cs.umn.edu/~ginipapers/Parker2014arms.pdf>, last accessed August 2016
- 36.75 The Robocup Foundation: RoboCup, <http://www.robotcup.org>, last accessed August 2016
- 36.76 L. Iocchi, D. Holz, J. Ruiz-del-Solar, K. Sugiura, T. van der Zant: RoboCup@Home: Analysis and results of evolving competitions for domestic and service robots, *Artif. Intell.* **229**, 258–281 (2015), doi:10.1016/j.artint.2015.08.002

- 36.77 DARPA: DARPA Robotics Challenge, [https://en.wikipedia.org/wiki/DARPA\\_Robotics\\_Challenge](https://en.wikipedia.org/wiki/DARPA_Robotics_Challenge), last accessed August 2016
- 36.78 euRathlon: euRathlon—an outdoor robotics challenge for land, sea and air, <http://www.eurathlon.eu>, last accessed August 2016
- 36.79 RoCKIn, RoCKIn project, <http://rockinrobotchallenge.eu>, last accessed August 2016
- 36.80 J. Hernández-Orallo: AI evaluation: Past, present and future, <http://arxiv.org/abs/1408.6908>, last accessed August 2016
- 36.81 M. Tedre: *The Science of Computing: Shaping a Discipline* (CRC, Boca Raton 2015)
- 36.82 R.N. Giere: *Explaining Science: A Cognitive Approach* (Univ. Chicago Press, Chicago 1998)
- 36.83 A.M. Isaac: Modeling without representation, *Synthese* **190**(16), 3611–3623 (2013)



# Biorobotics

## 37. Biorobotics

Edoardo Datteri

Starting from a reflection on the various roles played by simulations in scientific research, this chapter provides an overview of the biorobotic strategy for testing mechanistic explanations of animal behavior. After briefly summarizing the history and state of the art of biorobotics, it also addresses some key epistemological and methodological issues that need to be taken into serious consideration when setting up and performing biorobotic experiments. These issues mainly concern the relationship between the biorobot and the theoretical model under investigation, the choice of criteria for comparing animal and robotic behaviors, and the pros and cons of computer versus robotic simulations.

37.1	<b>Robots as Models of Living Systems</b> .....	817
37.1.1	Data-Oriented and Model-Oriented Simulations .....	817
37.1.2	The Structure of Biorobotic Methodology .....	819
37.2	<b>A Short History of Biorobotics</b> .....	825
37.2.1	Cybernetic and Artificial Intelligence ...	825
37.2.2	Contemporary Invertebrate and Vertebrate Simulation Studies .....	826
37.3	<b>Methodological Issues</b> .....	826
37.3.1	The Epistemic Requirements of <i>Good</i> Biorobots .....	826
37.3.2	On the Meaning of Behavior .....	830
37.3.3	Robots and Their Environment: Robotic versus Computer Simulations ..	832
37.4	<b>Conclusions</b> .....	833
	<b>References</b> .....	834

### 37.1 Robots as Models of Living Systems

#### 37.1.1 Data-Oriented and Model-Oriented Simulations

Computing devices are used in contemporary biological and psychological research not only as *number-crunching* tools for the analysis of experimental data, but also to simulate theoretical models of biological and cognitive systems. Two broad classes of computer simulation studies are found in the current scientific literature. A typical example of the first class is biomolecular simulation of the behavior of ion channels [37.1]. Ion channels are proteins arranged to form pores in the membrane of cells. Under particular chemical and physiological conditions, they undergo conformational changes that permit or block the movement of particular substances across the membrane. A detailed reconstruction of these changes is difficult to obtain via current molecular imaging technologies [37.1, p. 430]:

“A variety of experimental techniques can provide information about the dynamics of proteins and other biomolecules, but they are generally limited

in their spatial and temporal resolution, and most report ensemble average properties rather than the motion of individual molecules.”

Computer simulations may be used to overcome these difficulties. The detailed molecular structure of a particular class of ion channels may be represented by a program and the system may be allowed to compute the conformational changes that these channels would undergo under various chemical conditions, according to the physical laws governing atomic interactions. Simulations of this kind therefore [37.1, p. 429]:

“serve as a computational microscope, revealing biomolecular mechanisms at spatial and temporal scales that are difficult to observe experimentally.”

Biomolecular simulations exemplify one of the roles played by computer simulations in scientific research, namely the role of *generating data* about the behavior of a system. This is the use made of computer simulations in what we will here refer to as *data-ori-*

ented simulation studies. As illustrated by the proposed example, data-oriented simulations may be particularly useful when the required data are difficult or impossible to obtain by means of conventional observational or measurement techniques: in these cases, “simulations [...] replace experiments and observations as sources of data about the world” [37.2].

Simulations are not always used for this purpose, however. In many cases, they are used to *test a theoretical model* of the target system, rather than to obtain data about it. This is a *model-oriented* use of simulations (Fig. 37.1). A distinction between these two classes of simulation studies has also been proposed by *Guala* [37.3], who defined it in the following terms:

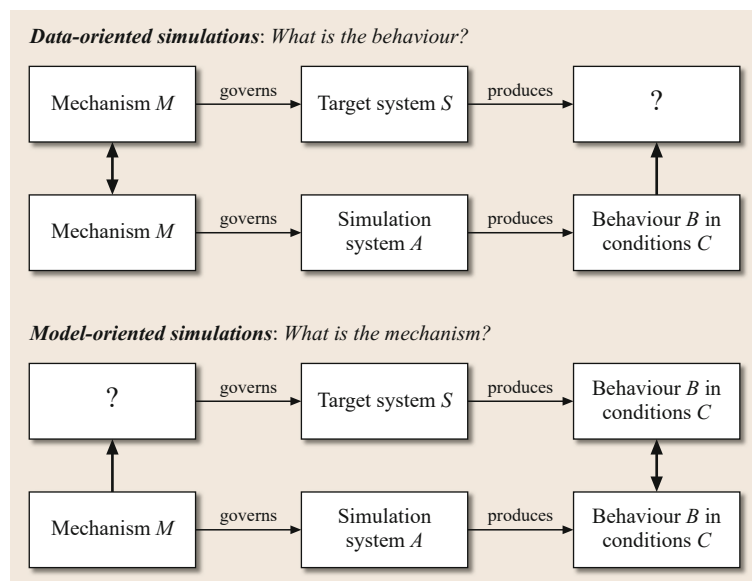
“Typically, simulations are used in one of two different ways: either (1) to bootstrap from the fact that a given effect (which we have observed in system *A*) can be produced by means of simulation *B*, to the fact that the relations governing the behavior of *B* also govern the behavior of *A*. Or (2) to argue that a certain effect observed by simulating with *B* will also be observed in the case of *A* because the two are governed by similar relations.”

Cases 1 and 2 correspond to model-oriented and data-oriented simulation studies, respectively.

In cognitive science and neuroscience, robots are often used as model-oriented simulations. For example, in so-called *biorobotic* studies, which are the main focus of this chapter, robots are used to test theoretical models of human and animal behavior. Interestingly, biorobotic, model-oriented simulation studies predate the creation of the first digital computer. What is per-

haps the first example of biorobotic simulation dates to 1913, when engineers John Hammond Jr. and Benjamin Miessner built a robot, known as the *electric dog*, which was able to track light sources by means of a simple sensory-motor mechanism. Notably, the mechanism implemented was similar in many respects to that which physiologist *Jacques Loeb* [37.4] had suggested was used by moths to seek light sources [37.5]. The fact that *by virtue of this mechanism* the robot was able to generate a light-seeking behavior – thus replicating the behavior of moths to some extent – was taken by Loeb himself as corroborating his claim that the hypothesized mechanism could also account for the light-seeking abilities of moths. The behavior of the *electric dog* robot was thus taken as a basis on which to test (and ultimately in this case, to accept) a theoretical model of insect behavior.

Since that first pioneering instance, robots have been used to test a variety of models of animal behavior: a history and a concise summary of the state of the art are provided in Sect. 37.2. One purpose of this chapter is to outline the structure of robotic model-oriented simulation experiments. It is worth noting that the expression *simulation experiments* is used here to refer to scientific experiments involving (robotic) simulation systems. It is not meant to imply that simulations are experiments in any significant sense of the term. Rather, it is here assumed that simulation systems are something that can be experimented on, given that one can assess their behavior under various experimental conditions and in response to interventions of various kinds. Moreover, it is assumed that simulation systems may be actively used as experimental tools at some step



**Fig. 37.1** Diagram illustrating the distinction between data-oriented and model-oriented simulations. As explained in the text, data-oriented simulation systems implement the mechanism governing the target system *S* (as illustrated by the *double arrow*) to find out what behavior *S* would display under conditions *C*. In model-oriented simulation experiments, matches or mismatches between the behaviors of *S* and *A* under conditions *C* are taken as a basis for testing whether the mechanism implemented in *S* coincides with the mechanism *M* implemented in *A* or not

in a procedure aimed at gathering data on, or testing a model of, a biological system (see Chap. 34 by *Imbert* in this volume and [37.6], for an analysis of whether simulations can be *literally* regarded as experiments).

Another purpose of this chapter is to raise some interesting methodological and epistemological issues biorobotics experiments give rise to, including the issue of understanding under what conditions one is justified in taking robotic behaviors as a basis for accepting or rejecting a biological or cognitive theoretical model (for other epistemological and methodological analyses of biorobotic methodology that address many of the issues discussed here, see [37.7–10]). These issues point to the existence of a gap between being able to *reproduce* a given behavior in a robotic or computer system and being able to *explain* it. Before going on to address them, and to examine the structure of biorobotic methodology in detail, it may be useful to provide a preliminary analysis of the notions of theory and explanation in cognitive science and neuroscience.

### 37.1.2 The Structure of Biorobotic Methodology

#### Behavioral Explananda

Explaining human and animal behavior is one of the primary goals of the cognitive sciences and neurosciences. The phenomena to be explained in these areas – whose linguistic formulation is often referred to as explanandum (plural explananda) in philosophical analyses of the concept of explanation – typically consist of *behavioral regularities*, that is to say, regular connections between a set of environmental conditions *C* and a behavior *B* (for the sake of brevity, in this chapter the term explanandum refers both to the statement describing the phenomenon to be explained and to the phenomenon itself). The term *C* refers to a more or less large set of environmental conditions (such as the presence of a light source, a minimum external temperature of 20°C, the absence of electromagnetic fields), while *B* stands for a more or less abstract description of the motor behavior of the target system (such as percentage success rate in hitting the light source over a given period of time, or the trajectory followed by the animal while approaching the light source). For example, Loeb's explanandum, which consists of a regular connection between the presence of a light source (*C*) and the generation of motor trajectories leading to the light (*B*), may be formulated in this way.

Cognitive and neuroscientific research often seeks to explain human and animal *capacities*, such as the ability to control arm movements in order to reach a target specified within a given sensory frame of reference (see for example [37.11]). Capacities may be reformu-

lated as behavioral regularities [37.12]: for example, the ability to reach a target with one's hand may be reformulated as a regular connection between the presence of a target (possibly plus other conditions of environmental normalcy) and the generation of movements bringing the arm endpoint into proximity of this target. Higher-level cognitive capacities may be formulated as behavioral regularities too. Contemporary research on spatial memory capacities in rats [37.13] seeks to explain, for example, why rats regularly make fewer and fewer errors in locating a reward in a maze on each successive trial, and why they make errors of a particular sort when some aspects of the environment are selectively changed. Many cognitive studies attempt to explain why humans regularly manage to solve particular classes of problems or why they show regular reactions to certain complex stimuli (for example, why it systematically takes a longer time to say the color of the word *red* written in green than to say the color of the word *green* written in green, the so-called Stroop effect, [37.14]). This is not to say that the explanation of *singular* behaviors – *why did system S display behavior B in experimental trial number 312?* – is never pursued. Indeed, that question would first be addressed by identifying the conditions *C* holding in trial number 312 and supposedly responsible for that singular (or exceptional) behavior. From that point on, the explanandum would then take the form: *why does S regularly produce B under conditions C?*

#### Behavioral Mechanisms

Explanations in the cognitive sciences and neurosciences typically include a description of a *mechanism M* supposedly responsible for a behavioral regularity *R* (for the sake of brevity, the term *M* will be used in this text to refer both to the description of a mechanism and to the mechanism itself, depending on the context). Detailed analyses of the structure of neuroscientific and cognitive mechanism descriptions are provided in [37.15–18]. For the purposes of the present discussion, it is sufficient to note that explanatory mechanism descriptions specify a number of components, characterize the behavior of each component in a more or less precise way, and define how these components interact with one other. For example, *Wolpert et al.* [37.11] explained motor control abilities in humans by describing a mechanism made up of various components (notably including an *inverse model* of the controlled object) and the mutual connections among these components (for example, by stating that some parameters of the *inverse model* component vary as a function of output from a *feedback controller* component). The behavior of each component is typically formulated as a regularity (which takes the form of an input-output regularity

in cognitive models). For example, the behavior of the *inverse model* component in [37.11] is characterized as a mapping from desired trajectories to motor commands.

Here a fundamental difference emerges between explanations in neuroscience and in cognitive science. Neuroscientific explanations use the theoretical vocabulary of the neurosciences to describe explanatory mechanisms: they define mechanisms in terms of the electrical or chemical behavior of neurons and neural areas. Cognitive science explanations describe explanatory mechanisms in terms of representations and information processing modules: the components of cognitive science mechanisms typically perform input-output mappings among representations possessed by the system (see [37.19] for a more specific analysis of the relationship between cognitive and neuroscientific mechanism descriptions).

Note that the term *theory* is often used in these fields to denote explanatory mechanism descriptions. For example, the theory on rodent navigation described in [37.20], and tested by means of computer model-based simulations, coincides with the description of a mechanism that supposedly enables rats to effect a number of spatial behaviors. The term *explanation* on the other hand is typically used to denote a structure composed of an explanandum and an explanatory mechanism description.

### Abstraction

To set the stage for the ensuing discussion of the methodological issues arising in biorobotics, it is worth reflecting on the *abstract* character of mechanistic explanations of behavior. Such explanations may be said to be abstract for at least two reasons. To introduce the first of these reasons, let us once more focus on Loeb's *explanandum*: when moths are in the proximity of a light source, they regularly generate movements leading to it. As it stands, this is a literally *false* assertion. Any generalization (i. e., a statement expressing a regularity) is strictly speaking false when it has exceptions, and moths do not *always* reach light sources. They will be unable to do that, for example, if their light receptors are damaged, if some other internal drive prevails over the light-seeking one, or if transparent glass is interposed between them and the light source. For analogous reasons, rats do not always improve their goal-seeking abilities in mazes on successive trials: the explananda addressed by contemporary studies on spatial memory in rats consist of, strictly speaking, literally false (exception-ridden) generalizations. They would be true only in the absence of a number of possibly perturbing factors, which typically are not, or only partially, specified in *C*. How then may we interpret

the claim that *S* regularly produces *B* under conditions *C*? Should one end up conceding that neuroscientists and cognitive scientists work with literally false (i. e., exception-ridden) explananda?

To adopt a more reasonable position, in line with the so-called semantic conception of scientific theories [37.21], we may say that neuroscientific and cognitive explananda state the behavior of *ideal* systems. According to this interpretation, claiming that system *S* regularly produces behavior *B* under conditions *C* amounts to making a counterfactual claim about an ideal system placed under ideal conditions: were *S* subjected *only* to conditions *C* – in other words, were *C* a correct and complete description of the conditions under which *S* stands – it would regularly display behavior *B*. Mechanism descriptions are abstract too. They clearly do not describe *all* the mechanisms found in the target system *S*, but only the mechanism *M* that is supposed to produce the behavior to be explained; and it is assumed that *M* will produce that behavior only if no perturbing condition intervenes. In other terms, claiming that system *S* exhibits *R* by virtue of mechanism *M* amounts to making a counterfactual assertion to the effect that, were *M* the *only* mechanism at work in *S*, it would produce *R*.

In sum, cognitive and neuroscientific theories are formulated to explain ideal behavioral regularities. They pursue this goal by modeling the behavior of an ideal system governed by the postulated mechanism *M* and by nothing else: hence, they are *abstract* explanations. This view is well expressed by *Frederick Suppe* [37.21, pp. 82–83]:

“The theory does not attempt to characterize the phenomena in all their complexity, but only attempts to do so in terms of a few parameters abstracted from the phenomena. For example, classical particle mechanics attempts to characterize mechanical phenomena *as if* they depended only on the abstracted position and momentum parameters. In point of fact, however, other unselected parameters usually do influence the phenomena; so the theory does not characterize the actual phenomena, but rather characterizes the contribution of the selected parameters to the actual phenomena, describing what the phenomena *would have been had* the abstracted parameters been the only parameters influencing them.”

Note that some contemporary philosophers of science forcefully claim, contrary to what I argue here, that neuroscientists and cognitive scientists address exception-ridden explananda (for example [37.15]). The *prima facie* plausibility of this hypothesis is given by the apparently exception-ridden character of most

neuroscientific *explananda* (the examples on moth locomotion and rat navigation being cases in point). The view proposed here – i. e., that neuroscientists seek to explain the behavior of ideal systems – is preferred, insofar as it is more consonant with scientific practice (see [37.22], for a discussion) and fully in line with the semantic view of scientific theories, whose virtues with respect to the so-called *received view* have been extensively emphasized in the philosophical literature. Whether the methodological analysis of biorobotic experiments proposed here may also fit with a different conception of scientific theories is a question that goes beyond the scope of the current chapter, whose aim is to provide (at the least) a plausible account of how, and under what conditions, biorobotic experiments can assist in the discovery of, and theorization on, biological and cognitive mechanisms.

### Explanation

Having analyzed the nature of biorobotic explananda and theories, it is worth focusing on the notion of *explanation*. Explaining *S*'s behavior *B* under *C* by reference to *M* amounts to making two closely related, but conceptually distinct claims. The first is that:

- 1 All systems governed only by *M* produce behavior *B* when only conditions *C* hold.

This assertion concerns what, in *Suppe*'s semantic analysis of scientific theories [37.21], is labeled the class of *theory-induced systems*, that is to say, the class of the ideal systems defined by the theory under investigation. Explaining *S*'s behavior by reference to *M*, however, implies also asserting the existence of a close relationship between *S* and the class of the theory-induced systems defined by *M*, a relationship that we express here as follows (see [37.21] for a more detailed analysis of the relationship between theory-induced systems and the systems whose behavior is to be explained):

- 2 *S* implements mechanism *M*.

This amounts to asserting that *S* has a number of components behaving and organized as specified in *M*. Of course in reality *S* will have other components and will be affected by boundary conditions not included in the ideal theory-induced systems (here we assume that the ideal theory-induced systems are *impoverished* versions of the concrete systems under investigation, obtained by abstracting out components and features that are deemed to be irrelevant for the present explanatory purposes; however, some theory-induced systems have features that are incompatible with the theory itself; see *Suppe*'s analysis of *idealization* as distinct from *abstraction* in [37.21]).

Claims 1 and 2 are conceptually independent of one another. It may be the case that *M* exhibits behavior *B* under *C* (as asserted by Claim 1), and that, at the same time, *S* does *not* implement *M* even though it displays behavior *B* under *C*: in that case, one concludes that *S* displays *B* under *C* by virtue of a different mechanism to *M*. It may also be the case that *S* actually implements *M*, but that *M* does not produce *B* under *C*: *M* will then be responsible for *other* behaviors exhibited by *S*, and will be irrelevant to explaining why *S* produces *B* under *C*. In both cases, by rejecting either Claim 1 or 2, one rejects the explanation stating that *S* produces *B* under *C* by virtue of *M*.

Distinguishing between these two claims enables us to understand how robots may be used to test cognitive and neuroscientific explanations. Indeed, to evaluate whether *S*'s behavior *B* under *C* may be explained by reference to *M*, one must have good reason to argue that *M* is implemented in *S* (Claim 2) and that *M* alone is responsible for *B* under *C* (Claim 1). As we shall see in the next section, biorobots can be especially useful for testing the latter claim. Before addressing this point, it should be noted that *dysfunctional* behaviors are typically explained mechanistically in cognitive and neuroscience. A dysfunctional behavior is one that differs, to some extent, from the behavior normally produced under particular conditions *C* by individuals belonging to a particular reference class. Explanation of these behaviors proceeds by identifying the mechanism *M* responsible for the *normal* behavior, and by finding out how *M* would need to be *damaged* to produce the dysfunctional behavior. For a thorough epistemological and methodological discussion on the explanation of dysfunctional behaviors in cognitive neuroscience, see [37.23].

### Biorobotics and the Study of Ideal Mechanisms

In the previous section, it is argued that two claims are made in asserting that *S*'s behavior *B* under *C* may be explained by reference to *M*. The first states that a system governed only by mechanism *M*, in the ideal setting in which only conditions *C* hold, would produce behavior *B* (or more concisely, that *M* can produce *B* under *C*). Testing a claim in this form only by reasoning about the structure of *M* and the contents of *C* is likely to be very difficult, even when *M* is relatively *simple*. In [37.24], a basic mechanism is proposed to explain lobsters' ability to reach the source of chemical streams in water (a behavior known as *chemotaxis*). The proposed mechanism is structurally similar to that of one of the first Braitenberg vehicles [37.25] and, incidentally, to the mechanism proposed by Loeb to explain the light-seeking abilities of moths. According to the proposed mechanism, the system has two chemical sensors

(chemoreceptors) located at the two sides of the animal. The higher the chemical concentration perceived by one receptor, the greater the movement on the opposite side of the animal: high concentration on the right side will induce rapid movement of the left motor organs, and vice versa. At first sight, this mechanism seems to be able to guarantee efficient chemotaxis. Suppose, for example, that the source of the chemical stream is located to the right of the animal, at a certain distance from it. In this circumstance, chemical concentration will probably be higher on the animal's right side. According to the proposed mechanism, the left motor organs will be moved faster than the right ones, making the system veer to the right, and thus towards the chemical source.

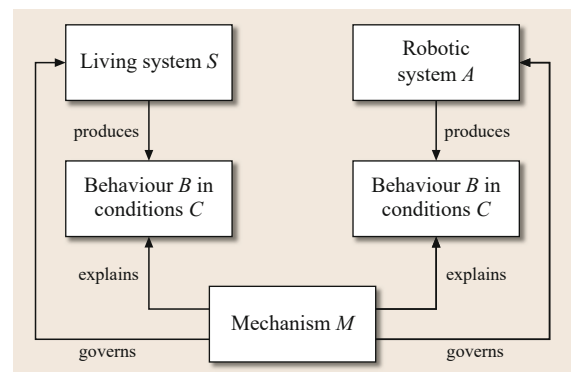
Is this argument sufficient to conclude that the proposed mechanism guarantees efficient chemotaxis? The answer largely depends on whether higher chemical concentration on the right (or left) reliably signals that the source is located on the right (or left) of the animal, a question that in turn depends on the shape of the chemical plume. If the plume is concave and highly irregular, a higher concentration will not always be a good indicator of the true position of the source. Without accurate information about the structure of the environment and the exact distribution of the relevant sensory stimuli one cannot easily predict the behavior of the system, even if the mechanism is relatively *simple*. This lesson, extensively illustrated by [37.25], has been also offered in the cybernetic era by the pioneer of electroencephalography William Grey Walter. This scientist famously built simple light-seeking devices whose sensory-motor mechanism was not very different from the mechanism later formulated by [37.24]. Yet the behavior of the devices in domestic environments turned out to be extremely difficult to predict: according to their creator, they even displayed forms of “uncertainty, randomness, free will” [37.26, p. 44]. *Herbert Simon* similarly pointed out that very simple mechanisms – whatever *simple* may mean in this context – can exhibit behaviors that are difficult to predict, if the environment is rich with unpredictable stimuli [37.27].

To sum up, whether a system governed by the mechanism described above can in fact replicate efficient chemotaxis in a realistic underwater environment is difficult to assess only by reasoning about the structure of the mechanism itself. Here biorobotic, model-oriented simulations may help. In the previous section, we claimed that mechanism descriptions in cognitive science and neuroscience may be said to be abstract for two reasons, of which so far we have discussed only one (i. e., abstraction from the numerous boundary conditions and concurrent mechanisms at work in *real* systems). The second reason concerns the *material substrate* of the target system. A fundamental assumption

of mechanistic theorizing in cognitive science and neuroscience is that whether a system governed only by  $M$  can exhibit  $R$  or not is a question that does not depend on the material the system is made of. Nothing prevents  $M$  from being implemented in both biological systems and artificial systems. Equivalently, there is no reason to deny *a priori* that any given living system and any given robotic or computer device can implement the *same* mechanism. Biorobotic methodology consists precisely of building a system  $A$  governed by  $M$  (i. e., a *robotic simulation* of  $M$ , see Fig. 37.2), and assessing – by means of an appropriately conducted *biorobotic experiment* – whether  $A$  produces behavior  $B$  under conditions  $C$ , that is to say, whether it replicates the behavior whose manifestation by  $S$  (see the left side of Fig. 37.2) is to be explained. If it does, one may be induced to conclude that the hypothesis that systems governed by  $M$  exhibit  $B$  under  $C$  is corroborated, thus providing partial support for the claim that the living system  $S$  produces  $B$  under  $C$ ; otherwise, one may be induced to reject this hypothesis.

One of the fundamental epistemological issues arising in biorobotic experimentation is easily identified. In this methodology, one draws conclusions regarding the behavior of an *ideal* theory-induced system (i. e., a system governed only by  $M$  in the ideal setting in which only conditions  $C$  hold) by analyzing the behavior of a *real* man-made system. In Sect. 37.3.1 we will reason about whether this inference may be justified, pointing out that – quite paradoxically – biorobotic experiments conducted with the aim of explaining biological behaviors must rely on the explanation of robotic behaviors, which is not always (as vividly illustrated by Braitenberg) as straightforward as it may seem.

Biorobots are thus used as *epistemic tools* to acquire information about the behavior that a system governed



**Fig. 37.2** Diagram showing the functional relationships holding among the mechanism  $M$  under investigation, the living system  $S$ , the robotic system  $A$ , and the behavior to be explained, in biorobotic studies on animal behavior

by  $M$  would exhibit under the ideal conditions  $C$ , that is to say, to explore the behavioral implications of  $M$ . Such a tool may be useful when it is difficult to arrive at these implications only by considering the structure of  $M$ , as illustrated by the study on lobster navigation. Indeed, in that study, the robotic simulation of the mechanistic hypothesis formulated by the authors turned out not to be able to perform efficient chemotaxis. After a number of control experiments aimed at excluding a particular class of artifacts due to the structure of the robot (discussed in Sect. 37.3.1), the authors concluded that the proposed mechanism *could not* account for the chemotaxis behavior, therefore rejecting Claim 1. The authors' diagnosis was that, as discussed above, higher chemical concentrations are not reliable indicators of the direction of the chemical source, due to the turbulence of the water and to the consequently irregular shape of the chemical stream.

This example helps to draw out even further the distinction between data-oriented and model-oriented simulations made in Sect. 37.1.1. In biorobotics, the robotic simulation  $A$  is used to ascertain the behavioral implications of the mechanism  $M$  supposedly implemented in  $S$  under conditions  $C$ . This is also true of the data-oriented biochemical simulations mentioned above: they are used to acquire information about the behavior that ion channels would display in some circumstances *according to* a hypothesis about their structure and to the best available theories on atomic interactions. The output of data-oriented simulations is therefore an implication of these theories. However, the theories underpinning the biomolecular simulation have *already* been accepted as the best current theories on the matter, and it is for this reason that their implications – outputted by the simulation – are interpreted as the behavior that the target system would display in the simulated circumstances. In biorobotic model-oriented studies, on the contrary, the underlying mechanistic models are *not* accepted at first, and their implications – analyzed against the data available about the target system – are taken as a basis for evaluating their plausibility. Mismatches between the behavior of the target system and of the simulation are undesired in data-oriented simulation studies. In model-oriented simulation studies, on the contrary, they constitute a valuable experimental result, as they may be taken as good reason to reject the mechanism under assessment.

#### Internal Comparisons and Lesion Studies

The overt motor behavior of the robot is not the only aspect of the robot that may be compared with  $S$ 's behavior. While the robotic simulation is running, its internal components continuously change state, and in some cases it may be interesting to make comparisons

with the state of the corresponding components in the nervous system of the target system, if suitable data are available. In [37.28, p. 199], for example, a simulation of a neural network mechanism supposedly underlying spatial memory in rats was described. The behavior of the artificial neurons, identified during the functioning of the robot, was compared with the behavior of given hippocampal neurons in the rat nervous system, which were thought to correspond to the artificial ones, and the two were found to match to some extent [37.28, p. 199]:

“The responses of simulated neuronal units in the hippocampal areas during its exploratory behavior are comparable to those of neurons in the rodent hippocampus.”

Detection of such *internal* matches may be taken as further evidence corroborating the simulated theoretical model.

It is also to be noted that biorobotic studies may involve experiments in which *artificial lesions* are created in given components of the robot and the resulting behavior assessed. Such experiments may be useful for testing mechanistic explanations of dysfunctional behaviors. As discussed earlier, dysfunctional behavior  $B^*$  under  $C$  is explained mechanistically as the result of a specific impairment of the mechanism  $M$  responsible for the corresponding *normal* behavior  $B$  under  $C$ . The explanatory mechanism, in these cases, consists of an impaired version  $M^*$  of  $M$ . To test whether  $M^*$  can produce  $B^*$ , one may build a robotic implementation of  $M$ , *injure* it so that it may be considered an implementation of  $M^*$ , and assess whether it exhibits behavior  $B^*$ . An example of a (nonrobotic) model-oriented lesion study is described in [37.29]. Experiments involving changes to particular aspects of a biorobot may also be useful for arguing that the same mechanism  $M$ , with minor modifications, can account for a variety of different phenomena. Suppose that mechanism  $M$ , simulated in a robot  $A$ , turns out to be able to exhibit behavior  $R$ ; and suppose that, by changing some parameters of it – without modifying the overall structure of the mechanism –  $A$  displays behavior  $R^*$  different from  $R$ . One may take this result as a basis on which to conclude that  $M$  has some potential for *unification*, according to the analysis of scientific unification provided by philosopher Philip Kitcher [37.30] (it is worth noting, however, that possessing a high unification potential is not, according to some authors, either necessary nor sufficient to provide a *good* explanation – see for example [37.31]; as far as necessity is concerned, in particular, it is not clear whether a mechanism description that can explain behavior  $R_1$  should depend on whether it can explain a different behavior  $R_2$ ). Finally, as discussed later in

Sect. 37.1.2, artificial lesion experiments may also help to test whether the robot accurately implements the theoretical model under scrutiny.

### Biorobotics and the Study of Neural Implementation

Showing that  $M$  can produce  $R$  (Claim 1), as discussed in the previous section, is necessary in order to claim that  $S$ 's behavior  $R$  may be explained by reference to  $M$ . But it is not sufficient, as it does not exclude that  $S$  exhibits  $R$  by virtue of a mechanism that is different from  $M$  but equally efficient (as mentioned earlier, there is a gap between being able to *reproduce*  $R$  in a machine and being able to *explain*  $R$ ). One must therefore also show that  $M$  is implemented in  $S$ , that is to say, that  $S$  has components that are organized and behave as prescribed by  $M$ : this is the second claim introduced in Sect. 37.1.2, *Explanation*. A variety of experimental techniques may be used for this purpose, depending on whether  $M$  is expressed using cognitive (representational) or neuroscientific theoretical vocabulary. Can biorobotic experimentation help in this case too?

It follows from the previous discussion that, if robot  $A$  implementing  $M$  displays behavior  $R$ , one cannot conclude that  $M$  is implemented in  $S$  too. Similarly, if  $A$  does not display  $R$ , one is not authorized to conclude that  $M$  is not implemented in  $S$ . It may be the case that  $M$  is implemented in  $S$  without producing  $R$ . In the biorobotic experiments on lobster chemotaxis discussed earlier, the robot consistently failed to reach the source of the chemical stream. Based on these results, the authors concluded that the simulated mechanism  $M$  did not guarantee efficient chemotaxis. This theoretical conclusion does not imply that  $M$  is not implemented in lobsters, however. Indeed, as the authors suggest at a certain point of their discussion,  $M$  could be regarded as *too simple* rather than completely *wrong*. Indeed, there are good reasons to believe that it might be an effective chemotaxis mechanism in the presence of *homogeneous chemical plumes*. But the plumes that *real* lobsters are able to track are irregular and filamentous. Therefore, one might conclude that lobsters' chemotaxis mechanism is totally different from  $M$ ; or, alternatively, one might well conclude that lobsters use  $M$  supplemented with other components that enable the system to locate the right track when the plume is too scattered. In the latter case,  $M$  would be implemented in  $S$  (although *not only*  $M$ ) while, at the same time,  $M$  as it stands would be irrelevant to explaining lobsters' chemotaxis. To generalize, a target system  $S$  might implement  $M$  even though  $M$  is not the mechanism responsible for the behavior to be explained.

### Constraining and Revising the Space of Possible Mechanisms

For the reasons illustrated in the previous section, one cannot obtain from a biorobotic experiment *per se* any strong reason to decide whether  $S$  implements  $M$  or not. However, weak reasons may be obtained in some cases. Suppose that, based on a number of biorobotic experiments, one concludes that  $M$  is the *only* mechanism producing  $R$  out of several alternative mechanisms, and that no more adequate mechanistic hypothesis can be conceived. This result would increase one's confidence that  $M$  is *the* mechanism producing  $R$  in  $S$ , that is to say, that  $M$  is implemented in  $S$ , although it may still not be viewed as providing strong evidence to support this claim.

To take another example, suppose that  $M$  is known to be *partially* implemented in  $S$ , i. e., that some components of  $M$  can be found in  $S$ , while other components of  $M$  are only fictional (i. e., no information is available as to whether they are in  $S$  or not). In this case, success in robotically replicating  $R$  may be taken as at least a weak reason for predicting that the fictional components will also be found in  $S$ . This case is exemplified by the biorobotic inquiry into mechanisms of spatial memory in rats reported in [37.32]. These authors' mechanistic hypothesis mentioned a number of neural structures that are known to exist in the nervous system of rats, notably including a population of so-called *place* cells, which fire selectively whenever the rat crosses a particular point of the environment [37.13, 33]. To obtain an efficient goal-seeking mechanism, the authors also postulated the existence of so-called *goal* cells, whose function is to signal proximity to goal locations; these cells were argued by the authors to be essential to goal-seeking behavior, even though no information on their existence in the rat nervous system was available at the time of publication of the study. In the biorobotic experiments, the robot implementing place and goal cells was found to be able to generate efficient goal-seeking behavior. This result provided at least a weak reason to believe that goal cells are not only fictional entities, but that they must be really implemented in the rat brain. In this case too, successful biorobotic replication of  $R$  may induce one to expect that some fictional components of  $M$  are to be found in the target system  $S$ . It is worth noting that the construction of a biorobot may also guide the localization of fictional components of  $M$  within the target system. As discussed earlier, the state of  $A$ 's components while the system is running may be monitored and compared with the internal state of  $S$  as established using an appropriate experimental technique. Thus, for example, one might record the activity of artificial goal cells during experiments, and search for neurons displaying compa-



rable activity in the rat brain in order to identify the neural counterparts of the fictional goal cells in the target system.

In sum, although the primary role of robotic model-oriented simulation experiments is to support reasoning on the relationship between  $M$  and the behavioral regu-

larity  $R$  to be explained, they may also assist indirectly in the analysis of the relationship between  $M$  and the target system  $S$ . The roles played by biorobotics in the explanation of human and animal behaviors will be further exemplified in the brief summary of the state of the art in biorobotics provided in the next section.

## 37.2 A Short History of Biorobotics

### 37.2.1 Cybernetic and Artificial Intelligence

As mentioned above, the history of biorobotics stretches back to the first decades of the twentieth century: the idea of building electric or electromechanic simulations of living system behaviors to discover the mechanisms producing them was pursued by many physiologists and psychologists even before the birth of the digital computer. An example, in addition to the *electric dog* described in Sect. 37.1, is the learning device built by engineer *Bent Russell* [37.34], which featured hydraulic mechanisms loosely inspired by the learning theories of Spencer and Thorndike (Russell's machine and many other model-oriented simulation studies carried out in the twentieth century are discussed at length by philosopher and historian of science *Roberto Cordeschi* in [37.5]). Other examples of simulations carried out before the age of the digital computer include various circuits built by psychologist *Clark Hull* and collaborators. An electronic device built by *Hull* and the electrotechnical engineer *Robert G. Krueger* proved able to replicate various forms of Pavlovian conditioning [37.35]. A nonelectronic simulation of learning was built by *Hull* and chemist *H.D. Baernstein* [37.36]. Baernstein's device inspired electrochemist *Thomas Ross* to build a robot able to memorize the structure of mazes on the basis of a sort of *mechanical* memory (reports on the device and its memory abilities appeared in the journal *Psychological Review*, [37.37, 38]). Some years later, cybernetics pioneer *Norbert Wiener* et al. extensively discussed the role of machines in the study of animal behavior in two articles published in the journal *Philosophy of Science* between 1943 and 1945 [37.39, 40], and *William Grey Walter* built the earlier-mentioned *turtles*, which were able to navigate, and react to light stimuli in, domestic environments [37.26].

The use of model-oriented *computer* simulations characterizes the information-processing approach to psychology pursued by Herbert Simon and Allen Newell in the artificial intelligence (AI) era. Their famous *logic theorist* and *general problem solver* programs were presented not only as machines displaying

remarkable problem-solving abilities, but also as simulations of theoretical models of intelligent *human* behavior couched in information-processing theoretical vocabulary [37.41, 42]. They proposed computer simulations as tools “both for constructing theories and for testing them” [37.43, p. 2011]. A similar model-oriented simulation approach was also pursued by *Nathaniel Rochester* et al. to test models of brain functioning [37.44]. More recent model-oriented computer simulations deployed to test neuroscientific models of animal behavior include the LIMAX system, used to theorize on the *logic* of learning in the *limax* slug [37.45], and the computer simulation built by *Hawkins* to test a model of learning in *Aplysia* [37.46].

Information-processing theories of intelligent behavior developed during the artificial intelligence era offered good bases for explaining and simulating various human-level problem-solving abilities exhibited in well-structured environments. However, they turned out to be inadequate for modeling sensory-motor behaviors produced by living systems in partially structured or chaotic environments, and under strict time constraints. AI robots were extremely slow and unable to produce timely reactions to environmental changes, not only due to the speed limitations of the computers available in the 1960s and 1970s, but also due to the structure of the algorithms used to process sensory data and plan motor behaviors (a paradigmatic example of an AI robot is Shakey, described in [37.47]). In the mid-1980s robots were built that exhibited efficient sensory-motor behavioral capacities in partially structured environments based on a parallel and distributed architecture dubbed *behavior-based architecture* [37.48]. Behavior-based algorithms had many aspects in common with the control architectures devised and implemented in the cybernetic era and later described by *Valentino Braitenberg* [37.25]. Their formulation, and the sensory-motor efficiency of the first behavior-based robots, renewed the interest of the robotics community in the construction of robotic systems able to reproduce aspects of animal behavior and in the robotic simulation of neural and psychological mechanisms.

## 37.2.2 Contemporary Invertebrate and Vertebrate Simulation Studies

Many model-oriented biorobotic studies have been conducted, especially from the 1990s onwards. Some of these concern invertebrate sensory-motor capacities. In [37.49, 50] a series of biorobotic studies aimed at explaining how female crickets find possible mates by following their calling song were described. Other notable examples include biorobotic inquiries into the mechanisms underlying the remarkable navigation abilities of the desert ant *Cataglyphis* [37.51], the navigation mechanisms of flying insects [37.52, 53], moths' ability to track chemical streams [37.54], and locust visuomotor coordination [37.55]. As mentioned earlier, biorobots have been used to study the mechanisms of visuomotor coordination in lobsters too [37.24, 56]. Biorobotic simulations have also been deployed to study sensory-motor behaviors in vertebrates. Examples include studies on locomotion in lampreys [37.57, 58] and salamanders [37.59], and on the mechanisms underlying general aspects of locomotion exhibited by a variety of animal species [37.60–63]. The mechanisms underlying spatial memory and navigation in rats have been investigated using biorobots in [37.20, 28,

32, 64, 65]. Sense of touch in animals has been explored by means of whisker-controlled robots in [37.66, 67]. Visuomotor coordination in the barn owl has been studied by [37.68]. There have also been biorobotic inquiries into primate behavior, notably including studies on postural control [37.69] and on cerebellar mechanisms of motor control in humans [37.70]. See [37.71] for a detailed discussion on the role of biorobotics in the study of primate behavior.

Model-oriented simulative approaches have additionally been used to advance understanding of the *development* of behavior (see for example the biorobotic study on the development of visuomotor coordination in cats described in [37.72]). Detailed reviews of biorobotic investigations on the *evolution* of behavior are offered by [37.73] and [37.74]. It is worth mentioning here the possibility of building *hybrid* simulations of mechanism descriptions, i.e., simulation systems in which some components of the target mechanism consist of biological tissues appropriately connected with other components that are artificial. Systems of this kind have been used to study the mechanisms of lamprey sensory-motor behavior [37.75–77]; see [37.78] for a methodological analysis.

## 37.3 Methodological Issues

### 37.3.1 The Epistemic Requirements of Good Biorobots

#### Biological Mimicry

In the previous section, we reviewed studies illustrating the role played by biorobotics in the explanation of human and animal behavior. In all of these studies, the behavior of a robotic system  $A$  is taken as a basis for assessing whether a system governed by  $M$  only would produce behavior  $B$  under  $C$  (as discussed in Sect. 37.1.2, *Explanation and Biorobotics and the Study of Ideal Mechanisms*). This assessment is essential, though not sufficient, in order to conclude that  $M$  is the mechanism enabling  $S$  to exhibit  $B$  under  $C$ . In the current section we ask under what conditions – if any – one is really *justified* in bringing robotic behaviors to bear on the mechanism  $M$  under scrutiny. More precisely, let us assume that  $A$ , in a sufficient set of experimental trials, has actually exhibited behavior  $R$ , thus reproducing the behavior of the target system (an assumption that is discussed in Sect. 37.2). What auxiliary assumptions are needed to justifiably take this result as a basis for concluding that  $M$  can generate  $B$  under  $C$ ? This key epistemological question has been addressed

in other methodological analyses of biorobotics [37.8, 9]. Awareness of these auxiliary assumptions may help to carry out methodologically sensible biorobotic experiments and to evaluate the soundness of studies reported in the literature.

Some of these assumptions follow quite directly from the discussion carried out so far. In order to take  $A$ 's behavior as bearing on whether a system governed only by  $M$  would display  $B$  in a setting in which only conditions  $C$  hold, one must ensure that (a)  $A$  is governed only by  $M$  and (b) only conditions  $C$  hold in the experimental setting. Let us focus on assumption (a) and postpone discussion of assumption (b) until Sect. 37.3.3. Assumption (a) places a constraint on the structure of robot  $A$ , a constraint that may seem quite hard to satisfy – not least, because robots are complex devices: to work properly any given robot is likely to need many peripheral components not mentioned in the mechanism description under scrutiny, as we will discuss in the following subsection. It is worth noting, however, that other authors seem to place even stronger constraints on the structure of *good* biorobots.

*Krichmar* et al. [37.28], for example, proposed a list of principles that should guide the design and building

of *good* biorobots, which notably includes the following one: “The device must be controlled by a simulated nervous system having a design that reflects the brain’s architecture and dynamics” [37.28, p. 198]. This principle, similarly to assumption (a), places a constraint on the mechanism governing the device. However, this constraint is defined in terms of the mechanisms at work in the target system (the brain). Reformulated using our terminology, it prescribes that *A* be governed by a mechanism *M* which is known to be (at least partially) implemented in the target system *S*. A similar constraint is formulated by [37.71, p. R910, emphasis added], who claim that “robots incorporating the biomechanics of the animal system under study become physical models to test hypotheses”. Let us generalize by distinguishing the following two constraints:

1. The robot *A* must be governed only by *M*.
2. *M* must be implemented in the target system *S*.

Let us label the combination of Constraints 1 and 2 *biological mimicry*. It goes without saying that biological mimicry is a much stronger constraint than the mere prescription that *A* must be governed by *M*. Note that the terms *validation* and *verification* are often used in the literature in reference to the evaluation of these two requirements [37.2]. Validation refers to the evaluation of the *goodness* of the theoretical model (Requirement 2), while verification is the process of checking whether the theoretical model has been accurately simulated (Requirement 1).

Is biological mimicry, as defined above, required to justify the use of *A* to test whether *M* can produce *B* under *C* or not? Arguably not. More precisely, Constraint 1 above is surely required: if *A* is not governed only by *M*, it is not clear why one should interpret *A*’s behaviors as implied by *M*. On the contrary, Constraint 2 is not required for *A* to be a *good* biorobot. Remember that *A*’s role in biorobotics is to provide evidence of the behavioral implications of *M* under *C* (Sect. 37.1.2, *Biorobotics and the Study of Ideal Mechanisms*). There is no reason to claim that *A*’s behaviors may be interpreted as implied by *M* only if *M* is implemented in *S*: a robot could well be used to explore the behavioral implications of purely notional mechanisms that are impossible to implement biologically. This brings to light a substantial methodological difference between data-oriented and model-oriented simulation studies. In order to use *A* to generate data about the target system *S*, *A* must be a *good* model of *S* – otherwise there are no good reasons to interpret *A*’s behavior as the behavior that *S* would exhibit under the same conditions. This constraint is concisely stated by Wendy Parker in the following terms [37.79]; see also [37.80, 81]:

“scientists typically select a simulating system on the basis of its being hoped or believed to be similar to the target system in ways deemed relevant, given the goals of the simulation study.”

On the contrary, similarity between the mechanism governing the machine and the mechanism at work in the target system is not required to authorize the experimental use of the machine in a model-oriented study.

Note that so far we have reasoned about the conditions under which one is justified in interpreting *A*’s behaviors as implications of *M*. We have argued that Constraint 2 is not required in order to make this use of *A*. This is not to say that simulation of a mechanism that is already known to be (at least partially) realized in *S* is not to be praised for *other* reasons not related to the justification of using *A*. One may well agree with [37.65, p. 1527, emphasis added], who assert that:

“a close fidelity to the known properties of the nervous system is *likely to be* a main ingredient of success in modeling studies aimed at reaching an understanding of higher brain function.”

They complain that, so far [37.65, p. 1501]:

“very little has been done to incorporate detailed models of cellular and synaptic properties into large scale simulations of interconnected networks in multilevel systems.”

Darwin robots, on the contrary, “are based on physiological and anatomical data” [37.65, p. 1498]. Indeed, (robotic) simulations of mechanisms that “reflect the brain’s architecture and dynamics” may offer interesting experimental opportunities to cognitive and neuroscience research. It could be that detailed information is available about the neural structures and morphological features of a given species, but explanations of how they are organized to produce behavior are lacking. In this particular case, one might propose a particular organization of those components (i.e., formulate a mechanism description *M* that includes them), simulate it using a robot, and verify whether the robotic system can produce the behavior under investigation. This case emphasizes the role that biorobotics may play in systematizing anatomical and physiological observations into potentially explanatory mechanisms.

It is for these reasons that many authors actually *aim* to build robotic systems, which at least partially meet Constraint 2. Pyk et al. [37.54, p. 197], for example:

“report on a project that aims at constructing [...] a system based on [their] understanding of the pheromone communication system of the moth.”

Lambrinos et al. [37.51, p. 40], in commenting on their biorobotic inquiry into the mechanisms underlying the navigation strategies of the desert ant *Cataglyphis*, pointed out that:

“the goal of this approach is to develop an understanding of natural systems by building a robot that mimics some aspects of their sensory and nervous system and their behavior.”

In a biorobotic study on forearm posture maintenance in humans, Chou and Hannaford [37.69] claimed that their goal was:

“to apply knowledge of human neuro-musculoskeletal motion control to a biomechanically designed, neural controlled, ‘anthroform’ robotic arm system”

Blanchard et al. [37.55] implemented a mechanistic model of obstacle avoidance in locusts that was “based closely on the anatomy and physiology”.

It follows from the present discussion that artificial device *A* need *not* replicate the behavior of the target living system *S* to be used to test a model *M* of *S*’s behavior: what is required is that *A* be governed by *M*, independently of whether it replicates *S*’s behavior or not. Behavioral match between *S* and *A* is one of the possible *outcomes* of the experiments, but it is not among the *epistemic requirements* needed to use *A* to test *M*. However, an electronic device accurately reproducing the behavior of the corresponding living system may be used as a component of a larger simulation system in a model-oriented study. The robotic device described in [37.82] is a case in point. As reported by the authors, it was able to replicate accurately the output of mammalian muscle spindles (sensors able to detect changes in the length of muscles) based on a detailed theoretical model of them. This device was not used to discover theoretical models of muscle spindle behavior as in model-oriented simulations (good models were already available), but was nevertheless intended to play a useful role in scientific research. To make this clearer, let us suppose that *M* is a neural model of motor control that has muscle spindles among its components. Building a robotic model-oriented simulation of *M* clearly requires the inclusion of a robotic component that accurately reproduces the behavior of muscle spindles. The device described in [37.82] may be useful for this purpose (rather than for the narrower purpose of testing the plausibility of a model of the muscle spindle itself, which is already available). Similar considerations apply to other attempts to build biologically accurate sensors, see for example [37.83, 84].

To sum up, what is needed to license the use of robot *A* as an experimental platform for investigating the im-

plications of mechanism description *M* is that *M* be the only mechanism governing *A*. *Good* biorobots need not feature mechanisms closely based on the physiological or psychological structure of the target system (although biological mimicry may yield valuable findings for cognitive and neuroscientific research). To be sure, biorobotic simulations can be particularly useful when no information is available on the biological implementation of *M* (thus, when Constraint 2 above is false): they may assist in evaluating whether *M* can generate in principle the behavior under investigation, thus whether a biological or psychological implementation of *M* is worth searching for. Let us now focus on Constraint 1: under what conditions, if any, may robot *A* be said to be governed only by the mechanism under scrutiny?

### Simulation Accuracy

The very aim of building a robot governed by a neuroscientific or cognitive mechanism seems puzzling at first sight: how can an artificial, *inorganic* system be governed by the same mechanism at work in a *biological* system? Cognitive and neuroscientific mechanisms are made up of neural and psychological components, and for this reason it may seem more appropriate to say that robots can at most realize *artificial translations* of them, but not exactly *them*. We earlier dispelled this doubt in Sect. 37.1.2, *Biorobotics and the Study of Ideal Mechanisms*. There is no *a priori* reason to deny that an artificial system may be governed by *the same* mechanism governing the behavior of a biological or psychological system, despite the fundamental difference in the matter of which the two systems are made. Indeed, as discussed earlier, a basic tenet of the mechanistic approach to scientific explanation pursued in cognitive science and neuroscience is that whether a system governed only by *M* can exhibit *R* or not is a question that does not depend on the material that the system is made of, but rather on the way that material is organized (i. e., on whether that system is organized as prescribed in *M*). For example, that one may explain spatial memory behaviors in rats by reference to hippocampal *place cells* depends on the fact that these cells fire *regularly* in correspondence with certain points in the environment, rather than on the fact that they are cells made of organic molecules. Appropriately replacing living place cells in the hippocampus of a rat with electronic circuits displaying the same regularities would not alter the animal’s overall spatial memory mechanism (replacements of this kind are made in bionic model-based simulations of sensory-motor behaviors, of which an example is the aforementioned study described in [37.77]).

In principle, then, there is no reason to deny that a robotic system *A* may be governed by the same

mechanism  $M$  supposedly governing biological or psychological system  $S$ , even though  $A$  and  $S$  are made of completely different materials. What matters, for  $A$  to be governed by  $M$ , is whether  $A$  has components behaving as specified in  $M$  (e.g., components correctly reproducing the regularity associated with place cells) and organized accordingly. What seems to be more difficult to establish is whether  $A$  is governed *only* by  $M$ , an essential requirement in order to interpret  $A$ 's behavior as the behavior that a system governed only by  $M$  would display under conditions  $C$  (as discussed in Sect. 37.1.2, *Explanation*). Indeed, neuroscientific and cognitive mechanism descriptions are typically much *simpler* than the mechanisms actually implemented in their robotic simulations. In these cases, the behavior that a system governed only by  $M$  would have under conditions  $C$  must be *inferred* from  $A$ 's behaviors under  $C$ . What auxiliary assumptions are needed to draw such an inference?

Let us try to address this question by making a provisional distinction between the mechanism *implemented* in  $A$  (MI) and the mechanism *governing*  $A$  (MG) in particular circumstances. This distinction may be informally stated as follows. The mechanism description MI implemented in  $A$  is a full description of the components making up the system, at some (e.g., electronic or software) level of description. It describes the intended behavior of the components and their organization in normal conditions, and it may be thought of as the best blueprint used by the builders of the robot during the course of implementation. The mechanism MG governing  $A$  under a particular set of circumstances  $C$  is the mechanism actually responsible for  $A$ 's behavior under  $C$ . The two mechanism descriptions need not always be identical.

First, if some components implemented in  $A$  are silent, or exert a negligible influence on  $A$ 's behavior under conditions  $C$ , they cannot be said to be responsible for  $A$ 's behavior, and should be therefore excluded from the mechanism governing  $A$  under  $C$ . In this case, the mechanism MG governing  $A$  under  $C$  is a restriction of the mechanism MI implemented in  $A$  (MG may be obtained from MI by ignoring all the ineffective components). Suppose, for example, that  $A$  implements a mechanism establishing a relationship between left light sensors and right motors and vice versa (as in [37.24]), plus (a) a wireless communication module for exchanging data with an external computer and (b) a circuit that produces a *wandering* behavior when environmental light is too low. All these components figure in the description MI of the mechanism implemented in the robot. However, the wireless communication module will probably never have an impact

on  $A$ 's motor behavior. On the contrary, component (b) may influence  $A$ 's behavior in some circumstances, for example, in the dark. When there is sufficient environmental light, however, this component will be silent and will not be effectively responsible for  $A$ 's behavior: turn it off, or extract it from the system, and  $A$ 's behavior will not change under those conditions. Therefore, *when there is sufficient environmental light*, MG does not include (b).

Second, in some cases the behavior of certain components implemented in the machine may violate the behavior specified in the blueprint. Suppose that MI prescribes that a transistor is placed between each light sensor and the motor on the opposite side, to amplify smaller sensory signals. Suppose, however, that under particular conditions (for example, excessive heat) the transistor displays an utterly anomalous behavior (for example, its pins get short-circuited). In such cases, the mechanism MG actually governing the robot is different from the mechanism MI specifying the *normal* behavior of the components.

It follows from these remarks that the same robot may be reasonably said to be governed by *different* mechanisms in different conditions, even though there is a sense in which it *implements only one* mechanism. Let us now link this claim with the epistemic requirements of *good* biorobots. We have argued that to inform about the behavior a system governed only by  $M$  would display under conditions  $C$ ,  $A$  must be governed *only* by  $M$ . And we have illustrated that robots possessing a rich mechanistic structure may be said to be governed by *simpler* mechanisms in some circumstances. What is essential for a *good* biorobot, in other words, is that  $M$  exactly coincides with the mechanism *responsible* for  $A$ 's behavior in circumstances  $C$  independently of the other components implemented in it. How may we establish whether this essential requirement is satisfied by  $A$  – in other words, how does one go about identifying *the* mechanism actually governing  $A$  under conditions  $C$ ? By nothing other than a process of *explanation*, totally analogous to the process leading to the identification of the mechanism governing the behavior of any system. The mechanism description MG governing  $A$  under  $C$ , according to the proposed analysis of the term *governing*, is a *mechanistic theory* of  $A$ , resulting from an explanation of  $A$ 's behavior under  $C$ . Any explanation will mention only the relevant factors making the difference in the behavior to be explained [37.85, 86]. Thus the *wandering* component is not likely to figure in an explanation of the behavior *in the light* of the phototaxis robot described above; nor the wireless communication module, if the behavior to be explained is restricted to the movements of the robot.

### Explaining Robotic Behaviors

These remarks illustrate a quite startling aspect of biorobotic methodology: the use of robotic simulation to test a mechanistic explanation of a biological behavior rests on *explanations of the behavior of the robot*, aimed at determining whether the robot is actually governed by the mechanism under scrutiny. Such explanation processes are essential to deciding whether  $M$  is to be accepted or rejected as a basis for explaining the target behavior. And the explanation of robotic behaviors is not always straightforward, even for those who have built the system: the behavior of man-made systems may be as hard to understand as the behavior of any physical system. The technical blueprint  $MI$  helps, but in some cases it may be not enough. Often even the builder, who knows the internal structure of the system better than anyone else, will have to revisit it, formulate hypotheses as to why the robot produced peculiar behaviors in certain circumstances, and carry out experiments to test these hypotheses. It is by a process of explanation that one identifies a fault in a transistor or the unexpected activation of a component that was presumed to be silent in particular experimental settings. Detailed reports of robotic explanation processes can be found in the biorobotic literature; see for example [37.24] and [37.49].

It is worth noting, however, that further auxiliary assumptions may be needed to interpret  $A$ 's behaviors as the behavior that a system governed only by  $M$  would produce under conditions  $C$ . Indeed,  $A$  may fail to be governed by  $M$  for reasons not discussed so far. Current technology might not permit the construction of components behaving exactly as prescribed by  $M$ , obliging the researcher to design components that at best exhibit an adapted version of the required behavior. Alternatively, the formulation of  $M$  might be too vague to enable one to judge whether  $A$  is governed by  $M$  or not. Consider that many cognitive and neuroscientific models have unfixed parameters, describe the behavior of components only qualitatively, and display gaps, such as missing components (these are referred to as *mechanism sketches* in [37.15]). All these underspecified aspects must be fixed in the course of implementation. Given that a vague mechanism description  $M$  may be fully specified in a variety of ways, thus obtaining various mechanism descriptions  $M_1, \dots, M_n$ , which potentially differ greatly from each other in terms of behavioral output, one may legitimately ask whether the particular  $M_i$  governing  $A$  faithfully reflects the vague theory  $M$  under scrutiny. And in some cases it may be far from clear whether  $A$ 's behavior should be brought to bear on a vague  $M$ . Suppose, for example, that  $A$  fails to exhibit the target biological behavior  $R$ . This result may be taken as a good reason to reject  $M_i$ , that is, to

conclude that the particular parameters chosen to derive  $M_i$  from the vague  $M$  cannot produce  $R$ . But it may not be sufficient to also reject  $M$ , insofar as *other* different parameters might well work.

In sum, to sensibly interpret  $A$ 's behavior as the behavior that a system governed only by  $M$  would display under conditions  $C$ , one must first explain  $A$ 's behavior and identify the mechanism that actually governs  $A$  under  $C$  (which does not always coincide with the best technical blueprint of the robot available). A *mechanistic theory* on the robot is crucially needed to infer theoretical conclusions about  $M$  based on  $A$ 's behavior. But in many cases other auxiliary premises are implicitly introduced, whose nature has yet to be precisely identified. For in some cases the best one can do is to build an *inaccurate* simulation of  $M$ . This may be due to limitations in current technology or to the fact that  $M$  is formulated in excessively vague terms. Or it may be also due to the fact that, in the experiments, some perturbing condition has actually interfered with the behavior of the robot, some component has unexpectedly broken down, or a module that was expected to be silent has interfered with  $A$ 's behavior. In some cases, such artifacts are simply unavoidable. Should one therefore discard the entire experimental setting? What auxiliary assumptions are needed to justify theoretical conclusions on  $M$  based on an inaccurate robotic simulation of it? These are interesting questions for epistemological research, which may lead to a deeper understanding of the methodological structure and epistemological requirements of *good* biorobotic studies.

### 37.3.2 On the Meaning of Behavior

Biorobotic experiments typically involve comparisons between the behavior of  $A$  and the behavior of the target system  $S$ . Matches or mismatches between the two (under the auxiliary assumptions discussed in the previous section) are taken as a basis for claiming that the mechanism description under scrutiny has been corroborated or disproved. To make sense of what is really done in biorobotic experiments, however, it is worth reflecting on what the *living system behavior* that is compared with robotic behaviors amounts to.

First, typically only *some aspects* of the behavior of the target system are compared with *some aspects* of the behavior of the robot. For example, in the aforementioned biorobotic studies on cricket phonotaxis [37.49] and on lobster chemotaxis [37.24] various aspects of the behavior of the robot were measured with great care in the experiments. However, theoretical conclusions on the target model  $M$  were drawn in both cases only by reasoning on the ability of the robot to reach the source of the stimulus, irrespectively for example of any match

or mismatch between the *trajectories* followed by *A* and *S* (as a matter of fact, in some cases it is impossible to make fine-grained comparisons between *A*'s and *S*'s behaviors due to the unavailability of data on *S*). Thus acceptance or rejection of *M* was based on analysis of only *some* aspects of *A*'s behavior.

This does not count as a methodological limitation of these studies: there is no reason to claim that, in a good biorobotic experiment, one must perform fine-grained behavioral comparisons of the behavior of the two systems. The aspects of *S*'s behavior that are actually taken into account in the comparison define the explanandum of the biorobotic study, thus shaping the class of the theoretical conclusions that can be legitimately drawn from the experiments. If, for example, one takes into account only the robot's ability to find the stimulus source, irrespective of the trajectory leading to it, one will at most draw theoretical conclusions on the ability of the underlying mechanism *M* to find the source of the stimulus. Any theoretical conclusion on *M* will be limited to whether it can explain *that particular aspect* of the behavior of *S*. A finer-grained comparison between the trajectories taken by the two systems will be needed only if the trajectories taken by *S* are part of the explanandum. The choice of the criteria for comparison crucially depends on what one wants to explain (see [37.87] for a detailed reflection on the metrics enabling one to compare animal and robotic behaviors based on a case study on the mechanisms of rat navigation).

These remarks invite us to reflect on the meaning of the term *behavior* as it is used in biorobotics. The behavior of the target system in a general sense is one thing, while the behavior of the target system contained in the explanandum of the study is another. Biorobotic explananda always select some aspects of target system behavior, and thus indirectly constrain the choice of the criteria to be used for assessing whether the robotic simulation replicates the target system's behavior or not. Note that, however, biorobotic explananda do not only consist of a *filtered* description of the behavior of the target system – that is to say, a description that only takes some aspects of that behavior into account. Recall that the behavior to be explained consists of the behavior *B* regularly produced by the target system *under conditions C* (and if *C* occur rarely, the behavioral regularity *R* to be explained will be rarely instantiated). Real moths do not always go towards a light source, and female crickets find the source of male crickets' calling songs only under particular conditions. Consider also that, as pointed out in Sect. 37.1.2, *Explanation*, biorobotic *explananda* are highly idealized: they state behaviors that are produced in the ideal case in which *only conditions C* hold. This is not to say that this ideal

case can never occur. The behavior to be explained may be obtained in highly controlled experimental settings, in which one tries to exclude or neutralize any boundary condition that is not included under *C*. The list of the conditions defined under *C* must be taken into account in the setting up of the biorobotic experiments, whose purpose is to assess whether systems governed only by *M* produce behavior *B* when only conditions *C* hold. Therefore, ideally, the robot *A* should be observed under experimental conditions that are as close as possible to the ideal conditions in which only *C* hold. The explanandum thus crucially defines the conditions under which the behavior of the robot is to be observed, in addition to the criteria to be used for analyzing it. As it will be stressed in the next section, the real conditions under which the behavior of the robot is observed are often far from ideal.

It is worth noting that the formulation of the behavioral regularity under investigation involves various acts of observation and measurement. However, it also involves careful selection of the environmental conditions under which these observations and measurements are carried out. This selection is definitely theory-laden: possibly perturbing conditions are excluded or neutralized in the experimental setup in light of theoretical considerations – possibly supported by previous experiments on the target system – that classify those conditions as *perturbing*. The choice of environmental conditions also depends on the researcher's particular interests: some conditions may be excluded because they have behavioral effects that the researcher wants to exclude from the explanandum to be explored. Contrary to the popular assumption that science starts from the observation of something to be explained, the formulation of an explanandum is typically a constructive process involving progressive refinements of the conditions under which the experimental acts of observation are carried out, a process that is strongly driven by theoretical and pragmatic considerations.

It should also be noted that biorobotic studies do not always start from the analysis of behavioral regularities displayed by particular living species. They often aim at theorizing on the mechanisms underlying very *general* behaviors, which cannot easily be identified with the behavior of particular living systems under particular conditions. For example, Voegtlin and Verschure [37.88] described a number of simulation experiments aimed at exploring mechanisms that produce a variety of learning-related behavioral phenomena. The behavioral task that these robotic and computer simulations were expected to perform was a foraging task, in which the agent was required to avoid collisions with obstacles while locating targets dispersed in the environment. Many living systems are

able to perform this task, human beings included, and the simulation study reported did not focus on any one of them in particular. In such cases, the target behavior does not coincide with the behavior of a single living species: it unifies the behavior of many living systems, which may differ from each other in other respects.

Other simulation studies follow the so-called *animat* approach defined by Webb [37.89] as:

“the invention and study of an artificial creature that does not correspond to any specific real animal, but is (nevertheless) intended to provide some insight into real issues in biological or cognitive science.”

It is worth stressing that there is a sense in which every biorobotic explanandum is *invented*. As discussed above, biorobotic explanations address behaviors exhibited by living systems in highly artificial and controlled conditions, which may even be totally different from the real conditions in which the organism lives. The behavior to be explained is carefully *carved out* of the behavior exhibited by the system in ordinary conditions based on theoretical and pragmatic considerations. The *animat* approach, as defined above, brings this constructive process of invention to an extreme. It authorizes investigations on the mechanisms underlying the behavior of imaginary systems, or the behavior that existing systems might display under conditions that can be never attained. Whether this approach can really contribute to cognitive science or neuroscience is a question that depends on the relationship between the *animat* explananda and the questions addressed in those research disciplines. As discussed above, studies on animal behaviors often investigate highly idealized or general explananda, whose relevance to cognitive or neuroscientific research is typically justified by showing how the behavior of particular biological species may be considered as an instance of such general or idealized behaviors. As argued by Webb [37.89], such a justification must be provided by *animat* researchers too, if they wish to effectively contribute to the study of the behavior of existing living systems: they must show that their explananda may be considered idealized or general versions of specific neuroscientific or cognitive explananda (on the basis of given criteria or rules for idealizing and de-idealizing living system behaviors).

### 37.3.3 Robots and Their Environment: Robotic versus Computer Simulations

It is a basic tenet of contemporary cognitive and neuroscience that a full understanding of the mechanisms governing human and animal behavior cannot be attained without carefully studying the structure of the

environment in which the target behavior emerges. This tenet is closely related to the idea, discussed above (Sect. 37.1.2, *Biorobotics and the Study of Ideal Mechanisms*), that *simple* mechanisms can display *complex* behaviors in *complex* environments. Thus, when observing the behavior of a living system, one should first carefully explore the environmental conditions in which it is produced by *S*. This remark, once again, ultimately concerns the selection of the proper explanandum to be addressed. One must carefully select the conditions *C* that define the behavioral regularity to be explained, by including those that really make a difference to the behavior of the target system and excluding those that are irrelevant. The formulation of a sufficiently inclusive explanandum may contribute to simplifying the explanation. Note that scientific *discoveries*, in biorobotics, may consist both of the formulation of *good* explanatory mechanism descriptions and in the refinement of previous explananda – in the latter case, one discovers that environmental factors that had previously been overlooked affect the behavior of the target system and for this reason are to be included in the explanandum.

Another basic tenet of contemporary neuroscientific and cognitive research is that the morphology of the target system must be carefully analyzed and taken into account in the explanation of behavior, given that *simple* sensory-motor mechanisms may exploit particular features of the system’s body to achieve *complex* behaviors [37.90]; see biorobotic examples in [37.49] and [37.91]. Note that this tenet is perfectly consistent with the abstract character of scientific explanations discussed in Sect. 37.1.2, *Biorobotics and the Study of Ideal Mechanisms*. There we argued that whether mechanism *M* can explain behavior *R* or not is a question that does not depend on the nature of the material the system is made of, but on the way it is organized (i. e., on whether it is organized as prescribed by *M*). However, it is one thing to deny that the explanatory adequacy of *M* depends on the material *S* is made of, and another to deny that it depends on the *body* of *S*. Indeed, morphology describes the physical *organization* of a system, and therefore a morphological description is *abstract* – with respect to the matter the system is made of – just as a neural or cognitive mechanism description is. Suppose, for example, that *S* is found to produce efficient locomotion by exploiting particular passive elastic properties of its legs via a relatively simple control mechanism. Still, the nature of the material of which the legs are made of does not affect the adequacy of this explanation (although whether or not the legs possess the required elastic properties may make a difference). Similarly to the example provided in Sect. 37.1.2, *Biorobotics and the Study of Ideal Mechanisms*, if *S*’s legs are replaced with legs



made of a completely different material, but possessing the same elastic properties, the explanation will not change.

We have argued that, to test whether  $M$  can explain  $R$ , one must assess whether a system whose components behave according to the regularities prescribed by  $M$ , and are organized as specified in  $M$ , exhibits  $R$  or not. It is often claimed that if  $M$  is based on the morphology of a system and  $C$  defines a nontrivial set of real-world conditions, a *robotic* simulation should be built: robots have a body and their behavior may be observed under real-world conditions. However, as is occasionally acknowledged in the biorobotic literature [37.88], there is no reason to deny, at least in principle, that an adequate test may also be effected by creating a purely computer-based simulation of a system *with the selected morphological properties* and observing its behavior under a *simulation of the real-world conditions* included under  $C$ .

One reason often put forward for preferring a robotic simulation over a purely computer one in biorobotic studies is that the physical properties of  $S$ 's body and the various conditions  $C$  defined in the explanandum may be technically difficult to simulate in a sufficiently realistic way (this is the case of the chemical plumes tracked by lobsters in water [37.24]). Furthermore, a detailed description of conditions  $C$  is often not available for simulation. Building an artificial body and immersing it in a real-world environment allows one to sidestep these difficulties. Note, however, that the use of a robotic simulation introduces distinctive methodological issues that are not raised by computer simulation experiments. As discussed earlier, in order to justifiably draw theoretical conclusions about  $M$  on the basis of robot behaviors, one must ensure that the robot really *has* the morphological properties mentioned in  $M$ , and that other properties of the

system do not exert a significant impact on its behavior during the experiments. In addition, the experimental setting must be similar enough to the ideal case in which *only* conditions  $C$  hold – otherwise, one cannot bring robotic behaviors to bear on the hypothesis that systems governed by  $M$  generate behavior  $B$  *under*  $C$ . In particular, it must be verified that no perturbing factor significantly affected robotic behavior during the experiments (consider that the class of factor that may perturb the *normal* functioning of the robot is likely to be very different from that perturbing the behavior of the target living system). This in turn requires, as discussed above, explaining the behavior of the robot.

Computer simulations are on a par with biorobots in this respect, as the need to explain what has happened during the experiments – whether virtual or otherwise – emerges in both cases. However, computer simulations enable one to finely control all the conditions holding in the simulated environment and, more importantly, to create a virtual environment in which one is sure that *only* conditions  $C$  hold. This considerably simplifies the process of explanation. Inquiring into the behavior of a robot, in contrast, may give rise to difficulties of the same order of magnitude of those involved in explaining the behavior of any concrete system – including those related to the discovery and detection of possible perturbing factors. In sum, the emphasis on the morphological and environmental factors responsible for the behavior under investigation is not a good reason, in principle, to prefer robotic simulations over computer simulations. The pros and cons of the former or the latter solution are to be evaluated on a case-by-case basis, taking into account the technical complexities involved in creating accurate simulations of the environment and of the body of the system, and the methodological complexities involved in explaining the behavior of the simulation (see also [37.92]).

## 37.4 Conclusions

In this chapter we have outlined the structure of biorobotic methodology for the study of the mechanisms underlying animal behavior, presented a brief summary of the state of the art in biorobotic research, and discussed some of the epistemological and methodological issues arising, which notably concern the relationship between the mechanistic theory to be tested and the robot, the choice of criteria for comparing the behavior of the robot with the behavior of the target system, and the pros and cons of robotic versus computer simulations. It is important to note that these issues should not be taken as reasons for being skeptical about

the potential of the methodology, which (as pointed out in the previous sections) has made many valuable contributions to the study of animal behavior. Indeed, many studies reported in the literature display a rigorous approach to justifying their theoretical conclusions by carefully reasoning about the relationship between the model to be tested, the experimental settings, and the structure of the robot used. Discussing these issues from an epistemological and methodological point of view may contribute to identifying criteria for setting-up and carrying out *good* biorobotic experiments and, more generally, to placing biorobotics as a strategy for

the study of animal behavior on firmer methodological grounds.

The issues addressed here point to the existence of a profound gap between being able to reproduce a given behavior in a robot and being able to explain it. This gap is not only due to the fact that, as discussed in this chapter, successful simulations do not offer strong evidential grounds for assessing whether the implemented mechanism is implemented in the living system or not (and we have argued that this question must be answered in order to properly explain why the behavior under investigation is produced *by the living system*). It is also due to the fact that reproducing the target behavior is not the same as understanding *how* – i. e., by virtue of what mechanism – that behavior has been produced *by the robotic system*. And, as discussed in Sect. 37.3.1, *Simulation Accuracy*, developing a *theory* on the biorobot is an essential prerequisite for accepting or rejecting a theoretical model on the basis of biorobotic experiments: merely reproducing the behavior is not enough. Consider also, in this regard, that computer and robotic devices enable researchers to progress from “toy models” of behavior [37.93] to the fine-grained simulation of complex theoretical models of the nervous system (the Blue Brain Project of the École Polytechnique Fédérale de Lausanne (EPFL) in Lausanne, aimed at building a large-scale, neuron-level simulation of the brain, is a case in point [37.94]). Still, it is an open question whether large-scale, extremely complex theoretical models – too large and detailed to be grasped by the human mind without external memory and computational

aids – really provide a basis for *understanding* the behavior they are able to produce in simulation. Progress in the computer and robotic replication of living system behaviors does not always go hand in hand with progress in the explanation of them.

We have identified some auxiliary assumptions that are needed to justifiably infer theoretical conclusions about the target mechanism description from the analysis of robotic behaviors. Some of these assumptions concern the choice of a *good* biorobot, of a *good* environmental setting, and of a *good* set of criteria for comparing animal and robotic behaviors. However, alternative auxiliary assumptions may be needed in many cases. For instance, for theoretical and practical reasons it may not be possible to build a perfectly accurate robot, to reproduce the right environmental settings, and to apply the proper criteria for comparisons. In these cases, the auxiliary assumptions identified here turn out to be false. But their falsity needs not neatly imply the impossibility of justifying any theoretical conclusion flowing from the experiments. Other assumptions may be involved in the rational acceptance or refusal of the target mechanism description based on experimental protocols that do not meet the constraints discussed here. Identifying the nature of these rational assumptions, possibly based on a close analysis of case studies, may significantly contribute to extending the regulative framework presented here and, ultimately, to further reinforcing the status of biorobotics as a methodologically sensible strategy for the modeling of animal mechanisms.

## References

- 37.1 R.O. Dror, R.M. Dirks, J.P. Grossman, H. Xu, D.E. Shaw: Biomolecular simulation: A computational microscope for molecular biology, *Annu. Rev. Biophys.* **41**, 429–452 (2012)
- 37.2 E. Winsberg: Computer simulation and the philosophy of science, *Philos. Compass* **4**(5), 835–845 (2009)
- 37.3 F. Guala: Models, simulations, and experiments. In: *Model-Based Reasoning: Science, Technology, Values*, ed. by L. Magnani, N.J. Nersessian (Springer, New York 2002) pp. 59–74
- 37.4 J. Loeb: *Comparative Physiology of the Brain and Comparative Psychology* (Putnam, New York 1900)
- 37.5 R. Cordeschi: *The Discovery of the Artificial: Behavior, Mind and Machines Before and Beyond Cybernetics* (Kluwer Academic, Dordrecht 2002)
- 37.6 E. Winsberg: Computer simulations in science. The Stanford Encyclopedia of Philosophy (Fall 2014 Edition), ed. by E.N. Zalta, <http://plato.stanford.edu/archives/fall2014/entries/simulations-science/>
- 37.7 B. Webb: Can robots make good models of biological behaviour?, *Behav. Brain Sci.* **24**, 1033–1050 (2001)
- 37.8 B. Webb: Validating biorobotic models, *J. Neural Eng.* **3**(3), R25–35 (2006)
- 37.9 E. Datteri, G. Tamburrini: Biorobotic experiments for the discovery of biological mechanisms, *Philos. Sci.* **74**(3), 409–430 (2007)
- 37.10 G. Tamburrini, E. Datteri: Machine experiments and theoretical modelling: From cybernetic methodology to neuro-robotics, *Minds Mach.* **15**(3–4), 335–358 (2005)
- 37.11 D.M. Wolpert, R.C. Miall, M. Kawato: Internal models in the cerebellum, *Trends Cogn. Sci.* **2**(9), 338–347 (1998)
- 37.12 R. Cummins: Functional analysis, *J. Philos.* **72**(20), 741–765 (1975)
- 37.13 E.I. Moser, E. Kropff, M.-B. Moser: Place cells, grid cells, and the brain's spatial representation system, *Annu. Rev. Neurosci.* **31**, 69–89 (2008)

- 37.14 J.R. Stroop: Studies of interference in serial verbal reactions, *J. Exp. Psychol.* **18**(6), 643–662 (1935)
- 37.15 C. Craver: *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience* (Oxford Univ. Press, Oxford 2007)
- 37.16 R. Cummins: *The Nature of Psychological Explanation* (MIT Press, Cambridge 1985) p. 208
- 37.17 S. Glennan: Rethinking mechanistic explanation, *Philos. Sci.* **69**(3), 342–3353 (2002)
- 37.18 J. Woodward: What is a mechanism? A counterfactual account, *Philos. Sci.* **69**(S3), S366–S377 (2002)
- 37.19 E. Datteri, F. Laudisa: Box-and-arrow explanations need not be more abstract than neuroscientific mechanism descriptions, *Front. Psychol.* **5**, 464 (2014)
- 37.20 D.S. Touretzky, A.D. Redish: Theory of rodent navigation based on interacting representations of space, *Hippocampus* **6**, 247–270 (1996)
- 37.21 F. Suppe: *The Semantic Conception of Theories and Scientific Realism* (Univ. of Illinois Press, Champaign 1989) p. 496
- 37.22 E. Datteri, F. Laudisa: Model testing, prediction and experimental protocols in neuroscience: A case study, *Stud. Hist. Philos. Sci.* **43**(3), 602–610 (2012)
- 37.23 T. Shallice: *From Neuropsychology to Mental Structure* (Cambridge Univ. Press, Cambridge 1988)
- 37.24 F.W. Grasso, T.R. Consi, D.C. Mountain, J. Atema: Biomimetic robot lobster performs chemo-orientation in turbulence using a pair of spatially separated sensors: Progress and challenges, *Robotics Auton. Syst.* **30**(1–2), 115–131 (2000)
- 37.25 V. Braitenberg: *Vehicles. Experiments in Synthetic Psychology* (MIT Press, Cambridge 1986)
- 37.26 W. Grey Walter: An imitation of life, *Sci. Am.* **182**(5), 42–45 (1950)
- 37.27 H.A. Simon: *The Sciences of the Artificial* (MIT Press, Cambridge 1969)
- 37.28 J.L. Krichmar, A.K. Seth, D.A. Nitz, J.G. Fleischer, G.M. Edelman: Spatial navigation and causal analysis in a brain-based device modeling cortical-hippocampal interactions, *Neuroinformatics* **3**(3), 197–221 (2005)
- 37.29 T. Shallice, G.E. Hinton: Lesioning an attractor network: Investigations of acquired dyslexia, *Psychol. Rev.* **98**(1), 74–95 (1991)
- 37.30 P. Kitcher: Explanatory unification and the causal structure of the world. In: *Scientific Explanation*, ed. by P. Kitcher, W.C. Salmon (Univ. of Minnesota Press, Minneapolis 1989) pp. 410–505
- 37.31 V. Gijsbers: Why unification is neither necessary nor sufficient for explanation, *Philos. Sci.* **74**(4), 481–500 (2007)
- 37.32 N. Burgess, A. Jackson, T. Hartley, J. O’Keefe: Predictions derived from modelling the hippocampal role in navigation, *Biol. Cybern.* **83**(3), 301–312 (2000)
- 37.33 J. O’Keefe, J. Dostrovsky: The hippocampus as a spatial map: Preliminary evidence from unit activity in the freely moving rat, *Brain Res.* **34**, 171–175 (1971)
- 37.34 S.B. Russell: A practical device to simulate the working of nervous discharges, *J. Animal Behav.* **3**, 15–35 (1913)
- 37.35 R.G. Krueger, C.L. Hull: An electro-chemical parallel to the conditioned reflex, *J. Gen. Psychol.* **5**, 262–269 (1931)
- 37.36 C.L. Hull, H. Baernstein: A mechanical parallel to the conditioned reflex, *Science* **45**(2), 339–342 (1929)
- 37.37 T. Ross: Machines that think. A further statement, *Psychol. Rev.* **42**, 387–393 (1935)
- 37.38 T. Ross: The synthesis of intelligence. Its implications, *Psychol. Rev.* **45**, 185–189 (1938)
- 37.39 A. Rosenblueth, N. Wiener, J. Bigelow: Behavior, purpose and teleology, *Philos. Sci.* **10**(1), 18–24 (1943)
- 37.40 A. Rosenblueth, N. Wiener: The role of models in science, *Philos. Sci.* **12**(4), 316–321 (1945)
- 37.41 A. Newell, H.A. Simon: GPS, a program that simulates human thought. In: *Computers and Thought*, ed. by E.A. Feigenbaum, J. Feldman (McGraw-Hill, New York 1963) pp. 279–293
- 37.42 A. Newell, H.A. Simon: *Human Problem Solving* (Prentice-Hall, Englewood Cliffs 1972)
- 37.43 A. Newell, H.A. Simon: Computer simulation of human thinking, *Science* **134**(3495), 2011–2017 (1961)
- 37.44 N. Rochester, J.H. Holland, L.H. Haibt, W.L. Duda: Test on a cell assembly theory on the action of the brain using a large digital computer, *IRE Trans. Inf. Theory* **IT-2**, 80–93 (1956)
- 37.45 A. Gelperin, J.J. Hopfield, D.W. Tank: The logic of Limax learning. In: *Model Neural Networks and Behavior*, ed. by A. Selverston (Plenum Press, New York 1985) pp. 237–261
- 37.46 R.D. Hawkins: A biologically realistic neural network for higher-order features of classical conditioning. In: *Parallel Distributed Processing: Implications for Psychology and Neurobiology*, ed. by R.G.M. Morris (Clarendon Press, Oxford 1989) pp. 214–247
- 37.47 N.J. Nilsson: *Shakey the Robot*, Tech. note No. 323 (AI Center SRI International, Menlo Park 1984)
- 37.48 R.C. Arkin: *Behavior-Based Robotics* (The MIT Press, Cambridge 1998)
- 37.49 R. Reeve, B. Webb, A. Horchler, G. Indiveri, R. Quinn: New technologies for testing a model of cricket phonotaxis on an outdoor robot, *Robotics Auton. Syst.* **51**(1), 41–54 (2005)
- 37.50 B. Webb: Using robots to understand animal behavior, *Adv. Study Behav.* **38**, 1–58 (2008)
- 37.51 D. Lambrinos, R. Möller, T. Labhart, R. Pfeifer, R. Wehner: A mobile robot employing insect strategies for navigation, *Robotics Auton. Syst.* **30**(1/2), 39–64 (2000)
- 37.52 N. Franceschini, J. Pichon, C. Blanes, J. Brady: From insect vision to robot vision, *Philos. Trans. Biol. Sci.* **337**(1281), 283–294 (1992)
- 37.53 N. Franceschini, F. Ruffier, J. Serres: A bio-inspired flying robot sheds light on insect piloting abilities, *Curr. Biol.* **17**(4), 329–335 (2007)
- 37.54 P. Pyk, S. Bermúdez i Badia, U. Bernardet, P. Knüsel, M. Carlsson, J. Gu, P.F.M.J. Verschure, E. Chanie, B.S. Hansson, T.C. Pearce: An artificial moth: Chemical source localization using a robot based neuronal model of moth optomotor anemotactic

- search, *Auton. Robots* **20**, 197213 (2006)
- 37.55 M. Blanchard, F.C. Rind, P.F.M.J. Verschure: Collision avoidance using a model of the locust LGMD neuron, *Robotics Auton. Syst.* **30**(1–2), 17–38 (2000)
- 37.56 J. Ayers, J. Witting: Biomimetic approaches to the control of underwater walking machines, *Philos. Trans. Ser. A, Math. Phys. Eng. Sci.* **365**, 273–295 (2007)
- 37.57 L. Manfredi, T. Assaf, S. Mintchev, S. Marrazza, L. Capantini, S. Orofino, L. Ascari, S. Grillner, P. Wallén, Ö. Ekeberg, C. Stefanini, P. Dario: A bioinspired autonomous swimming robot as a tool for studying goal-directed locomotion, *Biol. Cybern.* **107**, 513–527 (2013)
- 37.58 C. Wilbur, W. Vorus, Y. Cao, S. Currie: A lamprey-based undulatory vehicle. In: *Neurotechnology for Biomimetic Robots*, ed. by J. Ayers, J.L. Davis, A. Rudolph (MIT Press, Cambridge 2002) pp. 285–296
- 37.59 A.J. Ijspeert, A. Crespi, D. Ryczko, J.–M. Cabelguen: From swimming to walking with a salamander robot driven by a spinal cord model, *Science* **315**, 1416–1420 (2007)
- 37.60 F. Delcomyn: Walking robots and the central and peripheral control of locomotion in insects, *Auton. Robots* **7**(3), 259–270 (1999)
- 37.61 A.J. Ijspeert: Biorobotics: Using robots to emulate and investigate agile locomotion, *Science* **346**(6206), 196–203 (2014)
- 37.62 A.J. Ijspeert: Central pattern generators for locomotion control in animals and robots: A review, *Neural Netw.* **21**, 642–653 (2008)
- 37.63 H. Kimura, Y. Fukuoka, A.H. Cohen: Adaptive dynamic walking of a quadruped robot on natural ground based on biological concepts, *Int. J. Robotics Res.* **26**(5), 475–490 (2007)
- 37.64 J.G. Fleischer, J.A. Gally, G.M. Edelman, J.L. Kirchner: Retrospective and prospective responses arising in a modeled hippocampus during maze navigation by a brain-based device, *Proc. Natl. Acad. Sci. USA* **104**, 3556–3561 (2007)
- 37.65 G.N. Reeke, O. Sporns, G.M. Edelman: Synthetic neural modeling: The “Darwin” series of recognition automata, *Proc. IEEE* **78**(9), 1498–1530 (1990)
- 37.66 M.J. Pearson, A.G. Pipe, C. Melhuish, B. Mitchinson, T.J. Prescott: A robotic active touch system modeled on the rat whisker sensory system, *Adapt. Behav.* **15**, 223–240 (2007)
- 37.67 T.J. Prescott, M.J. Pearson, B. Mitchinson, J.C.W. Sullivan, A.G. Pipe: Whisking with robots: From rat vibrissae to biomimetic technology for active touch, *IEEE Robotics Autom. Mag.* **16**, 42–50 (2009)
- 37.68 M. Rucci, G.M. Edelman, J. Wray: Adaptation of orienting behavior: From the barn owl to a robotic system, *IEEE Trans. Robotics Autom.* **15**(1), 96–110 (1999)
- 37.69 C.–P. Chou, B. Hannaford: Study of human forearm posture maintenance with a physiologically based robotic arm and spinal level neural controller, *Biol. Cybern.* **76**(4), 285–298 (1997)
- 37.70 S. Eskiizmirli: A model of the cerebellar pathways applied to the control of a single-joint robot arm actuated by McKibben artificial muscles, *Biol. Cybern.* **86**, 379–394 (2002)
- 37.71 D. Floreano, A.J. Ijspeert, S. Schaal: Robotics and Neuroscience, *Curr. Biol.* **24**(18), R910–R920 (2014)
- 37.72 M. Suzuki, D. Floreano, E.A. Di Paolo: The contribution of active body movement to visual development in evolutionary robots, *Neural Netw. Off. J. Int. Neural Netw. Soc.* **18**(5/6), 656–665 (2005)
- 37.73 S. Nolfi, D. Floreano: *Evolutionary Robotics: The Biology, Intelligence, and Technology of Self-Organizing Machines* (MIT Press, Cambridge 2001) p. 336
- 37.74 J. Long: *Darwin’s Devices: What Evolving Robots Can Teach Us About the History of Life and the Future of Technology* (Basic Books, New York 2012) p. 281
- 37.75 A. Karniel, M. Kositsky, K.M. Fleming, M. Chiappalone, V. Sanguineti, S.T. Alford, F.A. Mussa-Ivaldi: Computational analysis in vitro: Dynamics and plasticity of a neuro-robotic system, *J. Neural Eng.* **2**(3), 250–265 (2005)
- 37.76 F.A. Mussa-Ivaldi, S.T. Alford, M. Chiappalone, L. Fadiga, A. Karniel, M. Kositsky, E. Maggioni, S. Panzeri, V. Sanguineti, M. Semprini, A. Vato: New perspectives on the dialogue between brains and machines, *Front. Neurosci.* **4**(1), 44–52 (2010)
- 37.77 P.V. Zelenin, T.G. Deliagina, S. Grillner, G.N. Orlovsky: Postural control in the lamprey: A study with a neuro-mechanical model, *J. Neurophysiol.* **84**(6), 2880–2887 (2000)
- 37.78 E. Datteri: Simulation experiments in bionics: A regulative methodological perspective, *Biol. Philos.* **24**(3), 301–324 (2009)
- 37.79 W.S. Parker: Does matter really matter? Computer simulations, experiments, and materiality, *Synthese* **169**(3), 483–496 (2009)
- 37.80 P. Humphreys: *Extending Ourselves: Computational Science, Empiricism, and Scientific Method* (Oxford Univ. Press, Oxford 2004) p. 182
- 37.81 E. Winsberg: Simulated experiments: Methodology for a virtual world, *Philos. Sci.* **70**(1), 105–125 (2003)
- 37.82 K.N. Jaax, B. Hannford: A biorobotic structural model of the mammalian muscle spindle primary afferent response, *Ann. Biomed. Eng.* **30**(206), 84–96 (2001)
- 37.83 B. Raman, A. Gutierrez-Galvez, A. Perera-Lluna, R. Gutierrez-Osuna: Sensor-based machine olfaction with a neurodynamics model of the olfactory bulb, *IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)* (2004) pp. 319–324
- 37.84 F.C. Rind: Motion detectors in the locust visual system: From biology to robot sensors, *Microsc. Res. Tech.* **56**(4), 256–269 (2002)
- 37.85 C. Imbert: Relevance, not invariance, explanatory, not manipulability: Discussion of Woodward’s views on explanatory relevance, *Philos. Sci.* **80**, 625–636 (2013)
- 37.86 J. Woodward: *Making Things Happen: A Theory of Causal Explanation* (Oxford Univ. Press, Oxford 2003)

- 37.87 J.C. Schank, C.J. May, J.T. Tran, S.S. Joshi: A biorobotic investigation of Norway rat pups (*Rattus norvegicus*) in an arena, *Adapt. Behav.* **12**(3/4), 161–173 (2004)
- 37.88 T. Voegtlin, P. Verschure: What can robots tell us about brains? A synthetic approach towards the study of learning and problem solving, *Rev. Neurosci.* **10**(3–4), 291–310 (1999)
- 37.89 B. Webb: Animals versus animats: Or why not model the real Iguana?, *Adapt. Behav.* **17**(4), 269–286 (2009)
- 37.90 R. Pfeifer, J. Bongard: *How the Body Shapes the Way We Think. A New View of Intelligence* (MIT Press, Cambridge 2007)
- 37.91 J.C. Spagna, D.I. Goldman, P.-C. Lin, D.E. Koditschek, R.J. Full: Distributed mechanical feedback in arthropods and robots simplifies control of rapid running on challenging terrain, *Bioinspiration & Biomim.* **2**(1), 9–18 (2007)
- 37.92 T. Ziemke: On the role of robot simulations in embodied cognitive science, *AISB Journal* **1**(4), 389–399 (2003)
- 37.93 W. Gerstner, H. Sprekeler, G. Deco: Theory and simulation in neuroscience, *Science* **338**(6103), 60–65 (2012)
- 37.94 H. Markram: The blue brain project, *Nat. Rev. Neurosci.* **7**, 153–160 (2006), doi:[10.1038/nrn1848](https://doi.org/10.1038/nrn1848)

---

# Models in Part H

## Part H Models in Physics, Chemistry and Life Sciences

Ed. by Mauro Dorato, Matteo Morganti

**38 Comparing Symmetries in Models and Simulations**

Giuseppe Longo, Paris, France  
Maël Montévil, Paris, France

**39 Experimentation on Analogue Models**

Susan G. Sterrett, Wichita, USA

**40 Models of Chemical Structure**

William Goodwin, Tampa, USA

**41 Models in Geosciences**

Alisa Bokulich, Boston, USA  
Naomi Oreskes, Cambridge, USA

**42 Models in the Biological Sciences**

Elisabeth A. Lloyd, Bloomington, USA

**43 Models and Mechanisms in Cognitive Science**

Massimo Marraffa, Rome, Italy  
Alfredo Paternoster, Bergamo, Italy

**44 Model-Based Reasoning in the Social Sciences**

Federica Russo, Amsterdam,  
The Netherlands

This part of the Handbook is devoted to the issue of the use of models in physics, chemistry and the life sciences. As is well known, the question of the role of scientific models and representations has been thoroughly explored, and the available literature is rather extensive. The novelty of the chapters collected in this part of the Handbook is constituted by their focus on domains of scientific inquiry that so far have not received particular attention (if they have received any attention at all). Indeed, in the past authors dealing with scientific models almost exclusively focused on the role of models in physics. The other sciences have been left in the background, perhaps under the implicit, but unjustified, assumption that, once the main philosophical questions of models in physics had been clarified, the results could be straightforwardly extended to the other domains, and in particular to geology, biology, chemistry, and the other natural sciences in general. This physicalistic *sciovinism* is an undeniable but not very beneficial heritage of twentieth century philosophy of science: given the fact that the major scientific revolutions in the first part of this century happened in the domain of physics, philosophers of science tried to solve questions about explanation, realism, evidence, laws, and models by just looking at physics and its mathematical models. However, the scientific enterprise has proven to be much more multifaceted and polychromic than the unified picture handed over to us by the former generations of philosophers has made us believe. Of course, there were significant exceptions also in the last century, but a full awareness of the internal complexity of science and scientific theorizing is probably yet to be achieved. To provide just one example, the widespread conviction that biology has no laws because it is an historical science essentially characterized by contingent factors – as well as the fact that mathematics allegedly lacks a relevant role in biology – has led many philosophers to ignore biology altogether in their reflections on science.

One of the main merits of this section of the Handbook therefore lies in its pluralistic perspective. Chemistry, biology, geology – while regarded as importantly connected to physics on several counts – are all considered on their own merits and treated in depth. The role played by models in these disciplines, that is, is evaluated by the contributors in the specific terms that are appropriate in the relevant domains and theories. And the same holds for the chapters focusing on cognitive science and the social sciences. Despite – and perhaps in virtue of – the fact that each of them presupposes a well-focused, specific perspective, the chapters that follow also give new important insights on the topic of models *in general*. The upshot is that the best way to

understand how models are constructed and applied to the natural world consists in trying to achieve a synthetic or – to put it in Plato's terms – a *synoptic* grasp of the different issues involved in the various disciplines, thereby looking at specific domains and case studies while, at the same time, not losing sight of the forest in favor of the various trees that constitute it.

The chapters contained in this part of the Handbook are all clear, remarkably comprehensive, full of new arguments, and enriched by a wealth of images, graphs, and other forms of visual support. We are sure that readers will find the careful look at new themes and the in-depth analysis of particular aspects of model building and model application particularly stimulating. Before closing this introduction, however, let us briefly summarize each of the chapters that constitute the present part of the Handbook.

In *Chap. 38*, *Giuseppe Longo* and *Maël Montevil* explain how computer simulations brought important novelties in the domain of knowledge construction. In their chapter, they distinguish between mathematical modeling, computer implementations of these models, and purely computational approaches. In all three cases, they suggest that the questions that may be posed concerning the processes under investigation receive different answers. These differences are explored by looking at the various theoretical symmetries that are at work in each framework.

In *Chap. 39*, *Susan Sterrett* carries out a very detailed examination of the role of analogue models in experimentation. Analogue models are actual physical setups used to model something else, useful when what one wishes to investigate is difficult to observe or experiment upon, due to size or distance in space or time. Sterrett describes and discusses several experiments involving analogue models, and the tools for constructing them and interpreting their results.

In *Chap. 40*, *William Goodwin* aims to identify the most significant philosophical insights that have emerged out of the increased interest in scientific models, and to reflect on these insights in the context of chemistry, a discipline that has been relatively neglected in the philosophical literature. Goodwin argues that in chemistry the centrality and significance of models to the scientific enterprise is manifest, and that chemistry is a clear, useful, and interesting context in which to consider general philosophical questions about the nature and role of models in science.

In *Chap. 41*, *Alisa Bokulich* and *Naomi Oreskes* discuss models in geomorphology and the earth sciences, very probably introducing many readers to this

topic for the first time. Their contribution sheds light on the nature of modeling and idealization in sciences that have been as unduly neglected by the philosophical community at large. The geosciences, in particular, deal with issues related to environment, climate and, more generally, a number of complex features determined by physical, chemical, and biological processes that operate at the surface of the earth. As such, they are obviously of the utmost relevance, if only because they promise to provide understanding, and perhaps better control, of the very milieu in which we live.

In [Chap. 42](#), *Elisabeth Lloyd* looks at models in the biological sciences. In particular, starting from the insight that much of evolutionary theory today relies on mathematical models, Lloyd discusses in detail various ways to describe the models that make up evolutionary theory. Special focus is put on the representation of genetic states and changes in a population, which is a crucial part of population genetics theory.

In [Chap. 43](#), *Massimo Marraffa* and *Alfredo Paternoster* study the role of models and mechanisms in cognitive science. The authors give a very well-informed presentation and discussion of the use of models in the cognitive sciences, with special focus on computational models. Their suggestion is that – in

spite of undeniable difficulties in their integration with dynamical aspects – computational models are still to be regarded as fundamental for the study of the mind, in particular in virtue of their clear explanatory significance.

Finally, in [Chap. 44](#), *Federica Russo* discusses the role of model-based reasoning in the social sciences. This chapter provides an overview of various forms of model-based reasoning in social research and discusses the use of experiments and simulations in the study of social contexts. Russo also investigates the links between model-based reasoning and other key philosophical notions, such as explanation, causality, truth, and validity.

While this brief overview cannot do justice to the wealth of novel ideas that are to be found in this part, we hope that it suffices to signal to the readers the extreme importance of taking into account the role of models in those sciences that may supervene on physics, but are not straightforwardly reducible to it. Thus, in expressing our gratitude to the authors for accepting our invitation and for their hard work on their chapters, we close this introduction here and leave the rest of this part of the *Springer Handbook of Model-Based Science* to our readers.



## 38. Comparing Symmetries in Models and Simulations

Giuseppe Longo, Maël Montévil

Computer simulations brought remarkable novelties to knowledge construction. In this chapter, we first distinguish between mathematical modeling, computer implementations of these models and purely computational approaches. In all three cases, different answers are provided to the questions the observer may have concerning the processes under investigation. These differences will be highlighted by looking at the different theoretical symmetries of each frame. In the latter case, the peculiarities of agent-based or object oriented languages allow to discuss the role of phase spaces in mathematical analyses of physical versus biological dynamics. Symmetry breaking and randomness are finally correlated in the various contexts where they may be observed.

38.1	<b>Approximation</b> .....	844
38.2	<b>What Do Equations and Computations Do?</b> .....	845
38.2.1	Equations .....	845
38.2.2	From Equations to Computations .....	845
38.2.3	Computations.....	847
38.3	<b>Randomness in Biology</b> .....	848
38.4	<b>Symmetries and Information in Physics and Biology</b> .....	849
38.4.1	Turing, Discrete State Machines and Continuous Dynamics .....	849
38.4.2	Classifying Information .....	851
38.5	<b>Theoretical Symmetries and Randomness</b> .....	852
	<b>References</b> .....	854

Mathematical and computational modeling has become crucial in the natural sciences, as well as in architecture, economics, humanities and more. Sometimes the two modeling techniques are conflated into, or identified with natural processes, typically over continuous or discrete structures, by considering nature either intrinsically continuous or discrete, according to the preferences of the modeler.

Here, we analyze the major differences that exist between continuous (mostly equational) and computational (mostly discrete and algorithmic) modeling, often referred to as computer simulations. We claim that these different approaches to modeling propose different insights into the intended processes: they actually organize nature (or the object of study) in deeply different ways. This may be understood by an analysis of symmetries and symmetry breakings, which are often implicit but strongly enforced by the use of mathematical structures.

We organize the world by symmetries. They constitute a fundamental *principle of (conceptual) construction* [38.1], from Greek geometry to twentieth century physics and mathematics. All axioms by Euclid may

be understood as “maximizing the symmetries of the construction” [38.2]. Euclid’s definitions and proofs proceed by rotations and translations, which are symmetries of space.

Symmetries govern the search for invariants and their preserving transformations that shaped mathematics from Descartes spaces to Grothendieck toposes and all twentieth century mathematics [38.3]. Theoretical physics has been constructed by sharing with mathematics the same principle of (conceptual) construction. Amongst them are symmetries, which describe invariance, and order, which is needed for optimality. Symmetries and order play a key role in theoretical physics, from Galileo’s inertia to the geodetic principle and to Noether’s theorems [38.4–6]. The fundamental passage from Galileo’s symmetry group, which describes the transformation from an inertial frame to another while preserving the theoretical invariants, to Lorentz–Poincaré group characterizes the move from classical to relativistic physics. The geodetic principle is an extremizing principle and a consequence of conservation principles, that is, of symmetries in equations (Noether).

Well beyond mathematics and modern physics, the choice of symmetries as organizing principle is rooted in our search for invariants of action, in space and time, as moving and adaptive animals. We adjust to changing environments by trying to detect stabilities or by forcing them into the environment. Our bilateral symmetry is an example of this evolutionary adjustment between our biological structure and movement: its symmetry plane is given by the vertical axis of gravitation and the horizontal one of movement. The Burgess fauna, some 520 million years ago [38.7] seems to present many cases

of *asymmetric* beasts among these early multicellular organisms, later negatively selected. In this perspective, the role we give to symmetries in mathematics and physics is grounded in prehuman relations to the physical world, well before becoming a fundamental component of our scientific knowledge construction.

By this, we claim that an analysis in theorizing and modeling of the intended symmetries and symmetry breakings is an essential part of an investigation of their reasonable effectiveness and at the core of any comparative analysis.

## 38.1 Approximation

Before getting into our main theme, let's first clarify an *obvious* issue that is not so obvious to many: discrete mathematical structures are not an approximation of continuous ones. They simply provide different insights. Thus, in no way will we stress the superiority of one technique over the other. We will just try to understand continuous versus discrete frames in terms of different symmetries.

It should be clear that, on the one hand, we do not share the view of many, beautifully expressed by René Thom, on the intrinsically continuous nature of the World, where the discrete is just given by singularities in continua. On the other hand, many mythical descriptions of a computational world, or just of the perfection of computational modeling, seem to ignore the limits of discrete approximation as well as some more basic facts in numerical analysis (the first author's first teaching job), which have always been well known. When the mathematical description yields some sensitivity to initial or border conditions, there is no way to approximate that are long enough continuous nonlinear dynamics by an algorithm on discrete data types. Given any digital approximation, the discrete and the continuous trajectories quickly diverge by the combination of the round-off and the sensitivity. However, in some cases (some hyperbolic dynamics), the *discrete trajectory may be indefinitely approximated by a continuous one*, but not conversely. The result is proved by difficult "shadowing theorems" [38.8]. Note that this is the opposite of the *discrete approximating the continuum*, which is taken for granted by many.

Here, we are hinting at a comparison between mathematical techniques that provably differ. These techniques say nothing about the actual physical process, as these are not continuous nor discrete, per se, in our mathematical sense, they are what they are. Yet, it is very easy to check an algorithmic description of a double pendulum against the actual physical device (on sale

for 50 euros on the web): very soon the computational imitation has nothing to do with the actual dynamics. The point is that there is no way to have a physical double pendulum to iterate exactly on the *same* initial conditions (i. e., when started on the same interval of the best possible measurement), as this device is sensitive to minor fluctuations (thermal, for example), well below the unavoidable interval of measurement. By principle and *in practice*, instead, discrete data types allow exact iteration of the computational dynamics starting on exactly the same initial data. Again, this is a difference in symmetries and their breaking.

In conclusion a mathematical analysis of the equations allows to display sensitivity properties, from *mixing*, a weak form of chaos, to high dependence on minor variations of the initial conditions. These are mathematical properties of deterministic chaos. We stress by this that deterministic chaos and its various degrees are a property of the *mathematical model*: by a reasonable abuse one may then say that the modeled physical process is chaotic, if one believes that the mathematical model is a *good/faithful/correct* representation of the intended process. But this is an abuse: the dice or a double pendulum knows very well where it will go: along a unique physical geodesics, extremizing a Lagrangian action, according to the Hamilton principle. It is our problem if we are not able to predict it due to the nonlinearity of the model, which *amplifies fluctuations*, and due to our approximated measurements.

As it happens, the interval of measurement (the unavoidable approximated interface between us and the world) is better understood by continua than over discrete data types (we will go back to this) and, thus, physicists usually deal with equations within continuous frames.

However, the power of discrete computations allows to compute, even forever. By this, an implemented computation gives fantastic images of deterministic chaos.

As a matter of fact, this was mathematically described and perfectly understood by Poincaré in 1892, yet it came to the limelight only after Lorentz's computational discovery of "strange attractors" (and Ruelle's work [38.9]). As deterministic chaos is an asymptotic notion, there is no frame where one can better see chaotic dynamics, strange attractor, or alike than on a computer. Yet, just push the restart button and the most chaotic dynamics will iterate exactly, as we observe and further argue below, far away from any actual physical possibility. And this is not a minor point: it

is *correctness of programs*, a major scientific issue in computer science. Of course, one can artificially break the symmetry, by asking a friend to change the 16th decimal in the initial condition. Then, the chaotic dynamics will follow a very different trajectory on the screen, an interesting information, per se. However, our analysis here is centered on symmetry breaking intrinsic to a theory, that is, on changes which have a physical meaning. This control, which is available in computer simulations, is thus an artifact from a physical perspective.

## 38.2 What Do Equations and Computations Do?

### 38.2.1 Equations

In physics, equations depend on symmetries, either in equilibrium systems, where equations are mostly derived from conservation properties (which are symmetry properties), or in far from equilibrium systems, where equations describe flows, at least in the stationary cases – very little is known in nonstationary cases. This is the physical meaning of most equational descriptions.

One *computes* from equations and, in principle, derives knowledge on physical processes. This is possible by obtaining and discussing solutions – or the lack of solutions: a proof of nonanalyticity such as Poincaré's Three Body Theorem for example, may be very informative. But these derivations are not just formal: they are mostly based on proofs of relevant theorems. The job of mathematical deductions, in physics in particular, is to develop the consequences of *meaningful* writings. Mathematics is not a formal game of signs, but a construction grounded on meaning and handled both by formal "principles of proofs" and by semantically rich "principles of constructions" [38.1]. Typically, arguments are given by *symmetry reasons* or are based on order properties (including well-ordering). Moreover, a mathematical proof may use the genericity of the intended mathematical object or generalized forms of induction that logicians analyze by very large cardinals, an extension of the order of integer numbers obtained by alternating limits and successor operations [38.10]. Once more, theoretical symmetries and meaning step in while proving theorems and solving/discussing equations; also the progress from Laplace's predictability of deterministic process to Poincaré's proof of deterministic though unpredictable processes is a breaking of the observable symmetries (see below for more).

As a matter of fact, we invented very original mathematical structures, from Galois' groups to differential geometry, in order to solve equations or discuss their

solvability. The use of enriched construction principles that are often based on or yielding new mathematical meaning has constantly been stimulated by the analysis of equations. This is part of the common practice of mathematical reasoning. However, well beyond the extraordinary diagonal trick by Gödel, it is very hard to *prove* that *meaningful* procedures are unavoidable in actual proofs, that is to show that meaning is essential to proofs. An analysis of a recent *concrete* incompleteness result is in [38.11]: this means that geometric meaning inevitably steps in proofs even of combinatorial theorems (of arithmetic!), where meaning takes the form of well-ordering which is a geometric judgement. Or very large infinite cardinals may be shown to be essential to proofs [38.12]. In this precise sense, formal deductions as computations, with their finitistic principles of proof, are provably incomplete.

In particular, physicomathematical deductions, used to discuss and solve equations, are *not* just formal computations, i. e., meaningless manipulations of signs. They transfer symmetries in equations to further symmetries, or prove symmetry changes or breaking (non-analyticity, typically). In category theory, equations are analyzed by drawing diagrams and inspecting their symmetries.

### 38.2.2 From Equations to Computations

The mathematical frame of modern computers was proposed within an analysis of formal deductions. In fact Gödel, Kleene, Church, Turing . . . invented computable functions in the 1930s in order to disprove the largely believed completeness hypothesis of formal/axiomatic systems and their formally provable consistency. It is not by chance that an immense mathematical physicist, *H. Weyl*, was one of the few who claimed that the formalist/computational project was trivializing mathematics and conjectured incompleteness, already in

1918 [38.13] (see also [38.1]). Turing, in particular, imagined the logical computing machine imitating a man in the least action of sign manipulation according to formal instructions (write or erase 0 and 1, move left or right of one square in a *child's notebook*), and invented by this the modern split between software and hardware. He then wrote an equation that easily defines an incomputable arithmetic function. Turing's remarkable work for this negative result produced the modern notion of program and digital computer, a discrete state machine working on discrete data types. As we said, computing machinery was invented as an implementation of formal proofs and is provably incomplete even in an arithmetic, let alone proper extension of it, based on principles richer than arithmetic induction (well-ordering, symmetries, infinite ordinals . . .).

Thus, beyond the limits set by the impossibility of approximation mentioned above, there is also a conceptual gap between proving over equations and computing solutions by algorithms on discrete data. The first deals with the physical *meaning* of equations and their symmetries and their breaking, it transfers this meaning to consequences by human reasoning grounded on *gestures* (such as drawing a diagram) and common understanding. It is based on the invention, if needed, of new mathematical structures, possibly infinitary ones from Galois' groups to Hilbert spaces to the modern fine analysis of infinitary proofs [38.14]. These may even be proved to be unavoidable in some cases, such as for well-ordering or the large infinite cardinals mentioned above, well beyond computations and formalisms (see the reference above). Do algorithms transfer "physical meaning" along the computation? Do they preserve symmetries? Are those broken in the same way we understand they are in the natural process under scrutiny?

Our claim is that algorithmic approaches (with the notable exception of interactive automated formal calculus, within its limits) involve a modification of the theoretical symmetries used to describe and understand phenomena in physics, in particular by continua. This means that algorithmic approaches usually convey less or a different physical meaning than the original equation approaches. In other words, the modification of the equations needed for a completely finitary and discrete approach to the determination of a phenomenon leads to losses of meaningful aspects of the mathematization and to the introduction of arbitrary or new features.

As far as losses are concerned, the most preeminent ones probably stem from the departure from the continuum, an invention resulting from measurement from Pythagoras' theorem to the role of intervals in physical measurement. As we already hinted, deterministic unpredictability does not make sense in the computing

world. A program determines and computes on exact data: when those are known, exactly (which is always possible), the program iterates exactly, thus allowing a perfect prediction as the program itself yields the prediction. The point is that deterministic unpredictability is due to the nonlinearity, typically, of the *determination* (the equations) and triggered by nonobservable fluctuations or perturbations *below* the (best) interval of measurement. Now, approximation in mathematics is handled by topologies of open intervals over continua, the so-called *natural topology* over the real numbers.

With regards to this, note that a key assumption bridging mathematics of continua and classical physics is that any sequence of measurements of increasing, arbitrary precision converge to a well-defined state. This is mathematically a Cauchy condition of completeness, which implies that the rational numbers are not sufficient to understand the situation. Cantor's real numbers have been invented exactly to handle these kinds of problems (among other reasons, such as the need to mathematize rigorously the phenomenal continuum in its broadest sense, say the continuum of movement).

Also, the fundamental relation between symmetries and conservation properties exhibited by Noether's theorems depend on the continuum (e.g., continuous time translations), so these results can no longer be derived on a discretized background. In short, these theorems rely on the theoretical ability to transform states continuously along continuous symmetries in equations (of movement, for example) since the intended conserved quantity cannot change during such a transformation. With a discrete transformation the observed quantities can be altered (usually the case in simulations) because there is no continuity to enforce their conservation.

Reciprocally, the changes due to the discretization introduce features that are arbitrary from a physical perspective. For example, a basic discretization of time introduces an arbitrary fundamental time-scale. In numerical analysis, the methodology is to have the (differential) equations as the locus of objectivity and to design algorithms that can be shown to asymptotically converge (in a pertinent mathematical sense, and hopefully rapidly in practice) towards the mathematical solutions of the physically meaningful equations. In these frames, the theoretical meaning of the numerical (or algorithmic) approaches is entirely derivative: such numerical approaches are sound only with respect to, and inasmuch as there are mathematical results showing a proximity with the original equations and the trajectories determined by them. The mathematical results (convergence theorems) define the nature of this proximity and are usually limited to specific cases so that entire research communities develop around the topic of the simulation of a specific family of equations

(Navier–Stokes or alike for turbulence, Schrödinger in quantum physics, ...). As a result, the methods to approach different (nonlinear) equations by computing rely on specific discretizations and their close, often ad hoc, analysis.

### 38.2.3 Computations

As we said, we are just singling-out some methodological differences or gaps between different modeling techniques. On the *side of algorithms*, the main issue we want to stress here is that equational approaches force uniform phase spaces. That is, the list of pertinent observables and parameters, including space and/or time, of course, must be given a priori. Since the work by Boltzmann and Poincaré, physicists usually consider the phase space, which is made out of position and momentum or energy and time, as sufficient for writing the equational determination. By generalizing the philosopher's (Kant) remark on Newton's work, the (phase) space is the very *condition of possibility* for the mathematical intelligibility of physics. Or, to put it as H. Weyl, the main epistemological teaching of relativity theory is that physical knowledge begins when one fixes the reference system (that is to say, the way to describe the phase space) and the metrics on it. Then Einstein's invariant theory allows to inspect the relevant invariants and transformations, on the grounds of Lorentz–Poincaré symmetry groups, typically, within a pre-given list of observables and parameters.

Now, there exists a rich practice of computational modeling which does not need to pass through equations, it skips this a priori structure. *Varenne* nicely describes the dynamic mixture of different computational contexts as a *simulat*, a neologism which recalls *agrégat* (an aggregate) [38.15]. This novelty has been introduced, in particular, by the peculiar features of object oriented programming (OOP), but other *agent oriented systems* exist.

As a matter of fact, procedural languages require all values to share the same representation – this is how computer scientists name observables and parameters (Technically, an existential quantifier is opened at the beginning of the program and then everyone shares all private information). *Objects* instead may interact even with completely different representations as long as their interfaces are compatible (The existentials are opened only at the point of performing the operation). Thus, objects behave autonomously and do not require knowledge of the private (encapsulated) details of those they are interacting with. As a consequence, only the interface is important for external reactions [38.16, 17].

In biological modeling, aggregating different techniques with no common a priori *phase space* is a ma-

major contribution to knowledge construction. Organisms, niches, ecosystems may be better understood by structuring them in different levels of organization, each with a proper structure of determination, that is, phase space and description of the dynamics. For example, networks of cells are better described by tools from statistical physics, while morphogenesis, for example organ formation, is currently and mostly modeled by differential equations in continua. Each of these approaches requires pre-given phase spaces, which may radically differ (and the communities of researchers in the two fields hardly talk to each other). In a computer, thanks to its high parallelism one may mix these different techniques in spite of their differences with some more or less acceptable approximations. Even more so, ad hoc algorithms may describe specific interactions independently of a unified equational description that may be impossible. Then *objects* may interact only on the grounds of the actual interface, both within a level of organization and between different levels, without reference to the proper or internal (to the object, to the level) causal structure.

In other words, OOP allows independent objects dynamics, reminiscent of individual cell dynamics. Then, proliferation with variation and motility, which is the default state of life [38.18], may be added to the models of morphogenesis that usually consider cells as inertial bullets, which they are not; that is, their proliferation, changes, and motility are not entailed by physical forces that contribute to shape organs (in particular, when organs function for the exchange of energy and matter). By the computational power of modern computers, agent or object based programming styles (such as OOP) may implement autonomous agency for each cell, have them simultaneously interact within a morphogenetic field shaping the dynamics or a network ruled by statistical laws.

In summary, in computer simulation one may *put together* all these techniques, design very complex *simulat* as aggregation of algorithms including stochastic equations, probabilities distributions and alike. In particular, OOP allows the simulation of discrete dynamics of individual cells in an organism or of organisms in an ecosystem. And this with no need to write global first equations: one directly goes to algorithms in their changing environment.

However, let the process or images on a computer run ... then push the restart button. Since access to discrete data is exact, as we said and keep stressing, the computer will iterate on the same initial conditions exactly, with the same discrete algorithms. Thus, it will go exactly along the same computation and produce exactly the same trajectories, images, and genesis of forms. This has no physical meaning as an unstable

or chaotic system would never *iterate identically*. It is even less biologically plausible, as biology is, at least, the “never identical iteration of a morphogenetic process” [38.18]. Now observe that *exact iteration* is a form of (time-shift/process-identity) symmetry; while non-identical iteration is a symmetry breaking (see below for more on randomness versus symmetry breaking).

### 38.3 Randomness in Biology

Theoretical physics proposes at least two forms of randomness: classical and quantum. They are separated by different probability theories and underlying logic: entanglement modifies the probability correlations between quantum events [38.19]. Even the outcome of the measurement of generic states is contextual which means that this outcome depends on the other measurements performed and cannot be assumed to be predefined [38.20, 21] – this situation is different from classical ones which are not contextual. A new form of randomness seems to be emerging from computer networks or, so far at least, it is treated by yet a different kind of mathematics [38.22]. In particular, some analyses of randomness are carried out without using probabilities.

In the same way that we said the world is neither intrinsically continuous nor discrete, randomness is not in the world: it is in the interface between our theoretical descriptions and *reality* as accessed by measurement. Randomness is *unpredictability with respect to the intended theory and measurement*. Both classical and quantum randomness, though different, originate in measurement.

The classical one is present in dynamics sensitive to initial or border conditions: a fluctuation or perturbation below measurement, which cannot be exact by physical principles (it is an interval, as we said) is amplified by the dynamics, becomes measurable and “[...] we have a random phenomenon” [38.23]. This amplification is mathematically described by the nonlinearity of the intended equations or evolution function with a subtle difference though. If a solution of the nonlinear system exists, then the analysis of the Lyapounov exponents, possibly, yields some information on the speed of divergence of trajectories, initially indistinguishable by measurement: a nonmeasurable fluctuation is amplified and produces an unpredictable and measurable event, yet the amplification is computable. In the case of nonexistence or nonanalyticity of solutions of the given differential equations, one may have bifurcations or unstable homoclinic trajectories (i. e., trajectories at the intersection of stable and unstable manifolds). The

Noise, of course, may be introduced artificially, but this makes a deep conceptual difference at the core of our analysis.

Note, finally, that stochastic equations, probability values and their formal or algorithmic descriptions are *expressions* and *measurement* of randomness, they *do not* implement randomness. And this is a key issue.

choice at bifurcation, thus the physical trajectory is then highly unpredictable, thus random, and may be also physically ascribed to fluctuations or perturbations below measurement. In this case, however, one generally does not have a criterion of divergence, such as Lyapounov exponents. The fluctuation or perturbation *causes* the unpredictable event, thus Curie’s principle is preserved: a physical effect cannot have a dissymmetry absent from its efficient cause – a symmetry conservation principle, or “symmetries cannot decrease”. Yet, at the level of *measured* observables one witnesses a symmetry breaking, as the causing dissymmetry cannot be observed.

Quantum randomness is grounded in noncommutativity of the measurement of conjugated variables (position and momentum or energy and time), given by a lower bound – Planck’s  $h$ . It is represented by Schrödinger’s equation that defines the trajectory of a probability amplitude (or law) in a very abstract mathematical space (a Hilbert space). As hinted above, measurement of entangled particles gives probabilities that are different from the classical contexts (Bell inequalities are not respected [38.24]).

In quantum physics, though, there is another fundamental difference: in classical and relativistic mechanics, from Aristotle to Galileo and Einstein, it is assumed that *every event has a cause*. As mentioned above in reference to Curie’s principle, the unpredictable but measurable classical event is *caused* by the (initial or border) undetectable fluctuation. Instead, in current interpretations of quantum mechanics (QM), random events may be *acausal* – the spin up/spin down of an electron, say, is pure contingency, it does not need to have a cause. This radically changes the conceptual frame – and many still do not accept it and keep looking in vain for hidden variables (hidden causes) along the classical paradigm.

Surprisingly enough, a quantum event at the molecular level may have a phenotypic effect in biology. This is the result of recent empirical evidence, summarized and discussed in [38.25]. Thus, a phenotype that is a structural property of an organism, possibly a new or

ganism, may result from an acausal event happening at a completely different level of organization (molecular versus organs or organisms). This microevent may be amplified by classical dynamics of molecules including their enthalpic oscillations and their Brownian motion. Brownian motion is omnipresent in cells' proteome, where macromolecules are very *sticky* and their chemical interactions are largely stochastic – though canalized by strong chemical affinities and cell compartmentalization. So, quantum and classical randomness may *superpose* in a highly constrained environment. Moreover, it is increasingly recognized that gene expression is mostly stochastic [38.26, 27].

This leads to the fully general fact that *macromolecular interactions and dynamics are stochastic, they must be described in terms of probabilities and these probabilities depend on the context.*

This context includes the global proteomic composition, the torsion and pressure on the chromatin [38.28], the cell activity in a tissue [38.29, 30], the hormonal cascades etc. up to the ecosystem. The up and down interactions between different levels of organization yield a proper form of biological randomness, a resonance between levels called bioresonance in [38.25]. Bioresonance destabilizes and stabilizes organisms; it both *yields* and *follows from* variability, as correlated variations contribute also to the changing structural stability of organisms. Note that variability produces adaptation and diversity at the core of biological dynamical stability: an organism, a population, or a species is *biologically stable*, while changing and

adapting because it is diverse as well. Both stability and diversity are also the result of randomness. Also, as we said randomness is highly canalized in biology by cellular compartments of molecules, tissues tensesgrity, organismal control (hormones, immune and neural systems etc.) and the ecosystem may downward influence these constraints (methylation and demethylation, which may regulate gene expression, can be induced by the environment) [38.31]. Variability and diversity are constrained by history as well: phenotypes are the result of an evolutionary history that canalizes but does not determine (at least in view of quantum events) further evolution. For example, as for historical *canalization* there are good reasons to believe that we, the vertebrates, will never get out of the *valley* of tetrapodes – at most we may lose, as some of us have, podia and keep just traces of them.

In conclusion, randomness has a constitutive role in biology as variability and diversity contribute to structural stability, beginning with gene expression. Above, we developed a comparative analysis in terms of symmetries of physical processes with respect to their equational and computational modeling. We hinted at the different ways randomness is understood in various physical and biological frames. In biology, this later issue becomes particularly relevant in view of the organizing role of randomness, also in the case of small numbers (a population of a few thousands individuals is biologically more stable when diverse). Further on, we will propose a general thesis relating randomness and symmetry breaking.

## 38.4 Symmetries and Information in Physics and Biology

### 38.4.1 Turing, Discrete State Machines and Continuous Dynamics

We already stressed the key role of invariants and invariant preserving transformations in the construction of mathematical and physical knowledge. The sharing of construction principles in these two disciplines, among which symmetry principles and order principles, are the reason of the reasonable, though limited, effectiveness of mathematics for physics: these disciplines have been actually co-constituted on the grounds of these common construction principles [38.1]. However, since so few physical processes can be actually predicted – frictions and many-body interactions, i. e. nonlinearity, are everywhere – the effectiveness of mathematics is based mostly on the reasonable intelligibility we have of a few phenomena when we can organize them in terms of in-

variants and their transformations, thus of symmetries well beyond predictability.

In the account above, changing fundamental symmetries produced the change from one theoretical frame to another, such as from classical to relativistic physics. Further useful examples may be given by thermodynamics and hydrodynamics. The irreversibility of time, a symmetry breaking, steps in the first by the proposal of a new observable, entropy; hydrodynamics assumes incompressibility and fluidity in continua, two properties that are irreducible to the quantum mechanical ones, so far.

There is a common fashion in projecting the sciences of information onto biological and even physical processes. The deoxyribonucleic acid (DNA), the brain, even the universe would be (possibly huge) programs or Turing machines sometimes set up in networks – note

that the reference to networks is newer, it followed actual network computing by many years later.

We do not discuss here the universe nor the brain. It may suffice to quote the inventor of computing by discrete state machines, *Turing* [38.32, p. 440]:

“[...] given the initial state of the machine and the input signal it is always possible to predict all future states. This is reminiscent of Laplace’s view that from the complete state of the universe at one moment of time, as described by the positions and velocities of all particles, it should be possible to predict all future states. The prediction which we are considering is, however, rather nearer to practicability than that considered by Laplace. The system of the *universe as a whole* is such that quite small errors in the initial conditions can have an overwhelming effect at a later time. The displacement of a single electron by a billionth of a centimeter at one moment might make the difference between a man being killed by an avalanche a year later, or escaping. It is an essential property of the mechanical systems which we have called ‘discrete state machines’ that this phenomenon does not occur. Even when we consider the actual physical machines instead of the idealized machines, reasonably accurate knowledge of the state at one moment yields reasonably accurate knowledge any number of steps later.”

Note that in popular references to unstable or chaotic dynamics, instead of quoting the famous *Lorentz’s butterfly effect* proposed in 1972 on the grounds of *Lorentz’s* work of 1961, one should better refer to the *Turing’s electron effect*, published in 1952.

As for the brain, *Turing* continues [38.32, p. 451]:

“The nervous system is certainly not a discrete-state machine. A small error in the information about the size of a nervous impulse impinging on a neuron, may make a large difference to the size of the outgoing impulse.”

As a matter of fact, the notions of spontaneous symmetry breaking, catastrophic instability, and random fluctuations are at the core of *Turing’s* analysis of continuous morphogenesis [38.33], far remote from his own invention of the elaboration of information by the *discrete state machine* (DSM, his renaming in 1950 of his logical computing machine of 1936).

It is worth stressing here the breadth and originality of *Turing’s* work. He first invented the split hardware/software and the DSM in logic. Then, when moving to biophysics, he invented a continuous model for mor-

phogenesis viewed just as physical matter (hardware) that undergoes continuous deformations, triggered by (continuous) symmetry breaking of an homogeneous field in a chemical reaction-diffusion system. The model is given by nonlinear equations: a linear solution is proposed, the nonlinear case is discussed at length.

A key property of *Turing’s* continuous model is that it is “a falsification” (his words [38.33, p. 37]) of the need for a (coded) *design*. This becomes clear from the further comments on the role of genes, mentioned below. In discussions reported by *Hodges* [38.34], *Turing* turns out to be against *Huxley’s* “new synthesis”, which focused on chromosomes as fully determining ontogenesis and phylogenesis [38.35]. He never refers to the already very famous 1944 booklet by *Schrödinger* [38.36], where *Schrödinger* proposes to understand the chromosomes as loci of a coding, leading to a Laplacian determination of embryogenesis as he says explicitly (“once their structure will be fully decoded, we will be in the position of Laplace’s daemon” says *Schrödinger* [38.36, Chap. 2]). As a matter of fact, in his 1952 paper *Turing* quotes only Child, D’Arcy Thompson and Waddington as biologists, all working on dynamics of forms, at most constrained (Waddington) but not determined nor *predesigned* by chromosomes. Indeed, *Turing* discusses the role of genes in chromosomes, which differ from his *morphogenes* as generators of forms by a chemical action/reaction system. He sees the function of chromosomal genes as purely catalytic and, says [38.33, p. 38]:

“Genes may be said to influence the anatomical form of the organism by determining the rates of those reactions that they catalyze [...] if a comparison of organisms is not in question, the genes themselves may be eliminated from the discussion.”

This proposal is remarkable as for the very fuzzy, ever changing notion of *gene* [38.37]. No predefined design, no coded or programmed Aristotelian homunculus in the chromosomes (the myth of the chromosomes as a program) are needed for *Turing*, the man who invented coding and programming. This is science: an explicit proposal of a (possibly new) perspective on nature, not the transfer of familiar tools (the ones he invented, in this case!) on top of a different phenomenology.

Note finally that when comparing his DSM to a woman’s brain in [38.32], *Turing* describes an *imitation game* while he talks of a *model* as for morphogenesis. This beautiful distinction, computational imitation versus continuous model, is closely analyzed in [38.38].



### 38.4.2 Classifying Information

Let's further analyze the extensive use of *information* in biology – molecular biology in particular. Information branches in at least two theories:

- Elaboration of information (Turing, Church, Kleene and many others, later consistently extended to algorithmic information theory: Martin-Loef, Chaitin, *Calude*, see [38.39]).
- Transmission of information (*Shannon*, Brillouin, see [38.40]).

In [38.41] we stressed the key differences between these two theories that are mixed up in molecular biology with unintelligible consequences in the description of the relationship of information to entropy and complexity. The two latter notions are relevant to biology, see [38.42], where Turing–Kolmogorov's elaboration theory is quoted as well as Shannon's theory. The author initially considers the second as more pertinent for biology. Later in the paper a notion of complexity as amount of information is given. This notion is actually based on the first theory and it is described as *covariant to entropy*. Finally, Shannon's theory pops up again in the paper – the more pertinent theory, according to the author, where complexity is *contravariant to entropy* – it is negentropy.

As scientific constructions, both information theories are grounded on fundamental invariants. So it has been since at least Morse's practical invention, with no theory of information transmission. Information is independent of the specific coding and the material support. We can transmit and encode information as *bip-bip*, by short and long hits, as flashes, shouts, smoke clouds etc. by bumping on wood or metal, by electricity in cables – we can do this in a binary, ternary, or other code etc. Information is the *invariant* with respect to the transformation of these coding and material supports: this is its fundamental symmetry. It is also the case in Turing's fundamental invention: the distinction between software and hardware. So, a richer theory of programming was born that was largely based on logic, typed, and typed-free languages, term rewriting systems etc. entirely independent of the specific encoding, implementation, and hardware. The computer's soul is so detached from its physical realization that Descartes dualism is a pale predecessor of this radical and most fruitful split. And when the hardware of your computer is dying, you may transfer the entire software, including the operating system, compilers and interpreters, to another computer. This symmetry by transfer is called *metempsychosis*, we think. Now, it does not apply anywhere in biology.

The DNA is not a code carrying information. There is no way to detach soft content from it and transfer it to another material structure: it cannot be replaced by metal bullets or bumps on a piece of wood. What gets transferred to ribonucleic acid (RNA) and then to proteins is a chemical and physical structure, a most relevant one, as the DNA is an extraordinary *chemical trace of a history*. And it transmits to other chemicals an entirely contingent physicochemical conformation. If a stone bumps against other stones in a river and de-forms them (in-forms them, Aristotle would say), talking of a transmission of information in the scientific invariant sense above has no meaning unless in reference to the Aristotelian sense. No informational invariants can be extracted but the ones proper to the physicochemical processes relative to stone bumping. Life is radically contingent and material: no software/hardware split. The prescientific reference to information, sometimes called *metaphorical*, has had a major misleading role. First, it did not help to find the *right invariants*. The physicochemical structure of cellular receptors, for example, has some sort of generality which yields some stereospecificity [38.43]. Yet, this is still strictly related to common chemistry that has nothing to do with an impossible abstract information theoretic description. The proposal of an invariant that was too abstract and matter-independent did not help in finding the right scientific level of invariance. Or more severely so, it forced exact stereospecificity of macromolecular interaction as a *consequence* of the information theoretic bias.

*Monod*, one of the main theoreticians of molecular biology, claims that the molecular processes are based on the oriented transmission of information [...] (in the sense of Brillouin). In [38.44], he derives from this that the “*necessarily* stereospecific molecular interactions explain the structure of the code [...] a boolean algebra, like in computers” and that “genes define completely the tridimensional folding of proteins, the epigenetic environment only excludes the other possible foldings”. Indeed, biomolecular activities “are a Cartesian mechanism, autonomous, exact, independent from external influences”. Thus, the analysis based on the search for how information could be transmitted forced an understanding inspired by the Cartesian exactness proper to computers. It also conveyed the Laplacian causal structure, Turing would say, proper to information theories. It induced the invention of exact stereospecificity which is *necessary* to explain the boolean coding! That is, stereospecificity was logically, not empirically, derived. Indeed, robust evidence had already shown the stochasticity of gene expression (see [38.27, 45, 46] and [38.47] for a recent synthesis) since 1957 [38.48].

We now know that the protein folding is not determined by the coding (yet, Monod did consider this possibility). Macromolecular interactions, including gene expression, are largely random: they must at least be given in probabilities, as we said, and these probabilities would then depend on the context. No hardware-independent boolean algebra governs the chemical cascades from DNA to RNA to proteins, also because, as we already recalled these cascades depend on the pressure and tensions on the chromatin, the proteome activities, the intracellular spatial organization, the cellular environment, and many other forms of organismal regulations, see for example [38.28, 49].

In summary, the informational bias introduced a reasoning based on Laplacian symmetries, far away from the largely turbulent structure of the proteome, empowered also by chaotic enthalpic oscillations of macromolecules. This bias was far from neutral in guiding experiments, research projects, and conceptual frames. For example, it passed by the role of endocrine disruptors of the more than 80 000 molecules we synthesized and used in the twentieth century, an increasingly evident cause of major pathologies, including cancer [38.50–52]. These molecules were not supposed to interfere with the exact molecular cascades of key-lock correspondences, a form of stereospecificity. The bias guided the work on genetically modified organisms (GMO), which have been conceived on the grounds of the “central dogma of molecular biology” and of Monod’s approach above: genetic modifications would completely guide phenotypic changes and their ecosystemic interactions [38.53].

One final point: Information theories are *code independent*, or they analyze code in order to develop general results and transmission stability as *code insensitive* (of course cryptography goes otherwise: but secrecy and code breaking are different purposes, not exactly relevant for organisms). Information on discrete data is also *dimension independent*: by a polynomial translation one may encode discrete spaces of *any finite dimension* into one dimension. This is crucial to computing since it is needed to define Turing’s universal machine, thus operating systems and compilers.

Biology instead is embedded in a physical world where the space dimension is crucial. In physics, heat

propagation and many other phenomena, typically field theories, strictly depend on space dimension. By *mean field theories* one can show that life, as we know it, is only possible in three dimensions [38.1]. Organisms are highly geometric in the sense that *geometric* implies *sensitivity to coding and dimensions*. In this sense, continuous models more consistently propose some intelligibility: in *natural* topologies over continua, that is when the topology derives from the interval of physical measurement, dimension is a topological invariant. It is a fundamental invariant in physics to be preserved in biology, unless the reader believes that he/she can live encoded in one dimension, just exchanging information like on the tape of a Turing machine. A rather flat universe . . . yet, with no loss of information. But where one has only information, not life.

Missing the right level of invariance and, thus, the explanatory symmetries, is a major scientific mistake. Sometimes, it may seem just a *matter of language*, as if language mattered little, or a matter of informal metaphors, as if metaphors were not carrying meaning, forcing insight and guiding experiments. They actually transfer the conceptual structure or the intended symmetries of the theory they originate from, in an implicit thus more dangerous and unscientific way. Just focusing on language, consider the terminology used when referring to DNA/RNA as the *universal code for life* since all forms of life are based on it. This synchronic perspective on life – all organisms yield these molecules and the basic chemical structure of their interactions, *thus* there is a universal code – misses the historical contingency of life. There is no universality in the informational sense of an invariant code with respect to an independent hardware. Life is the historical result of contingent events, the formation somewhere and somehow of DNA or RNA or both, sufficiently isolated in a membrane, which occurred over that hardware only. Then, the resulting cell reproduced with variation and diversified to today’s biological diversity. Life history has one contingent material origin, then there was a diversification of that matter, of that specific hardware and no other. Invariance, symmetries and their breaking are different from those proper to *information*, in this strictly material, evolutionary perspective.

## 38.5 Theoretical Symmetries and Randomness

In this section, we would like to elaborate on a *thesis*, already hinted at in [38.6]. In physical theories, where the specific trajectory of an object is determined by its theoretical symmetries, we propose that randomness appears when there is a change in some of these symme-

tries along a trajectory and, reciprocally, that changes of symmetries are associated to randomness.

Intuitively, theoretical symmetries allow to understand a wide set of phenomenal situations as equivalent. At the end of the day, the trajectory that a physi-

cal object will follow, according to a theory, is the only trajectory which is compatible with the theoretical symmetries of a given system. Symmetries, in this context, enable to understand conservation properties, the uniqueness of the *entailed* trajectory, and ultimately the associated prediction, if any.

Now what happens when, over time or with respect to a pertinent parameter, a symmetry of the system is broken? A symmetry corresponds to a situation where the state or the set of possible states and the determination of a system does not change according to specific transformations (the symmetries). After the symmetry breaking, the state(s) becomes no longer invariant by these transformations; typically, the trajectory goes to one of the formerly symmetric states and not to the others (a ball on top of a mathematical hill falls along *one* of the equivalent sides). Since the initial situation is exactly symmetric (by hypothesis), all the different *symmetric* states are equivalent and there is no way to single out any of them. Then, in view of the symmetry breaking, the physical phenomena will nevertheless single out one of them. As a result, we are confronted with a nonentailed change: it is a random change.

This explanation provides a physicomathematical meaning to the philosophical notion of contingency as non-necessity: this description of randomness as symmetry breaking captures contingency as a lack of entailment or of necessity in an intended theory. Note that usually the equivalent states may not be completely symmetric as they may be associated to different probabilities, nevertheless they have the same status as *possible* states.

For now, we discussed the situation at the level of the theoretical determination alone, but the same reasoning applies *mutadis mutandis* to prediction. Indeed, we access a phenomenon by measurement, but measurement may be associated to different possible states, not distinguishable individually. Thus, these states are symmetric with respect to the measurement, but the determination may be such that these (nonmeasurably different) states lead to completely different measurable consequences. This reasoning is completely valid only when the situation is such for all allowed measurements, so that randomness cannot be associated to the possible crudeness of an arbitrary specific measurement.

Reciprocally, when we consider a random event, it means that we are confronted with a change that cannot be entailed from a previous observation (and the associated determination). When the possible observations can be determined (known phase space), this means that the different possibilities have a symmetric status before the random event (precisely because they are all predefined possibilities) but that one (or several of them) is singled out by the random event in the sense

that it becomes the actual state. We recognize in this statement the description of a symmetry that is broken during the random event.

Let us now review the main physical cases of randomness:

- Spontaneous symmetry breaking in quantum field theories and theories of phase transitions (from a macroscopic viewpoint) are the most straightforward examples of the conjecture we describe. In these cases, the theoretical determination (Hamiltonian) is symmetric and the change of a parameter leads the systems equilibrium to shift from a symmetric state to an asymmetric one (for example isotropy of a liquid shifting to a crystal with a specific orientation). Randomness then just stems from the *choice* of a specific orientation, triggered by fluctuations in statistical mechanics.
- Classical mechanics can, in spite of its deterministic nature, lead to unpredictability as a consequence of the symmetrizing effect of measurement on one side (there are always different states which are not distinguished by a measurement), and a determination that leads those states to diverge (which breaks the above symmetry). This reasoning applies to chaotic dynamics but also to phase transitions where, from a strictly classical viewpoint, fluctuations below the observation determine the orientation of the symmetry changes.
- In classical probabilities applied to *naive* cases such as throwing a dice or to more sophisticated frameworks such as statistical mechanics, our reasoning also applies. When forgetting about the underlying classical mechanics the probabilistic framework is a strict equivalence between different possibilities, except for their expected frequencies which may differ: those are given by the associated probabilities. In order to theoretically define these probabilities, some underlying theoretical symmetries are required. In our examples, the symmetries are the symmetry between the sides of a dice and for statistical mechanics, the symmetry between states with the same energy for the microcanonical ensemble. From a strictly classical viewpoint, these symmetries are assumed to be established on average by the properties of the considered dynamics. In the case of dice, it is the rotation, associated with the dependence on many parameters which leads to a sufficient mixing, generating the symmetry between the different sides of the dice. In the case of statistical mechanics, it is the property of topological mixing of chaotic dynamics (a property met by these systems by definition). This property is assumed in order to justify the validity of

statistical mechanics from the point of view of classical mechanics. In both cases, a specific state or outcome corresponds to a breaking of the relevant symmetry.

- In quantum mechanics, the usual determination of the trajectory of a state is deterministic, randomness pops out during measurement. The operator corresponding to the measurement performed establishes a symmetry between its different eigen vectors, which also correspond to the different outcomes corresponding to the eigen values. This symmetry is partially broken by the state of the system which provides different weights (probabilities) to these possibilities. The measurement singles out one of the eigen vectors which becomes the state of the system and this breaks the former symmetry.

We can conclude from this analysis and these examples that randomness and symmetry breaking are tightly associated. We can put this relationship into one sentence: *A symmetry breaking means that equivalent directions become no longer equivalent and precisely because the different directions were initially equivalent (symmetric) the outcome cannot be predicted.* As discussed elsewhere [38.6, 54], we assume that theoretical symmetries in biology are unstable. It follows that randomness, understood as associated to symmetry breaking, should be expected to be ubiquitous; however, this approach also leads to propose a further form of randomness. In order to show that randomness can be seen as a symmetry breaking, we needed to assume that the set of possibilities was determined before the event. In biology, the instability of the theoretical symmetries does not allow such an assumption in general. On the

opposite, a new form of randomness appears through the changes of phase spaces and this randomness does not take the form of a symmetry breaking *stricto sensu* inasmuch as it does not operate on a predefined set. In other words, these changes cannot be entailed but they cannot even be understood as the singling out of one possibility among others – the list of possibilities (the phase space) is not pre-given.

In brief, theoretical symmetries in physics enable to single-out a specific trajectory in a phase space, formed by a combination of observables. Thus, a symmetry breaking corresponds to the need of one or several supplementary quantities to further specify a system on the basis of already defined quantities (which were formerly symmetric and thus not useful to specify the situation). In biology, instead, the dynamic introduces new observable quantities which get integrated with the determination of the object as the latter is associated with the intended quantities and symmetries. This dynamic of the very phase space may be analyzed a posteriori as a symmetry breaking. Thus, randomness moves from within a phase space to the very construction of a phase space, a major mathematical challenge.

**Acknowledgments.** We would like to thank Kim Bruce for reminding and updating us on the Foundations of Object Oriented Languages, an area which he contributed to trigger, in part by joint work with GL, and by starting the FOOL series conferences twenty years ago.

G. Longo's work was supported in part by Marie Curie FP7-PEOPLE-2010-IRSES Grant RANPHYS. M. Montévil's work is supported by région Île-de-France, DIM ISC.

## References

- 38.1 F. Bailly, G. Longo: *Mathematics and the Natural Sciences; The Physical Singularity of Life* (Imperial College, London 2011)
- 38.2 G. Longo: Theorems as constructive visions. In: *Proof and Proving in Mathematic Education*, ed. by G. Hanna, M. de Villiers (Springer, Dordrecht 2012) pp. 51–66
- 38.3 F. Zalamea: *Synthetic Philosophy of Contemporary Mathematics* (Urbanomic Sequence, Falmouth 2012)
- 38.4 B.C. Van Fraassen: *Laws and Symmetry* (Oxford Univ. Press, New York 1989)
- 38.5 Y. Kosmann-Schwarzbach: *Les théorèmes de Noether: Invariance et lois de Conservation au XXe siècle* (Editions Ecole Polytechnique, Palaiseau 2004)
- 38.6 G. Longo, M. Montévil: *Perspectives on Organisms: Biological Time, Symmetries and Singularities*, Lecture Notes in Morphogenesis (Springer, Dordrecht 2014)
- 38.7 S.J. Gould: *Wonderful Life* (Norton, New York 1989)
- 38.8 S.Y. Pilyugin: *Shadowing in Dynamical Systems*, Lecture Notes in Mathematics, Vol. 1706 (Springer, Berlin 1999)
- 38.9 D. Ruelle, F. Takens: On the nature of turbulence, *Commun. Math. Phys.* **20**(3), 167–192 (1971)
- 38.10 J. Barwise: *Handbook of Mathematical Logic* (Elsevier, Amsterdam 1978)
- 38.11 G. Longo: Reflections on concrete incompleteness, *Philosophia Mathematica* **19**(3), 255–280 (2011)
- 38.12 H.M. Friedman: Finite functions and the necessary use of large cardinals, *Annu. Math.* **2** **148**, 803–893 (1998)
- 38.13 H. Weyl: *Das Kontinuum* (Veit, Leipzig 1918)
- 38.14 M. Rathjen: The art of ordinal analysis, *Proc. Int. Congr. Mathematicians*, Vol. 2 (2006) pp. 45–69

- 38.15 F. Varenne: La reconstruction phénoménologique par simulation: Vers une épaisseur du simulat. In: *Formes Systemes et Milieux Techniques Après Simondon*, ed. by D. Parrochia, V. Tirloni (Jacques André, Lyon 2012) pp. 107–123
- 38.16 W. Cook: Object-oriented programming versus abstract data types. In: *Foundations of Object-Oriented Languages*, Lecture Notes in Computer Science, Vol. 489, ed. by J.W. Bakker, W.P. Roeveer, G. Rozenberg (Springer, Heidelberg 1991) pp. 151–178
- 38.17 K. Bruce, del. B. Pierce: Comparing object encodings. In: *Theoretical Aspects of Computer Software*, Lecture Notes in Computer Science, Vol. 1281, ed. by M. Abadi, T. Ito (Springer, Heidelberg 1997) pp. 415–438
- 38.18 G. Longo, M. Montévil, C. Sonnenschein, A.M. Soto: In search of principles for a theory of organisms, *J. Biosci.* (2015), doi:[10.1007/s12038-015-9574-9](https://doi.org/10.1007/s12038-015-9574-9)
- 38.19 V.P. Belavkin: Quantum probabilities and paradoxes of the quantum century: *Infin. Dimensional Anal, Quantum Probab. Relat. Top.* **3**(4), 577–610 (2000)
- 38.20 A.A. Abbott, C.S. Calude, K. Svovil: Value-indefinite observables are almost everywhere, *Phys. Rev. A* **89**, 032109 (2014)
- 38.21 A. Cabello: Experimentally testable state-independent quantum contextuality, *Phys. Rev. Lett.* **101**, 210401 (2008)
- 38.22 G. Longo, C. Palamidessi, T. Paul: Some bridging results and challenges in classical, quantum and computational randomness. In: *Randomness Through Computation*, ed. by H. Zenil (World Scientific, London 2010) pp. 73–92
- 38.23 H. Poincaré: *La Science et l'hypothèse* (Ernest Flammarion, Paris 1902)
- 38.24 A. Aspect: Bell's inequality test: More ideal than ever, *Nature* **398**(6724), 189–190 (1999)
- 38.25 M. Buiatti, G. Longo: Randomness and multilevel interactions in biology, *Theory Biosci.* **132**(3), 139–158 (2013)
- 38.26 M.B. Elowitz, A.J. Levine, E.D. Siggia, P.S. Swain: Stochastic gene expression in a single cell, *Science* **297**(5584), 1183–1186 (2002)
- 38.27 R. Arjun, R. van Oudenaarden: Stochastic gene expression and its consequences, *Cell* **135**(2), 216–226 (2008)
- 38.28 A. Lesne, J.-M. Victor: Chromatin fiber functional organization: Some plausible models, *Eur. Phys. J. E. Soft Matter* **19**(3), 279–290 (2006)
- 38.29 M. Bizzarri, A. Giuliani, A. Cucina, F. D'Anselmi, A.M. Soto, C. Sonnenschein: Fractal analysis in a systems biology approach to cancer, *Semin. Cancer Biol.* **21**(3), 175–182 (2011)
- 38.30 C. Barnes, L. Speroni, K. Quinn, M. Montévil, K. Saetzler, G. Bode-Animashaun, G. McKerr, I. Georgakoudi, S. Downes, C. Sonnenschein, V. Howard, A. Soto: From single cells to tissues: Interactions between the matrix and human breast cells in real time, *PLoS ONE* **9**(4), e93325 (2014)
- 38.31 S.F. Gilbert, D. Epel: *Ecological Developmental Biology: Integrating Epigenetics, Medicine, and Evolution* (Sinauer Associates, Sunderland 2009)
- 38.32 A.M. Turing: Computing machinery and intelligence, *Mind* **59**(236), 433–460 (1950)
- 38.33 A.M. Turing: The chemical basis of morphogenesis, *Philos. Trans. R. Soc. Lond. Ser. B, Biol. Sci.* **237**(641), 37–72 (1952)
- 38.34 A. Hodges: *Turing: A Natural Philosopher* (Phoenix, London 1997)
- 38.35 J. Huxley: *Evolution. The Modern Synthesis* (Allen Unwin, London 1942)
- 38.36 E. Schrödinger: *What is Life?* (Cambridge UP, Cambridge 1944)
- 38.37 E.F. Keller: *The Century of the Gene* (Harvard Univ. Press, Cambridge 2002)
- 38.38 G. Longo: Critique of computational reason in the natural sciences. In: *Fundamental Concepts in Computer Science*, ed. by E. Gelenbe, J.-P. Kahane (Imperial College Press/World Scientific, London 2009) pp. 43–70
- 38.39 C. Calude: *Information and Randomness: An Algorithmic Perspective* (Springer, Heidelberg 2002)
- 38.40 C.E. Shannon: A mathematical theory of communication, *Bell Syst. Tech. J.* **27**, 379–423 (1948)
- 38.41 G. Longo, P.-A. Miquel, C. Sonnenschein, A.M. Soto: Is information a proper observable for biological organization?, *Prog. Biophys. Mol. Biol.* **109**(3), 108–114 (2012)
- 38.42 J.M. Smith: The idea of information in biology, *Q. Rev. Biol.* **74**, 395–400 (1999)
- 38.43 G. Kuiper, B.O. Carlsson, K.A.J. Grandien, E. Enmark, J. Häggblad, S. Nilsson, J. Gustafsson: Comparison of the ligand binding specificity and transcript tissue distribution of estrogen receptors  $\alpha$  and  $\beta$ , *Endocrinology* **138**(3), 863–870 (1997)
- 38.44 J. Monod: *Le Hasard et la Nécessité* (Seuil, Paris 1970)
- 38.45 J.J. Kupiec: A probabilistic theory of cell differentiation, embryonic mortality and DNA C-value paradox, *Specul. Sci. Techno.* **6**, 471–478 (1983)
- 38.46 J.J. Kupiec, P. Sonigo: *Ni Dieu ni gène: Pour une Autre Théorie de l'hérédité* (Editions du Seuil, Paris 2003)
- 38.47 T. Heams: Randomness in biology, *Math. Struct. Comp. Sci.* (2014), doi:[10.1017/S096012951200076X](https://doi.org/10.1017/S096012951200076X)
- 38.48 A. Novick, M. Weiner: Enzyme induction as an all-or-none phenomenon, *Proc. Natl. Acad. Sciences* **43**(7), 553–566 (1957)
- 38.49 M. Weiss, M. Elsner, F. Kartberg, T. Nilsson: Anomalous subdiffusion is a measure for cytoplasmic crowding in living cells, *Biophys. J.* **87**(5), 3518–3524 (2004)
- 38.50 R.T. Zoeller, T.R. Brown, L.L. Doan, A.C. Gore, N.E. Skakkebaek, A.M. Soto, T.J. Woodruff, F.S. Vom Saal: Endocrine-disrupting chemicals and public health protection: A statement of principles from the endocrine Society, *Endocrinology* **153**(9), 4097–4110 (2012)
- 38.51 A.M. Soto, C. Sonnenschein: Environmental causes of cancer: Endocrine disruptors as carcinogens, *Nat. Rev. Endocrinol.* **6**(7), 363–370 (2010)
- 38.52 B. Demeneix: *Losing our Minds: How Environmental Pollution Impairs Human Intelligence and Mental Health* (Oxford Univ. Press, Oxford 2014)

- 38.53 M. Buiatti: Functional dynamics of living systems and genetic engineering, *Rivista di Biologia* **97**(3), 379–408 (2003)
- 38.54 G. Longo, M. Montévil: From physics to biology by extending criticality and symmetry breakings, *Prog. Biophys. Mol. Biol.* **106**(2), 340–347 (2011)

# 39. Experimentation on Analogue Models

Susan G. Sterrett

Analogue models are actual physical setups used to model something else. They are especially useful when what we wish to investigate is difficult to observe or experiment upon due to size or distance in space or time; for example, if the thing we wish to investigate is too large, too far away, takes place on a time scale that is too long, does not yet exist or has ceased to exist. The range and variety of analogue models is too extensive to attempt a survey. In this chapter, I describe and discuss several different analogue model experiments, the results of those model experiments, and the basis for constructing them and interpreting their results. Examples of analogue models for surface waves in lakes, for earthquakes and volcanoes in geophysics, and for black holes in general relativity, are described, with a focus on examining the bases for claims that these analogues are *appropriate* analogues of what they are used to investigate. A table showing three different kinds of bases for reasoning using analogue models is provided. Finally, it is shown how the examples in this chapter counter three common misconceptions about the use of analogue models in physics.

39.1	<b>Analogue Models: Terminology and Role</b> .....	858
39.1.1	Analogue Models and Scale Models .....	858
39.1.2	The Role of Analogue Models in Philosophy of Science .....	860
39.1.3	Analogue Models in History of Science .....	861
39.2	<b>Analogue Models in Physics</b> .....	868
39.2.1	Lessons from the Nineteenth Century ..	868
39.2.2	Sound as an Analogue of Light: The Power of Experimentation on Analogues .....	868
39.2.3	Water as an Analogue of Electricity: Limitations of Generalizing from Analogues .....	869
39.2.4	Some Recent Results Using Analogue Models .....	870
39.3	<b>Comparing Fundamental Bases for Physical Analogue Models</b> .....	873
39.3.1	Three Kinds of Bases for Physical Analogue Models .....	875
39.4	<b>Conclusion</b> .....	876
	<b>References</b> .....	877

The array of analogue models used in science is extensive; an attempt to comprehend their range, in size and kind, would have to be abandoned sooner or later. The imagination, intellectual ingenuity, and technical expertise that have been expended in conceiving, constructing, and using these various disparate models, each requiring a methodology of construction and deployment appropriate to its nature and use, are dizzying.

Analogue models have been devised and used in physics for quite some time: one of the most common analogies in physics, the analogy between sound and light, was invoked in the mid-nineteenth century to build a sonic analogue of the Doppler effect for light, which was then used to investigate and establish results for both sound and light [39.1–3]. The analogy was later invoked in the twentieth century to explain Vavilov–Cerenkov radiation, also known as Cerenkov

radiation [39.3,4]. Cerenkov radiation is the electromagnetic radiation emitted when an electron travels in a medium faster than the speed that light travels in that medium. In his Nobel lecture, *Cerenkov* explained [39.4]:

“This radiation has an analogy in acoustics in the form of the so-called shock waves produced by a projectile or an aeroplane travelling at an ultrasonic velocity (Mach waves). A surface analogy is the generally known bow wave.”

More recently, in the twenty-first century, physicists have developed, loosely speaking, analogue space-time and analogue gravity [39.5–7]. Although the initial proposals for analogue models for space-time were based on an analogy between light and sound, once the idea of exploring analogue models of gravity began attract-

ing more interest, a variety of analogue models based on different analogies were proposed [39.8]. Thus the idea of an analogue based on the analogy between light and sound was expanded to many different kinds of analogues. *Faccio* points out a commonality that can be seen across all of them, though: All of them can be “re-connected to some form of flowing medium” [39.8, p. v]. *Visser* elaborates further [39.9]:

“In all the analogue spacetimes, the key idea is to take some sort of *excitation* travelling on some sort of *background*, and analyze its propagation in terms of the tools and methods of differential geometry.”

Arising in part from the interest generated by the work on these analogue models, physicists (*Carusotto* and *Rousseaux*) have formulated the notion of a “generalized Cerenkov emission” process [39.10].

Another analogy commonly drawn in physics is the analogy between electrical circuits and mechanical systems. The analogies date from the nineteenth century; it appears they were first invoked to make mechanical models of electrical circuits, the models being seen as a way of using knowledge about mechanical systems to provide a better understanding of electrical behavior and concepts [39.11, 12]. However, the use of electrical circuits specifically designed to model mechanical systems later became standard:

1. Measurements of the flow of current in an appropriately constructed circuit were used to accurately compute quantities used in the mechanical analysis of the corresponding structure.
2. Varying elements in the circuit corresponded to varying parameters in the mechanical system, so the effect of differences in a design or a system’s initial conditions could be explored.

More generally, electronic circuits were used as analogues of anything that could be formalized as a solution of certain classes of differential equations, and ever more sophisticated machines were developed to deal with ever larger classes of differential equations and

problems ([39.13, 14] and [39.15, p. 222ff]). Other examples of analogues used for computation are mechanical analogues such as the geared devices built in the seventeenth century [39.16], the soap bubble analogue computers invoking minimization principles that were used to efficiently solve difficult mathematical problems in the twentieth century [39.17] and biological analogue computers of the twenty-first century such as amoeba-based computing (ABC) analogue models [39.18].

Other analogue models used experimentally to carry out serious research could be named in astrophysics, cosmology, statistics, economics, geophysics, electromagnetism, fluid mechanics, fluid dynamics, solid mechanics, solid dynamics, structural engineering, coastal engineering, the behavior of volcanoes, and many other fields.

To be clear, these are actual, physical objects or setups, usually human made, designed to be used as analogue models. The modeling process for employing a physical object or setup as an analogue model includes the identification of a mapping that allows one to correlate something observed or measured in the analog model with something else (its correlative, such as a corresponding quantity) in the thing modeled. The modeling process also includes a justification of the mapping of some sort, usually invoking a principle or equation to establish the mapping. What is modeled is usually another physical object, process, or phenomenon. The model’s limitations in representing certain phenomena in the thing modeled, and any corrections that need to be made due to such limitations, are usually discussed when the analogue model is used for a particular problem. Such qualifications are not meant to undermine or recommend against using the model; they are part of the model and modeling process.

While numerical models implemented on electronic digital computers may have supplanted some of these specific uses, analogue models continue to be used in most of these fields today, and new analogue models and methods of using them continue to be invented and further developed.

## 39.1 Analogue Models: Terminology and Role

### 39.1.1 Analogue Models and Scale Models

It will be helpful to clarify the terminology of *analogue model* and *scale model* as used in this chapter.

#### Analogue Models

The word *analogue* has two connotations relevant in discussions on models: (i) analogous or parallel to; and (ii) continuous, as contrasted with digital. It sometimes

happens that a model is analogue according to both meanings. In this chapter, we will use *analogue* to mean analogous or parallel to. Thus *experimentation on an analogue model* is used to mean *experimentation on something analogous to the thing modeled*. It is important to be clear about what is meant in saying that a model is *analogous to* the thing modeled.

To say one thing is analogous to another is always to say it is so *with respect to* a particular analogy, whether



or not this is made explicit; there may be many different possible analogies one could draw between two physical things or processes. Thus, just as it does not make sense to ask whether or not one thing is analogous to another without specifying the analogy between them one means to be inquiring about, so it does not make sense to say that one thing is an analogue model of another thing without specifying the analogous relation that is the basis for the correspondences being drawn between the model and what is modeled. Thus, it is implicit in the notion of an analogue model that there is some definite analogous relationship that one means to be referring to, between the model and the thing modeled.

### Scale Models

A *scale model* (in the sense that engineers and scientific researchers use the term *scale model*) can be considered a special case of an analogue model. (Or, conversely, an analogue model can be considered a generalization of the notion of a scale model.) One way of understanding the relationship between analogue and scale models is by considering how the methodology of physically similar systems applies to each of them.

Using the method of physically similar systems, similarity of two systems is established by showing that each member of a certain (nonunique) set of dimensionless parameters that characterizes the behavior of the two systems has the same value in the model as in the thing modeled; in practice exact similarity is often not achievable (Chap. 18). Instead, certain of the dimensionless parameters are prioritized, or one aims for the dimensionless parameters to be only approximately equal. That is, it is said that a system  $S$  and a system  $S'$  are similar with respect to a behavior  $B$  (e.g., kinematically similar, dynamically similar, similar with respect to buckling behavior, similar with respect to electrical flows, and so on) when a set of dimensionless parameters (ratios) that characterizes that behavior has the same values in  $S$  as in  $S'$ . Despite the fact that, in practice, it is often possible to meet this criterion only partially or approximately, the concept of physically similar systems whose similarity is established by dimensional analysis (via establishing equality of the relevant dimensionless parameters), which was originally developed to provide a basis for making use of scale model experiments, still forms the foundation for the use of analogue models and has the virtue that it does not, as most other methods do, require complete knowledge of the equations and conditions that determine the behavior  $B$  of interest [39.19].

Now, to see the point that a scale model is a special case of an analogue model: each dimensionless parameter is a ratio, so it is only the value of a quantity *in relation to* other quantities that determines the value

of the dimensionless parameters used to establish similarity between two systems  $S$  and  $S'$ . To use a simple example, it is Mach number (the ratio of the velocity of a flow or a moving object to the velocity of sound in the medium at the fluid conditions that obtain at a certain time), and not the value of a quantity such as a velocity itself that indicates whether flow is supersonic or subsonic. The Reynolds number (density  $\times$  velocity  $\times$  length, divided by viscosity) is generally indicative of the flow regime (laminar, transitional, or fully developed (turbulent) flow). People often use the term *scale model* when thinking about scaling *linear* dimensions in particular, and thus are thinking in terms of *ratios of lengths* rather than some other (dimensionless) ratio; then, the point about sameness of ratios becomes a point about sameness of ratios of lengths, and hence about the significance of geometrical similarity to the occurrence of some phenomenon. That is, for two systems  $S$  and  $S'$ , *if all we are interested in is a feature or behavior that depends solely on ratios of linear dimensions, then geometrical similarity between model and thing modeled suffices for an object to serve as an analogue model of the thing modeled*. This is a special case of a physically similar system in which the relevant dimensionless parameter is a ratio of lengths [39.20]. A *scale model* used in architectural layout is a paradigm example of this kind of similarity, and can be considered a special case, even a degenerate case, of a physically similar system.

Unfortunately, the architectural model as a paradigm of a scale model has become so closely associated with the very idea of a scale model that it can interfere with understanding how broadly the concept of *scale model* applies. The concept of *scale model* is not limited to scaling of a linear dimension alone; other quantities can be scaled too [39.20]. Despite this, the term *scale model* is often used in the more restricted sense of scaling the linear dimension of the situation.

Even when restricting the meaning of *scale model* to the scaling of linear dimensions, it is certainly not necessary that a scale model be smaller in size than what it models, nor even that a scale model have the same geometrical proportions as what it models. Scale models of very small things have been built which are larger than what they model, so as to permit ease in manipulation and observation (e.g., a scale model of a cell); and full-size scale models used for prototype testing are common as well (e.g., to test airflow patterns around, or heat convection associated with, a certain shape). Even in full-size scale model testing, the method of physical similarity is applicable in designing the experimental conditions to be applied, and in interpreting the results of the experiment [39.20, 21]. The use of distorted scale models calls for some explanation.

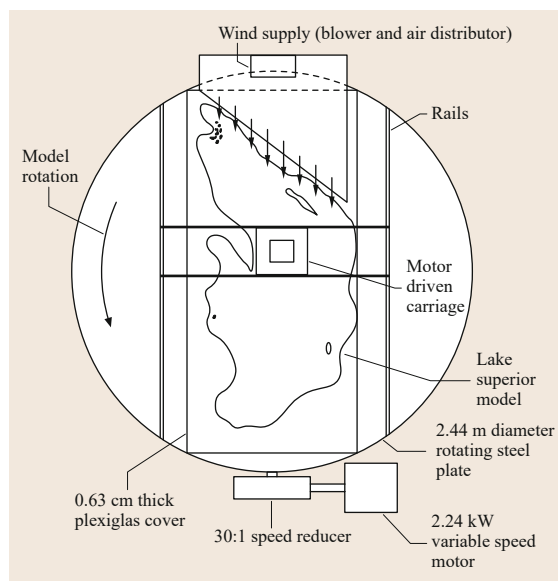
### Distorted Scale Models

Distorted scale models, which are scale models that fail to be geometrically similar to the situation modeled (by design, and in a certain very specific way) [39.21], have been used in scale modeling for over a century; an example may illustrate the nature of, and reason for using, such models.

One example of a distorted scale model is a physical model of Lake Superior [39.22] that was “built to satisfy the Froude number and Rossby number requirements of dynamic similitude.” (The Froude number is indicative of the ratio of inertial forces to the gravity forces of flow, and is important in studies where surface waves are important; the Rossby number is indicative of the ratio of inertial forces to the Coriolis force.) The model was used to generate quantitative results: the coriolis force (in the actual Lake Superior) was modeled by rotating the laboratory model about a vertical axis, and the lake bottom in the model was *warped* so as to provide the correct scaled depth while the model was rotating [39.22, p. 25]. The wind flow over the lake was modeled in the laboratory model using a blower with an air distributor. The researchers’ experimentation on this analogue model of Lake Superior involved blowing *wind* over it in different directions; they recorded the results in the analogue model by “photographing aluminum particles spread on the water surface” [39.22].

The plan view and side view of the model are shown in Figs. 39.1 and 39.2, respectively. The model is a scale model, yet it is not geometrically similar to Lake Superior; the researchers explain: “Because of the large ratio of horizontal to vertical distances in Lake Superior, the model was [intentionally made so as to be] vertically distorted” [39.22]. The fact that the vertical linear dimensions of the model are scaled differently than the horizontal linear dimensions are scaled is, of course, taken into account when the corresponding quantities to be associated with the actual Lake Superior are calculated from the values of the quantities observed in the model. The ratio of time in the laboratory model to time in the actual Lake Superior is 1/9480, so that “1 day in the prototype is equivalent to 9.1 s in the laboratory model,” for instance.

Such distortion of the vertical dimension in modeling ship performance at sea or fluid behavior in canals is common, but they are not the only use of distorted models. One particularly complex hydraulic model that used distortion was constructed and used to study the impact that large woody debris in a stream had in reconfiguring the creek bed itself [39.23]. More recently, the use of distorted laboratory models in investigating the structural response of (flexural) plates has been evaluated analytically and judged to provide a reliable means of laboratory investigation [39.24].

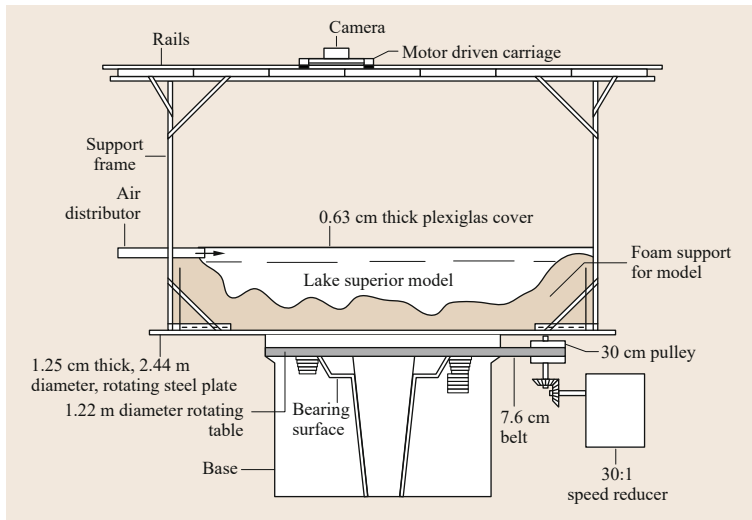


**Fig. 39.1** Plan view of Lake Superior distorted model; the ratio of horizontal distance in the laboratory model to horizontal distance in the actual Lake Superior is 1/300 000 (after [39.22])

### 39.1.2 The Role of Analogue Models in Philosophy of Science

The wide variety of analogue models currently used in serious scientific research mentioned in the beginning of this chapter is not, however, reflected in the discussions of analogue models one finds in the history and philosophy of science literature. When analogue models are mentioned in philosophy of science, they are usually seen as curiosities suitable for illustrative, entertainment or pedagogical purposes, rather than as a serious research methodology. When, on occasion, their role in serious scientific research is recognized, it is usually for a role played in the science of a past era, and often for a qualitative or heuristic purpose at that.

An indication of the extent of confusion and ignorance about analogue models and scale models that exists in mainstream philosophy of science is found in the account in the entry entitled *Models in Science* found in one of the most prominent encyclopedias of philosophy, coauthored by two leading philosophers of science, *Roman Frigg* and *Stephan Hartmann*. Scale models are not included under analogical models in that chapter, and it is claimed that “[t]ypical examples [of scale models] are wooden cars or model bridges” [39.25]. There is no recognition in the article of the notion of physically similar systems or any other methodology of scale models, in spite of the fact that methods of dimensional analysis applied to scale models are the topic of count-



**Fig. 39.2** Side view of Lake Superior distorted model, “Because of the large ratio of horizontal to vertical distances in Lake Superior, the model was vertically distorted.” The ratio of vertical distance in the laboratory model to vertical distance in the actual Lake Superior is 1/1000 (after [39.22])

less books and papers in a wide variety of journals in physics and other scientific disciplines. Instead, the topic of the methodology of scale models is dismissed with the misguided reasoning that [39.25]

“Scale models seem to be a special case of a broader category of representations that Peirce dubbed icons: representations that stand for something else because they closely resemble it.”

and that “no theory of iconicity for models has been formulated yet.” Likewise, the philosopher *Ronald Giere*, widely recognized in philosophy of science for championing the recognition of the role of models in science, uses examples such as “Watson’s original tin and cardboard model of DNA” and “Rutherford’s solar system model of the atom” as examples of scale models and analogue models, respectively ([39.26] and [39.27, p. 747]). Two well-known discussions in philosophical venues that specifically address experimentation on laboratory water tank analogue models [39.28, 29] do not discuss the foundations of the approach used by researchers who actually employ those models, that is, dimensional analysis and the method of physically similar systems.

Hence there is little help to have from the mainstream philosophical literature as far as understanding bases for the scientific reasoning involved in actual experimentation on analogue models by researchers. To be fair, there are a few works that do make some philosophical points about the methodology, the assumptions and the limitations of various bases for experimental methodologies employing analogue models, but they seem unconnected, in that the mainstream discussions in philosophy of science that ought to take note of them seldom do so (Chap. 18 and [39.30–37]). The emphasis

in this chapter will be on the methodologies employed by the researchers who have effectively used laboratory experimentation on analogue models: our interest here is especially in the basis for the inferences drawn using these analogue models.

### 39.1.3 Analogue Models in History of Science

#### Geophysics

*Analogue Models in Geophysics: An Historical Narrative.* An example typical of narratives that view analogue models in terms of their role in the past is historian *Naomi Oreskes’ From Scaling to Simulation: Changing Meanings and Ambitions of Models in Geology* [39.38]. According to her narrative, physical scale models were used in the nineteenth and early twentieth century, but in the latter part of the twentieth century [39.38, p. 93]:

“The word model took on a different meaning: a computer simulation. For earth scientists, this is the dominant meaning it holds today.”

The main storyline in her narrative has to do with the cause of shifts in epistemic goals of geologists in the late twentieth century. Our interest in this chapter is simply in the methodology of the analogue models employed. As Oreskes indicates, the method of physically similar systems was applied to the question of how to scale experimental analogue models for the behaviors of interest in studying geologic structure in the 1930s by *Hubbert*, in *Theory of Scale Models as Applied to the Study of Geologic Structures* [39.39]. It is a means of obtaining *quantitative* results about something by taking measurements on an analogue physical model of it.

*Hubbert* cites Galileo's *Two New Sciences*, then works by Newton, Stokes, Helmholtz, and Reynolds [39.39, pp. 1516–1517]. *Hubbert* also describes work done by Koenigsberger and Morath applying physical similarity to geology [39.39, p. 1518]. Prior to that, many people had built small tabletop models to investigate geological processes, but there was good reason to be skeptical about the validity of these small models; there was a need, *Hubbert* said, for [39.39, p. 1463]:

“an objective criterion to enable one to determine what the correct properties of a model should be for the best similarity, when the properties of the original are known, or whether it is even possible to build a correct model from available materials.”

Koenigsberger and Morath did “the earliest explicit application of the method of dimensional analysis to tectonic structures” that was in 1913 [39.39, p. 1518]. So, *Hubbert* stresses, what he is advocating is not new. *Hubbert* did not use the more mathematically concise and elegant method of physically similar systems that Buckingham presented in 1914 and which is described elsewhere in this volume (Chap. 18), but he did use the theory of dimensional analysis, systematically developing and carefully elaborating the dimensionless ratios associated with providing geometrical, kinematical, and geometrical similarity between the model and what is modeled by it. Finding the requirement of dynamic similarity too strict to be practical in many cases, he then goes on to discuss the kinds of approximations that *are* appropriate in special cases. He considers special cases in which inertial forces are very small, ones in which gravitative forces are negligible, and ones in which resistive forces are negligible, explaining which criterion can be violated without affecting the results too much in each case.

As have so many others who have seen themselves as advocates of the method, that is, as urging the use of the method of physical similarity to handle previously unsolved problems in their profession, *Hubbert* sounds almost evangelical in his advocacy [39.39, p. 1519]:

“[...] the evidence is in that in remote parts of the world the geological professional is already awakening to the importance of so powerful a tool as that afforded by the method of dimensional analysis and correctly made scale models, for the solving of problems that have not yielded satisfactorily to methods of attack previously employed.”

The reasons geophysics needed to use scale models were very much the same as the reasons scale models were being used in other areas such as mechanics and hydrodynamics: the phenomena “are so complicated as

a whole as to render complete mathematical analysis difficult or impossible” [39.39, p. 1460]. Then:

“where mathematical analysis is inadequate, and where for one reason or another direct experimentation is precluded, the best remaining alternative is to construct and study a scale model.”

Writing in 1937, he cites the fields of aerodynamic, hydraulic, mechanical, and electrical engineering for their success, then notes that [39.39, p. 1461]:

“The geological problems of mountain making and of diastrophism in general are peculiarly of the type that do not lend themselves readily to analysis, and the size of the elements involved place them beyond the range of direct experimentation. In this case also there remains the alternative of studying such phenomena by means of experiments performed upon properly built small scale models.”

In the preface to a later paper *The Strength of the Earth*, in which he resolved an apparent paradox in geophysics by appealing to physical similarity, *Hubbert* explained the value of the approach [39.40]:

“By means of the principles of physical similarity, it is possible to translate geological phenomena whose length and time scales are outside the domain of our direct sensory perception into physically similar systems within that domain.”

He shows the value of developing the characteristics of the materials that would be needed to make a physically similar laboratory model, and shows that even the knowledge of what the physically similar laboratory model would be like is informative. For the purpose of resolving the apparent paradox, understanding scaling relations for the case of the earth is all that is needed [39.40, p. 1653]:

“We learn that the resemblance of the behavior of rocks on a length scale of thousands of miles and a time scale of millions of years is not to that of rocks with which we are familiar but rather to that of the viscous liquids and weaker plastics of our personal experience.”

However, the fact that such qualitative lessons can be drawn does not obviate the need for building the models to learn about tectonics in many cases, and even in the same paper in which *Hubbert* makes the general observation just quoted, he reviews various experimental scale laboratory models that had been built using the method of physical similarity [39.40, p. 1653].

Translating geological phenomena occurring *outside the domain of our direct sensory perception* into that domain is of course extremely significant in the

field of geophysics, due to the large sizes and long time scales involved. Given that philosophy has so seldom included this method of experimentation among serious scientific reasoning, the attention Oreskes gives to Hubbert's work on applying physical similarity to geophysics, though not her main point, is a valuable rarity in the literature of history and philosophy of science.

**Analogue Models in Geophysics: The Case of Volcanology.** One area of geophysics where physical similarity has been employed is volcanology. There are different kinds of volcanoes; volcanoes can differ in configuration and in the mechanisms by which they were formed, for instance. Further, the configurations are seldom static: a given volcano's configuration can change during the process that is of research interest (e.g., eruption, spreading). Some processes take place over time periods that are very short, involving very high velocities, and others take place over long time periods, for example, slow changes between eruptions. Concurrent processes are often studied separately in order to understand the mechanisms involved. Sometimes the study focuses on the peculiarities of a specific volcano, and sometimes the subject of the investigation is about general processes and not specific to any volcano. Thus no single example of an analogue model from volcanology is likely to be representative. An example of the use of physical similarity that illustrates how its application can involve very different analogue models of the same volcano is the use of various scaled experiments of different mechanisms involved in the ongoing evolution of Mt Etna in Italy, especially volcano spreading and dike propagation.

To investigate the process of volcanic spreading, cones of sand on layers of sand and silicone were used [39.41]. Volcanic spreading is a long-term process and it involves more than one factor, but the effect of the weight of the volcano on the substratum is one of them. Identifying what the model does not do is part of explaining the model, and the researchers state up front that

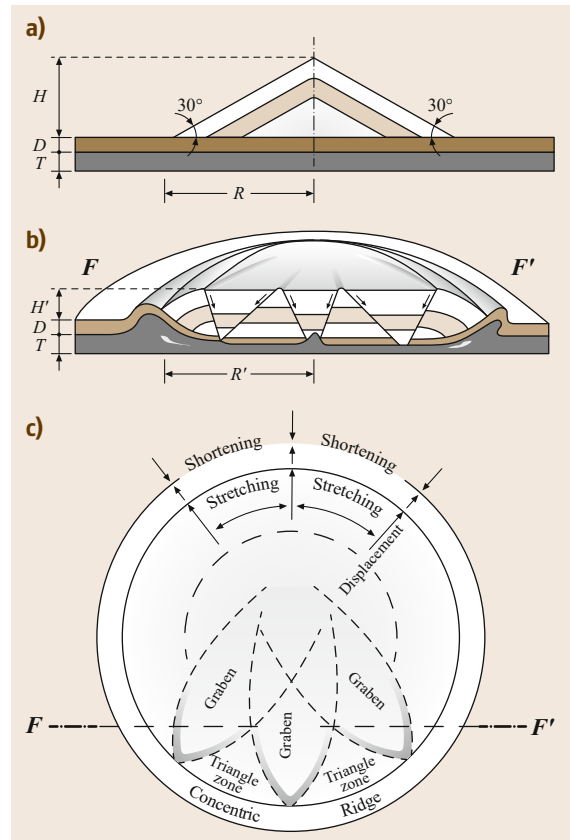
“Our experiments do not model the effect of the intrusive complexes; they cannot be used as exact scale analogs of volcanoes where the intrusive complexes give the dominant contribution to deformation [...]”

citing other experiments that do so. They add: “[...] our experiments do not model the effect of subsidence due to crustal flexure under the load of volcanic edifices” and note that effect *is* in fact important for some specific volcanoes and kinds of volcanoes. Their model considers the volcano already cooled, so they are not modeling thermal effects in the experiment, either. Nor,

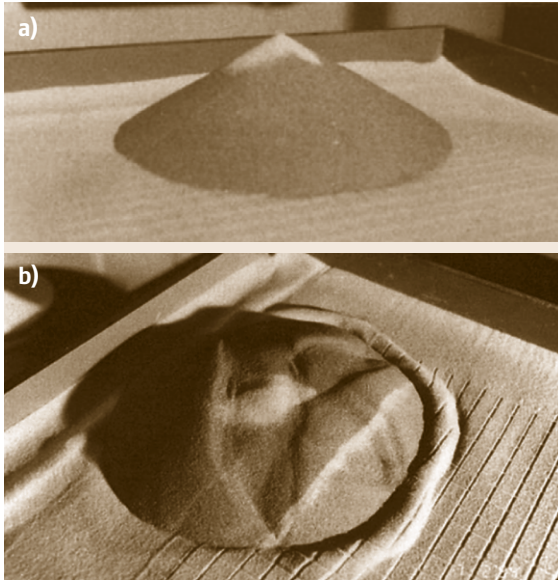
they add, do they take into account “any contribution of magma forces to the destabilization process” [39.41].

Their explanation of the value of an analogue model of volcano spreading that neglects so many mechanisms is that it can show a relation between the mechanism they wish to model and an effect that can be observed in both the model and the thing modeled. Figures 39.3 and 39.4 show the schematic and photographic views of the experiment. Drawing on previous results by others, they use dry sand in the small laboratory model as an analog of brittle rocks in the actual Mt Etna volcano [39.41, p. 13808]. The spreading in the model (analogue volcano) experiment takes less than a day; a record of the experiment is made using overhead time lapse photography of the surface of the spreading sand cone [39.41, p. 13807].

The model is constructed by taking the approach of preserving the dimensionless ratios important to the behavior of spreading, as best they can, and prioritizing



**Fig. 39.3a–c** Schematic of experiment of spreading volcano, from initial state of laboratory model to 10 h from initial state. (a) Initial stage (cross section); (b) state after deformation (cross section); (c) state after deformation (top view) (after [39.41])

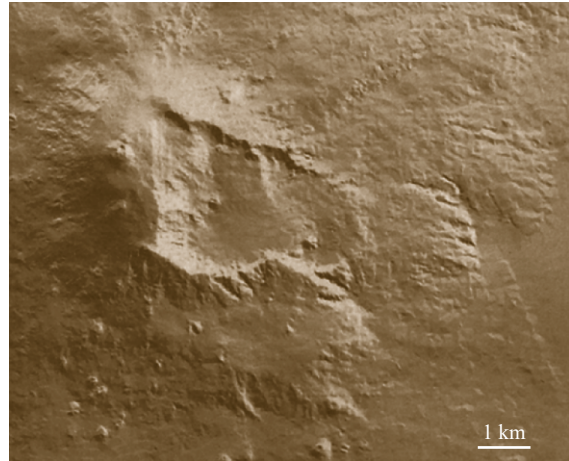


**Fig. 39.4a,b** Analogue volcano made of sand, from initial state (a) to end of spreading experiment 10 h later (b). From Merle and Borgia [39.41], who find the pattern in the end state above remarkably similar to pattern in actual volcano shown in Fig. 39.5

some ratios over others. The choice of dimensionless parameters and ratios here is not done from scratch by analysis of the specific problem they are investigating, but draws on Hubbert's analysis of these kinds of problems in geophysics. One interesting aspect of this experiment is that some of the ratios change significantly during the experiment itself. Table 39.1 shows the dimensionless variables used. (A note about the notation in the chart in Table 39.1: Although the Greek letter  $\Pi$  is used to denote dimensionless variables there, not all dimensionless variables indicated in Table 39.1 have the same role as a dimensionless  $\Pi$ -group or a dimensionless parameter as that term is commonly used in dimensional analysis (Chap. 18 and [39.33, 34]).

**Table 39.1** Average dimensionless numbers for experiments in actual volcano (*Field*) and in laboratory analogue volcano (*Experiment*) (after [39.41]). Because the configuration changes dramatically during the volcano spreading experiment, some of these ratios change in value during the experiment. Each experiment is characterized in terms of the values these dimensionless variables take on for that experiment

Dimensionless variable	Definition	Value	
		Field	Experiment
$\Pi_1$	Height/radius of volcano	0.15–0.2	0.58 → 0.18
$\Pi_2$	Brittle substratum/height of volcano	0–1.5	0.1 → 0.22
$\Pi_3$	Brittle/weak substratum	0–15	1 → 2
$\Pi_4$	Volcano/substratum density	1–1.4	1
$\Pi_5$	Gravitational/viscous forces	790	1200
$\Pi_6$	Frictional/viscous forces	82–327	160
$\Pi_7$	Inertial/viscous forces	$10^{-20}$	$10^{-12}$



**Fig. 39.5** Summit of Mt. Etna volcano as a digital elevation model (after [39.41], courtesy of Macedonio and Pareschi, University of Pisa)

The analogue spreading volcano experiment is run using different substratum layers (brittle layer only, brittle layer and ductile layer, ductile layer only) and a buffering solid boundary that buttresses the cone. The experiments are characterized in terms of the values of the dimensionless variables [39.41, p. 13809]. The authors remark on the ability of such simple experiments to model features of the natural volcano that had not been modeled before.

More recently, a completely different mechanism suspected to be occurring in the same volcano (although this *same* volcano, Mt Etna, had erupted in the meantime, including flank (side) eruptions) was modeled by an experiment that modeled a very different kind of mechanism: magma emplacement in the volcano [39.42]. In this model, a viscous material was injected into a cone of granular material. An aerial view of Mt Etna is shown in Fig. 39.6. The experimental apparatus for the laboratory analogue model used to study the consequences of magma emplacement is shown in Fig. 39.7.

The experimental apparatus and methods used produce measurements of high precision. The size of the model, the materials used in the model, and the scaling of quantities of interest are developed using dimensional analysis, again citing the work of *Hubbert* [39.42, Section 3.2].

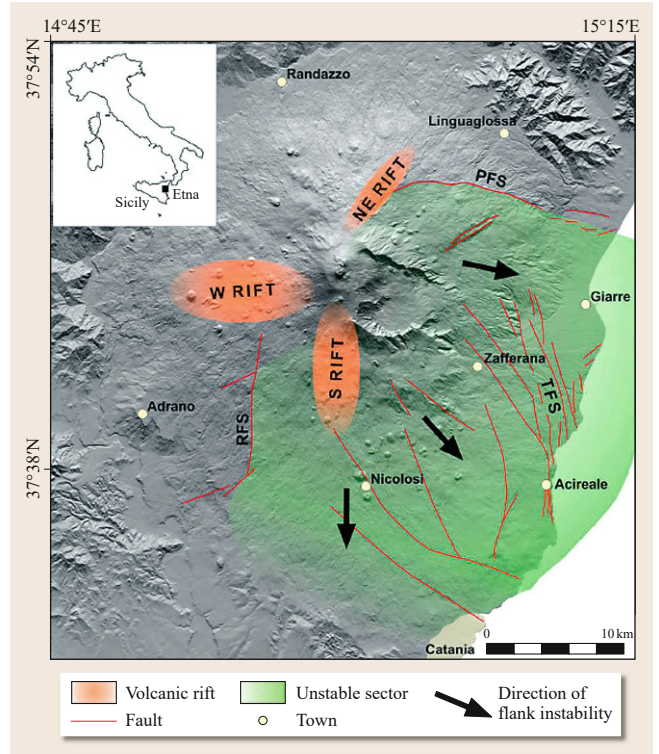
The research into the mechanisms at work in the Mt Etna volcano using analogue volcanoes is of more than theoretical interest, as thousands of deaths have already resulted from Mt Etna's eruptions. Predicting the future is certainly of interest in the employment of this analogue model; the safety of many could depend upon an understanding of how this specific volcano behaves. The authors note that the use of an analogue model to precisely model the displacements of a specific volcano *quantitatively*, as was done in their study, is a new use of analogue models in volcanology, and that this work could lead to an *advanced generation of analog models* that could be compared with those of the actual volcano, and could aid simulation studies [39.42, pp. 18–19].

*Analogue Models in Geophysics: The Lessons of History.* Oreskes' narrative, though a welcome rarity in mentioning the historical role of physical similarity, contains a statement about the role of analogue models that could mislead readers into thinking that (or reinforce existing prejudices that) physical analogue models are dispensable in geophysics. As this is a rather common misconception in philosophy of science today, it is useful to confront it here. Oreskes writes [39.38, p. 113]:

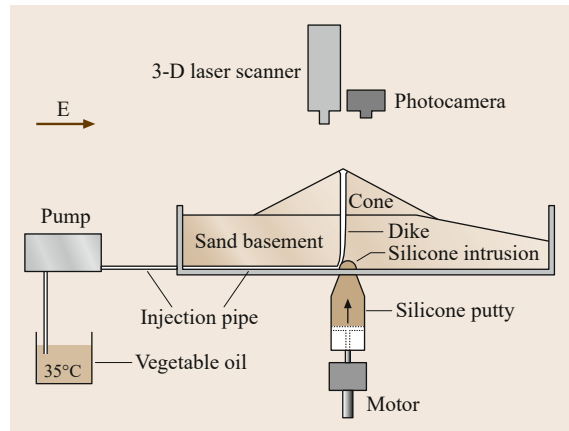
“If one could calculate the required properties of materials in a scale model, then there was actually no need to build the model itself. One could simply calculate the property of interest.”

It is not clear what is the basis of such a claim could be; building and running experiments with analogue models have been shown to be important in many cases in geophysics from *Hubbert's* day all the way up to the present. *Oreskes* continues: “In principle, a computer simulation can be used in precisely the same manner as a mimetic physical model to demonstrate circumstances capable of producing known effects” [39.38, p. 113]. This statement gives pride of place to computer simulation in geophysics, which is not deserved. The “in principle” qualification, which is actually an extremely significant qualification, needs to be given sufficient weight.

First, it needs to be emphasized that experimentation on analogue models has *not* been supplanted by computer simulations; it is surprising how often the misconception that they have been is voiced. Granting that sophisticated computing methods implemented on



**Fig. 39.6** Mount Etna volcano, aerial view. The green area is unstable. The fault systems are indicated by RFS, PFS, and TFS (after [39.42])



**Fig. 39.7** Experimental apparatus used to model magma emplacement by the injection of silicone putty or vegetable oil into an analog volcano (after [39.42])

digital computers are now used for many of the tasks for which analogue models were at one time used, this still does not mean that analogue model experiments are now dispensable. In many cases – perhaps even in most cases – a great deal more knowledge would be needed

in order to construct a computer simulation than would be needed to construct and use an analogue model experimentally in order to yield new knowledge. This is clearly the case in geophysics. Computer simulation is often preferred for reasons of cost and adaptability, but it cannot be considered a satisfactory substitute for experimental analogue models in general. Analogue models can be extremely expensive, due to the laboratory personnel and facilities involved in constructing, instrumenting, and carrying out experiments on models with a high degree of precision. Yet, even costly analogue models are still used to this day, as they often reveal phenomena that a computer simulation built using current knowledge does not. This has been as true in other areas such as aeronautics as it has been in geophysics and physical earth sciences: the demise of the wind tunnel has been predicted quite a few times over the last century, but, in spite of such predictions, wind tunnels are still considered indispensable today. So, too, are the analogue models – some quite costly – used in geophysics today.

Second, it needs to be emphasized that most computer simulations *rely upon* information gained by observation and experimentation, especially experimentation on analogue models. The current practice, in geophysics as in so many other fields, is to use both kinds of models in conjunction; over the long term, each methodology can help inform and improve the other ([39.42] and [39.43, p. 1317]). But there is an asymmetry: while analogue model experiments can be and in the past were performed without benefit of computer simulations, most computer simulations relied heavily on knowledge gained from analogue model experiments – whether today’s users of such sophisticated computer packages realize it or not. One practical benefit of computer simulations that accounts for their popularity and widespread use is the ease with which a model can be modified. The advantage of cost and adaptability of computer simulations led to their adoption in cases where the mechanisms were well understood, and this was followed by over-reaching claims about what computer simulations were capable of replacing. That these claims were over-reaching is seen in retrospect; we can now see that, in geophysics as in many other fields, experimentation on analogue models has not only *not* been replaced, but still *holds an irreplaceable role* in investigation.

*Role of Analogue Models vis a vis Numerical Computer Simulation.* These points about the role of analogue models – that is, their use in conjunction with, rather than their displacement by, numerical simulation computer methods in the post-computer era – is readily seen by looking at actual examples of recent re-

search using analogue models. Many examples would serve for this; here we shall look at the details of an example from a recently published (2014) investigation carried out by researchers at Caltech and the University of California and published in a major venue (*Journal of Geophysical Research: Solid Earth*): the experimental investigation of strong ground motion due to thrust fault earthquakes [39.43].

In this investigation, the topic is not the geometry of the changes caused by the earthquakes, but the rupture velocity. Previously, there had been a question of whether the rupture proceeds below or above the velocity at which the seismic shear wave propagates; though observations that supershear ruptures had occurred in natural earthquakes began piling up, their existence went against well-established belief. According to *Gabuchian et al.* [39.43], the role of analogue laboratory models in the discovery and acceptance of the occurrence of a high-speed “rupture velocity” that *does* exceed this critical velocity was profound: they write that [39.43, p. 1317]:

“it was the experimental discovery of supershear ruptures occurring repeatedly and reproducibly under highly instrumented and controlled laboratory conditions [...] that stimulated the recent flurry of theoretical activities on the subject.”

They report that the theoretical activities were themselves changed in important ways as a result of the laboratory results; [39.43]

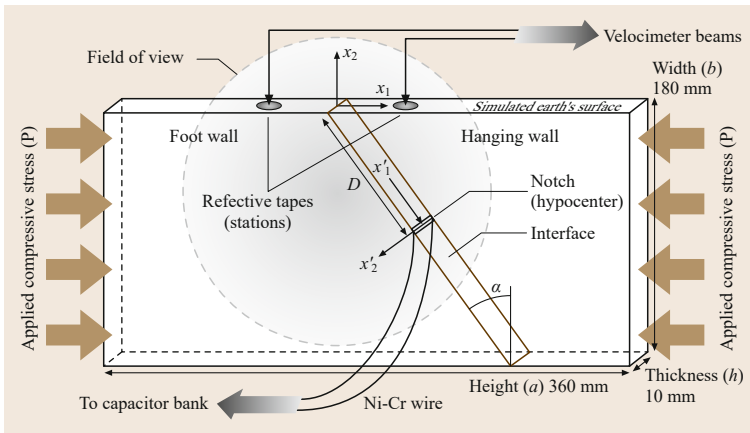
“it motivated seismologists to remove [certain speed restrictions], to revisit a number of historic earthquake events, and to reexamine irregular field observations in search for such a phenomenon.”

*Gabuchian et al.* argue that analogue experiments are crucial in developing and validating numerical models [39.43], and that, even more importantly, they are [39.43, p. 1317]:

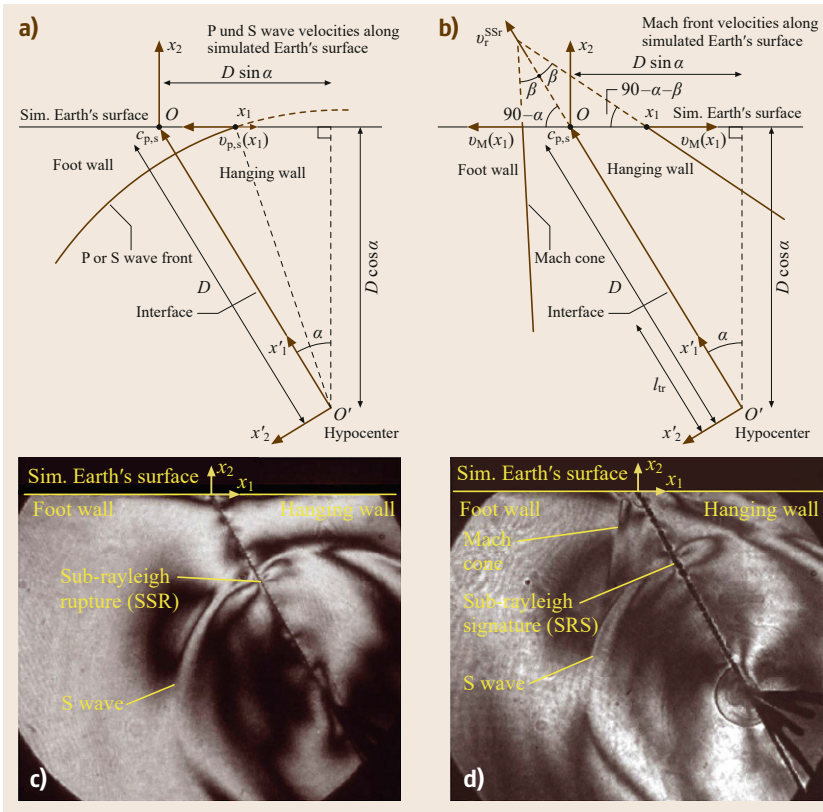
“the only way to provide fresh observations of previously unknown phenomenon (discovery) that can then be investigated in numerical models and in seismological data. Indeed, many of the effects of dip-slip faults [...] were predicted by the analog foam rubber models of *Brune* [39.44]. Thus, laboratory experiments, numerical models, and seismic observations can be used together and iteratively to more fully investigate the physics of faulting.”

*Gabuchian et al.*’s earthquake investigation [39.43] uses an analog material called homalite, which is a high-density photoelastic polymer material. The material is prepared and the setup instrumented in such a way that it provides a high-precision model of lab-





**Fig. 39.8** Experimental specimen made of homalite (a high-density photoelastic polymer material). A dynamic rupture is triggered at the interface; the upper surface on the right side of the interface serves as a *simulated earth's surface* (after [39.43])



**Fig. 39.9** Photoelastic images of experimental results. Panels (a) and (b) are schematic drawings for sub-Rayleigh (a) and supershear ruptures (b). Panels (c) and (d) are photoelastic images of sub-Rayleigh (c) and supershear events (d). A dynamic rupture is induced on the diagonal surface so that the propagation of the rupture in the homalite specimen can be studied. The Mach cone is clearly visible in the *right hand image* (after [39.43])

oratory earthquakes, with respect to the quantities measured. The results are also visually accessible. Many analogue models in geophysics use materials more like the material modeled, as in the volcanic spreading example. The use of a photoelastic polymer such as homalite in which shocks are caused thermally or electrically – rather than being mechanically induced – reflects both the recent development of new laboratory techniques and the recognition that fundamentally dif-

ferent processes may need to be studied now in order to understand the consequences of the phenomena of *dynamic fracture*. Experimental reproducibility is addressed in their studies, too [39.43, Section 4.3]. The interest here is definitely not limited to explaining the past, but in using analogue models to understand the future as well as the past.

A sketch of the experimental specimen used in Gabuchian et al. experiments is shown in Fig. 39.8.

As indicated in Fig. 39.8, one of the surfaces of the specimen serves as a simulated *earth's surface*. The design and sizing of the specimen involved some mathematical analysis; avoiding reflected waves and buckling were two requirements that had to be met. The photoelastic images revealing the difference between the experimental results in the subcritical and supercritical cases are very striking and shown in Fig. 39.9. We shall not go into the details of the results here, except to point out the Mach cone in the

supercritical case (on the right-hand side of Fig. 39.9), and that one can visually compare the photoelastic image of the supercritical case to that of the photoelastic image of the subcritical case (left side of Fig. 39.9).

The significance of this research for our purposes is that, as described above, this research using analogue volcanoes shows a significant feature, or phenomenon, that was not uncoverable using existing methods of analyzing earthquakes [39.43].

## 39.2 Analogue Models in Physics

Analogue models are “a constant presence in the world of physics and an invaluable instrument in the progress of our knowledge of the world that surrounds us,” in the words of the editors of a recent (2013) collection on the use of analogue models in contemporary theoretical physics [39.8], though, as they point out, it “would be impossible to give a comprehensive list of these analogue models” [39.8, p. v]. One especially interesting example in theoretical physics is the use of analogue models of space-time briefly mentioned in the opening of this chapter. A variety of such analogue models have been proposed, including analogues that employ surface waves, Bose–Einstein condensates, graphene sheets, optical fibers, optical glass, and laser pulse analogues. Some of these have so far only been used in probing questions about gravity and space-time theoretically, but some have also been used to actually construct analogue models in the laboratory [39.8].

### 39.2.1 Lessons from the Nineteenth Century

The use of analogue models in investigating cosmology, for example, analogue space-times, or analogue gravity, may seem quite distant from the more familiar analogue models in geophysics and nineteenth century mechanics, but, conceptually, it actually looks like a most natural outgrowth of them. In the late nineteenth century, while engineers were developing similarity methods to improve their designs of and predictions about ships and structures, physicists explicitly employed analogies to help them think through theory and come up with experiments about light, heat, sound, electricity, and magnetism. One of the most well known of these was the analogy between light and sound that was based on the fact that both were waves; another was the analogy between fluid, heat, and electrical currents that was based on the fact that the partial differential equations describing all three such *flows* were of the same form.

### 39.2.2 Sound as an Analogue of Light: The Power of Experimentation on Analogues

The analogy between light and sound was especially fruitful in the development of the correct understanding of the Doppler effect. (The Doppler effect is the change in frequency observed due to relative motion between the source of a wave and an observer. If the relative motion between source and observer is toward each other, the observed frequency increases; if the relative motion between source and observer is away from each other, then the observed frequency decreases.) To put it more precisely, the relevant factor is the ratio of the velocity of the relative motion between observer and source to the velocity  $c$ , where  $c$  is the velocity of sound for a change in pitch, or the velocity of light for colour shift. Because the velocity of light is so high, the velocity of motion required to create an observable change in pitch is much, much less than that required to create an observable change in colour [39.3, p. 18]. Mach devised and carried out laboratory experiments in which changing the relative motion between the observer and the sound source resulted in changes in observed frequency [39.2], concluding in 1860 that “the fluctuation in the pitch is dependent on no other circumstance than the direction and speed [of the source] with respect to the observer” ([39.1] and [39.3, p. 18]).

Could the laboratory experiments using sound be considered an analogue for light in a serious scientific sense? Eventually, Mach became convinced that they could be; in 1878 he published an article on it, no longer hesitant about extending the Doppler principle from his experiments on sound to the realm of optics. The Doppler effect for light, he argues, follows from the characteristics of light that are common to both sound and light waves. He does not need to assume that light is a mechanical wave in order to extend

his results from sound to light. The characteristics that light and sound have in common that are relevant to applying the Doppler principle are things such as being propagated in time with a finite velocity, having spatial and temporal periodicity, and being able to be algebraically summed. This characterization does not assume the existence of a medium for light. Mach considers the experiments on sound to be confirmatory for light, and argues that on the basis of them he can conclude with confidence that the Doppler principle applies to light [39.1, 3].

To lay out the reasoning that Mach eventually uses in claiming that his experiments on the Doppler effect for sound waves in the laboratory are confirmatory of the Doppler effect for light propagation in astronomy: His theory is that it is the relative motion of wave (or signal) source and observer that is responsible for the Doppler effect. Mach identifies the characteristics common not only to light and sound, but also to all oscillatory motion, that he believes are sufficient for the occurrence of the Doppler effect. He shows that any oscillatory motion having these characteristics will give rise to the Doppler effect, according to his line of reasoning. The mechanism does not matter; he is explicit about this point, in part because he wishes to emphasize that the existence of the Doppler effect for sound does not depend on features of the medium of transmission. Mach then experimentally confirms that, as his reasoning predicts, the Doppler effect arises for sound in a laboratory setup that allows him to manipulate the relative motion of signal source and observer. Based upon the fact that light has in common with sound those characteristics he has shown are responsible for giving rise to the Doppler effect according to his line of reasoning, he concludes that the experimental confirmation of his experiments for sound in the laboratory applies to light in astronomical observations [39.3]. As we shall see, this kind of approach using analogue models is very much like approaches still in use in physics today.

### 39.2.3 Water as an Analogue of Electricity: Limitations of Generalizing from Analogues

The nature and limits of analogical reasoning, including reasoning about experiments on analogues, was also a concern of nineteenth century physics. In a paper entitled *On Discontinuous Movements Of Fluids* (which is discussed in more detail in Chap. 18, in this volume), Helmholtz points out both the invaluable role that analogue models can fulfill, and the limitations they may display.

As for the limitation of analogical reasoning on the basis that two things instantiate the same equation,

*Helmholtz* notes that “the partial differential equations for the interior of an incompressible fluid that is not subject to friction and whose particles have no motion of rotation” are precisely the same as the partial differential equations for “stationary currents of electricity or heat in conductors of uniform conductivity” [39.45]. Yet, he notes, even for the same configurations and boundary conditions, the behavior of these different kinds of currents can differ [39.45, p. 58]. The explanation he gives is that in some situations, “the liquid is torn asunder,” whereas electricity and heat flows are not. Based upon observations, the difference in behavior between fluid currents on the one hand and electrical and heat currents on the other is due to “a surface of separation” that exists or arises in the case of the fluid.

*Helmholtz* identifies another method [39.46, p. 68]:

“In this state of affairs [the insolubility of the hydrodynamic equations for many cases of interest] I desire to call attention to an application of the hydrodynamic equations that allows one to transfer the results of observations made upon any fluid and with an apparatus of given dimensions and velocity over to a geometrically similar mass of another fluid and to apparatus of other magnitudes and to other velocities of motion.”

The method Helmholtz is referring to, which he presented in this now-classic paper (originally published in German in 1873), thus differs from deducing predictions from theory. The method he presents there does *make use of* the fact that the same equation applies to both situations to provide a basis for using one situation as an analogue for another. However, Helmholtz derives the dimensionless parameters that must be made the same between analogue model and what is modeled. The topic is discussed in more detail in Chap. 18. What Helmholtz describes is a special case of a scale model, for he specifies that the bodies are to be geometrically similar, and involve fluids to which the hydrodynamic equations apply. This is a use of an analogue model in which the basis for drawing the analogy, although it *makes use of the fact* that there is an equation instantiated by analogue and thing modeled, *does not rely on that fact alone*: there are dimensionless parameters that must be held the same between analogue and thing modeled.

This kind of concern about basing the use of an analogue model on the fact that the analogue model and what it models both instantiate the same equation is reflected in critiques of their use in physics today. The concern Helmholtz raised in the nineteenth century (about the surfaces of separation that arise in fluid flow) regarding the limitations of analogues between differ-

ent kinds of flow (heat, electrical, hydrodynamic) arises today not only when there is a difference in material or what is flowing, but also when the scales between laboratory model and what is modeled are so different that one cannot assume that the same forces or mechanisms are at work in analogue model and what is modeled.

Whether the physicists using analogue approaches in twenty-first century physics realize it or not, the new methods they are developing and the concerns they are raising about them have precursors in the nineteenth century.

### 39.2.4 Some Recent Results Using Analogue Models

The use of analogue models of space-time in the twenty-first century involves drawing analogies between flows of various sorts, too. However, the interest does not seem to be historically continuous with the nineteenth century efforts.

#### Unruh's 1981 *Experimental Black-Hole Evaporation?*

The current interest in analog models of gravity is usually traced to a paper by William Unruh published in 1981. In that very short paper, *Unruh* addressed what he called “one of the most surprising discoveries of the past ten years”: black hole evaporation. He noted that [39.47]

“experimental investigation of the phenomenon would seem to be virtually impossible, and would depend on the highly unlikely discovery of a small black hole (a relic of the initial stages in the life of the universe perhaps) of the earth.”

However, he said, “a physical system exists which has all the properties of a black hole as far as thermal radiation is concerned, but in which the physics is completely understood” [39.47].

The physical system he referred to was a sound wave propagated in supersonic flow. He restricted consideration to cases of “the background fluid smoothly exceeding the velocity of sound” which, he notes, can be assured by the use of “a suitably shaped nozzle” [39.47, p. 1352, n. 7]. Indeed, such a “suitably shaped” nozzle exists; the DeLaval nozzle (Fig. 39.10) was invented

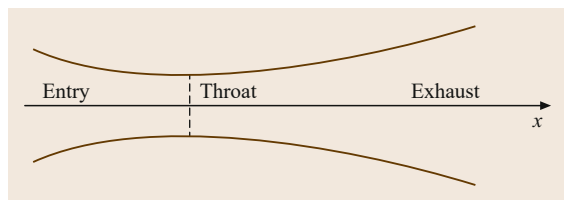


Fig. 39.10 DeLaval nozzle (after [39.48, p. 380])

in the nineteenth century for steam applications, and is used in rocket design and many other applications today.

This insightful invention can be used to create conditions for smooth supersonic flow (i.e., a region in which flow is supersonic but a shock wave does not occur) [39.48, p. 380].

Normally, in a pipe or convergent (decreasing cross-sectional area) nozzle, once flow reaches the critical flow, or *choked* conditions, the velocity of the flow at the throat of the nozzle does not increase, even if the pressure upstream increases. However, if the convergent nozzle has an appropriately designed divergent nozzle attached to its outlet, it is possible for the velocity of the flow to increase in the divergent (increasing area) portion of the nozzle, after passing through the throat (minimum area) of the nozzle. The striking thing about this situation (as many engineers are aware) is that changes in the downstream pressure do not affect the rate of flow in the nozzle. Considering the phenomenon in terms of pressure signals, one way to think of this is that the information that the pressure downstream has changed (i.e., a pressure signal or pressure pulse) cannot travel back upstream to the throat of the nozzle [39.49]. This feature of flow in a DeLaval nozzle is depicted in Fig. 39.11.

This physical situation is an analogue model, in that, as *Unruh* put it: “The model of the behavior of a quantum field in a classical gravitational field is the motion of sound waves in convergent fluid flow” which, he added [39.47, p. 1353]:

“forms an excellent theoretical laboratory where many of the unknown effects that quantum gravity could exert on black hole evaporation can be modeled [...] the phonons emitted are quantum fluctuations of the fluid flow and thus affect their own propagation in exactly the same way that graviton emission affects the space-time on which the various relativistic fields propagate.”

He had some doubts about how detectable the emission would be in that physical system, though.

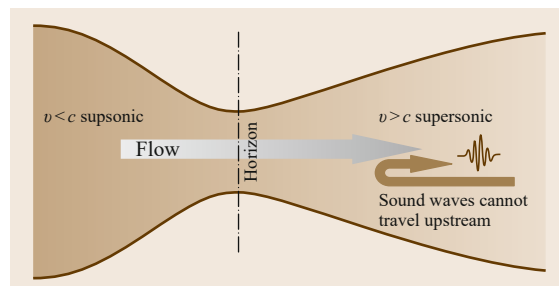


Fig. 39.11 Scheme of a DeLaval nozzle showing important features of flow behavior (after [39.49, p. 89])

### Reasoning About Analogues of Gravity Before and After 1981

Visser notes that, actually, notions of analogues of gravity “have, to some extent, been quietly in circulation almost since the inception of general relativity itself” citing Walther Gordon’s introduction of “a notion of *effective metric* to describe the effect of a refractive index on the propagation of light” and “notions developed in optics to represent gravitational fields in terms of an *equivalent refractive index*” [39.6]. *Max Born’s* sonic analogue of the kinematics of special relativity might be cited as another example of sorts [39.50, p. 251], also quoted and discussed in [39.3]:

“[...] if we use sound signals to regulate the clocks, Einstein’s kinematics can be applied in its entirety to ships that move through motionless air. The symbol  $c$  would then denote the velocity of sound in all formulae. Every moving ship would have its own units of length and time according to its velocity, and the Lorentz transformations would hold between the systems of measurement of the various ships. We should have before us a consistent Einsteinian world on a small scale.”

Nonetheless, Unruh’s short 1981 paper is regarded as marking a new era in the revival of analogue notions of gravity, although the thinking concerning the analogue involved in it has progressed since then. Whereas, in the 1981 paper, Unruh had proposed an analogue involving sound traveling in a fluid flowing through a nozzle, by 2002, *Schutzhold* and *Unruh* [39.51] proposed “gravity waves of a flowing fluid in a shallow basin” as an analogue to study black holes in a curved space-time. The surface wave setup permitted manipulations not available in the earlier proposal, due to being able to alter the depth of the water. The reasoning *Schutzhold* and *Unruh* use here appeals to “similar equations”, that is, that the behavior of interest in the analog and the behavior of interest in what the analogue models are described by equations of the same form [39.51, emphasis added]:

“Analogues, which obey similar equations of motion to fields around a black hole raise the possibility of demonstrating some of the most unusual properties of black holes in the laboratory. This is the basic idea of the black and white hole analogs [...] originally proposed by Unruh [in 1981] [...]. The sonic analogs established there are based on the observation that *sound waves in flowing fluids are (under appropriate conditions) governed by the same wave equation as a scalar field in a curved space-time*. The acoustic horizon, which occurs if the velocity of the fluid exceeds the speed of sound within the

liquid, acts on sound waves exactly as a black hole horizon does on, for example, scalar waves.”

However, there is also reasoning about aspects of the analogue model (surface waves in fluid) that is *not* part of the analogy, in a way that aims to show how the fuller knowledge we have about the analogue model might be drawn upon [39.52, p. 2908]:

“In the case of a fluid, one knows that the fluid equation of motion is inapplicable at high frequencies and short wavelengths. At wavelengths shorter than the interatomic spacing, sound waves do not exist and thus the naive derivation of the temperature of [sonic analogues of black] holes will fail. But unlike for black holes, for [sonic analogues of black] holes, the theory of physics at short wavelengths, the atomic theory of matter, is well established. For black holes, a quantum theory of gravity is still a dream. Thus, if one could show that for [sonic analogues of black] holes the existence of the changes in the theory at short wavelengths did not destroy the existence of thermal radiation from a [sonic analogue of a black] hole, one would have far more faith that whatever changes in the theory quantum gravity created, whatever nonlinearities quantum gravity introduced into the theory, the prediction of the thermal radiation from black holes was robust.”

This is basically an attempt to identify the *relevant characteristics* of a system of which the thermal radiation is a consequence. The approach is reminiscent of Mach’s approach in investigating the Doppler effect: if features other than the relative motion of source and observer did *not* make a difference to the existence of the Doppler effect, that would increase confidence (or show to those who were sceptical) that the Doppler effect depended only on the relative motion of source and observer. In this case (sonic analogues of black holes), if features *other than those related to instantiating the equation* that the black hole and the sonic analogue for it had in common seemed *not* to make a difference to the existence of a certain effect, then one’s confidence that the effect followed in virtue of a system instantiating the equation would be strengthened. But it could go the other way, too [39.52, p. 2908]:

“On the other hand, if the introduction of the atomicity of matter invariably destroyed the thermal radiation for [sonic analogues of black] holes, one would strongly suspect that the thermal nature of black holes would not survive the complications introduced by quantum gravity.”

Unruh had shown concern from the start that Hawking’s derivation of black hole evaporation relied upon

assumptions he describes as *absurd*. There was thus value in being able to sort out which features or characteristics of the situation are actually responsible for the Hawking effect. Visser had argued [39.53] that Hawking radiation is a kinematic effect independent of dynamics; In Schutzhold and Unruh's 2002 paper, they remark on what this means for the value of analogue models of gravity [39.51]:

“Although the kinematics of the waves propagating within the black and white hole analogs are governed by the same equation as those in a curved space-time, the dynamics of the effective metric itself are not described by the same laws as gravity (i. e., the Einstein equations) in general. In this way the analogs allow one to separate the dynamical effects of gravity (following from the Einstein equations) from more general (kinematic) phenomena.”

Our interest here is in the methodologies that are used to underwrite the use of analogue models to serve as models of what they are used to model. However, it is worth noting that this use of analogue models – that is, using them to help sort out what the phenomenon of interest is dependent upon for its existence – is a valuable help that analogue models can provide when there is dispute about the dependency. It is not a new kind of reasoning, of course, for we saw that Mach used it in his experiments on the Doppler effect, especially to argue against a view held by others (e.g., Petzval) that the phenomenon arose from features of the medium of transmission. Nor does reasoning about sorting out dependencies necessarily require the use of analogue models or experimentation; mathematicians often show that a result can be proven with fewer assumptions than

currently known proofs, without resorting to laboratory experiments to do so.

#### *Sorting out Dependencies: Measurements of Stimulated Hawking Emission in an Analogue System.*

When Hawking radiation was finally measured in an analogue model of a black hole, it was for exactly this benefit – sorting out dependencies – that it was especially valued, and the interest was in observing the existence of the phenomenon. Unruh argued for his view that the result of the experiment counted as a genuine measurement of Hawking radiation in 2010; there are also comprehensive reports on the experiment by the experimental team headed by *Silke Weinfurter* et al. [39.54, 55]. The laboratory setup is shown in Fig. 39.12 [39.54, Figure 8.2]: a region of high-velocity flow, including (surface wave) horizons, is created by placing a streamlined obstacle in the water flume [39.54, Figure 8.4], as seen in Fig. 39.13.

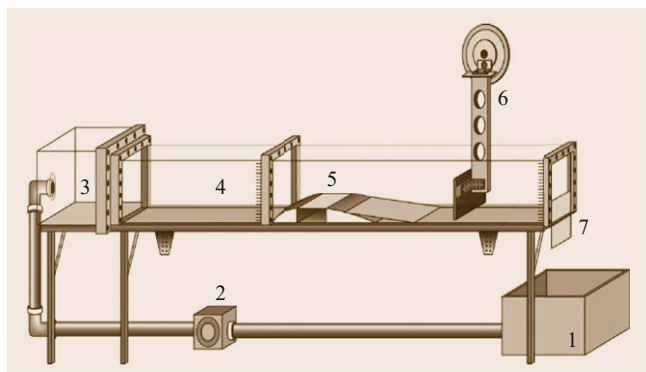
Against this flow, long waves are propagated, which become blocked and converted into short waves, thus creating a laboratory analogue of the behavior of interest [39.55, p. 120312–2]:

“It is this blocking of the ingoing waves that creates the analogy with the white hole horizon in general relativity. That is, there is a region that the shallow water waves cannot access, just as light cannot enter a white hole horizon. Note that while our experiment is on white hole horizon analogs, they are equivalent to the time inverse of black hole analogues.”

*Rousseaux* illustrates the effect occurring in nature: Fig. 39.14 below, of a white hole formed where a river enters the sea, is one such example from his work [39.56]. Other examples are fluid being poured into a sink, and a whale's *fluke-print* [39.56, p. 99]:

“As a whale swims or dives, it releases a vortex ring behind its fluke at each oscillation. The flow induced on the free surface is directed radially and forms an oval patch that gravity waves cannot enter [...].”

In the 2011 paper by Weinfurter et al., *Measurement of Stimulated Hawking Emission in an Analogue*



**Fig. 39.12** Experimental apparatus in Weinfurter et al.: Analogue model of Hawking emission (after [39.54]): (1) holding reservoir, (2) pump and pump valve, (3) intake reservoir, (4) flume, (5) obstacle, (6) wave generator, (7) adjustable weir



**Fig. 39.13** Obstacle in Weinfurter et al. experiments (after [39.54]): (1a) and (1b) curved parts motivated by airplane wing; (2) flat aluminum plate to further reduce flow separation; and (3) flat top aluminum plate to reduce wave tunneling effects



**Fig. 39.14** White holes in nature. Where a river enters the sea, the sea waves may be *blocked* (after [39.56])

*System*, the authors note that they already had numerical studies indicating that “the [Hawking] effect is independent of short-wavelength physics” [39.55]. The motivation for experimentation on the analogue model is that, were they to be able to show that there was thermal emission in their physical setup, this would “indicate the generic nature of the Hawking thermal process.” This is because the water tank/flume physical system “exhibits turbulence, viscosity, and nonlinearities”; their argument seems to be that the existence of thermal emission *in spite of* these would show that the process is a feature that follows from the wave kinematics of the physical setup, not forces arising due to these other features of the setup. And, if that were true, the result would be a very general one, applying to waves of any sort.

The measurement result is not as simple as providing the value of a quantity; the researchers identified a certain dimensionless parameter having to do with the ratio of the amplitudes of the waves that were absorbed to those that were emitted, which is important in

describing thermal (Hawking) emission, and they investigated how it scaled with frequency. This was a check that what they were observing did have the character of the theorized emission, that is, this ratio scaled with frequency in just the way one would expect thermal (Hawking) emission to.

The role of the results from the analogue model do not have their significance as stand-alone results; as Weinfurter et al. point out, there are certainly other, theoretical reasons for suspecting the thermal emission to be independent of quantum gravity or Planck-scale physics. There are other reasons for departing from the view Hawking had when he first discovered it, that is, for abandoning the view that thermal emission was “a feature peculiar to black holes” [39.54, p. 179], notably *Visser’s* work in 1998 mentioned above, in which he argued that Hawking radiation is a “purely kinematic effect” that is “generic to Lorentzian geometries containing event horizons” [39.53, 57]. The experimentation on an analogue model demonstrated how to create and measure such thermal emission in a laboratory setup with classical waves, lending support to the theoretical conclusions that the phenomenon is a far more general feature of waves in systems of a certain sort. *Weinfurter* et al. close the paper in which they report their results with an indication of even more valuable work that experimentation on analogue models of black holes might do [39.55]:

“It would still be exciting to measure the spontaneous emission from a black hole. While finding small black holes to test the prediction directly is beyond experimental reach, such measurements might be achievable in other analogue models, like Bose–Einstein condensates, or optical fibre systems. [citing such models being developed by others.]”

### 39.3 Comparing Fundamental Bases for Physical Analogue Models

Analogue models are often described in terms of an equation that both a physical model and what it models have in common; on such an account, an analogue model is described as one of two physical setups instantiating a certain equation. Common examples are a hydraulic system and an electrical system (each of which can serve as an analog model of the other), or an oscillating electric circuit modeling an oscillating mechanical device. This is certainly one kind of basis for an analogue model. However, as the discussions in this chapter drawing on actual scientific practice indicate,

instantiating the same equation is not always the basis used to develop, justify, and reason using analogue models. The actual logic involved is often far more sophisticated, and sometimes does not require as much knowledge about the phenomenon one wished to bring about as the approach of exhibiting an equation that is instantiated by both the model and the thing modeled does. Gathering together the insights above, along with points made in the sources cited, three kinds of bases for analogue models can be identified. These are depicted in Table 39.2. Besides the account just men-

**Table 39.2** Three bases for experimentation on a physical analogue model

Basis for analogy between physical analogue model and thing modeled	Knowledge relied upon
<p>1. Equation in common The equation governing a quantity or phenomenon in the model is the same equation governing the corresponding quantity or phenomenon in what is modeled, although the quantities in the equation refer to different physical things, processes, or systems when used to describe the model and what is modeled. Experimentation on the analog model (e.g., flow currents in a fluid body) is used to inform the researcher about what will happen in the analogous setup modeled (e.g., electrical currents in an electrically conducting body), or vice versa.</p>	<p>In the example of a fluid body as an analogue of an electrically conducting body, the <i>knowledge relied upon</i> includes the partial differential equation for fluid flows governing the behavior of flow currents in the analog model, the analogous equation for electrical current in the electrically conducting body, and that they can be put in the same form so as to permit drawing correspondences between the fluid quantities (e.g., flow velocity, pressure) and electrical quantities (e.g., current, voltage). Note: Behavior of analogue model and thing modeled can diverge. Key example: Helmholtz's <i>On Discontinuous Movements Of Fluids</i> [39.45] in which a surface of separation arises in the case of fluids, but not in the case of electrical flows – even though the partial differential equation takes the same form for both flows. Nevertheless, there are cases for which this discontinuity is known not to arise (the pressure does not become negative in the case of interest). Which cases are which is one piece of knowledge relied upon in using this kind of basis for the analogue model.</p>
<p>2. Relevant characteristics in common Characteristics essential to behavior of interest are the same in the model as in the thing or systems modeled. Sometimes the features are derived from equations and/or analyses of mechanisms.</p>	<p>Relies upon knowledge as to which characteristics are essential to the behavior of interest. Examples: (a) <i>E. Mach</i> identified characteristics essential a wave that did not depend upon the existence of a medium of transmission and showed the Doppler effect due to relative motion of source and observer was a consequence of these features, then verified by experiments on analogue model [39.1–3] (b) <i>S. Weinfurter</i> et al. identified classical features of Hawking process that did not depend upon quantum gravity or Planck-scale physics (e.g., wave pair formation), showed Hawking radiation a consequence of these features of waves, and then subsequently verified the phenomenon by experiment on the analogue model [39.54, 55]</p>
<p>3. Physically similar systems A (nonunique) set of dimensionless parameters that characterize the system with respect to a certain kind of behavior is identified; similarity of system behavior between <math>S</math> and <math>S'</math> is established when these parameters have the same value in <math>S</math> as in <math>S'</math> (see Chap. 18, this volume)</p>	<p>Relies upon knowledge as to which quantities (e.g., viscosity, density) are relevant to the behavior of interest. (Generally this is <i>less</i> information than required in 1. above) (Note: There is also reliance on the basic assumption that the behavior is rule-governed, in that it is assumed that there is a relation (possibly unknown) between the quantities relevant to the behavior of interest, and that the relation(s) can be expressed by a physical equation.)</p>



tioned, which is listed below as *equation in common*, there are two others: *characteristics essential to the behavior of interest are the same in the model as in the thing or system modeled*; and *physically similar systems*.

### 39.3.1 Three Kinds of Bases for Physical Analogue Models

These three kinds of bases are listed below and the points made about them are organized in Table 39.2.

#### Equation in Common

When the basis for the analogy between the analogue model and what is modeled is an appeal to the fact that an equation governing a quantity or phenomenon in the model is the same equation governing the corresponding quantity or phenomenon in the thing modeled, we will say the basis is having an *equation in common*. The equation may refer to different physical things, processes, or systems in the analogue model than in what is modeled.

To illustrate with an example using a hydraulic circuit as an analogue of an electrical circuit, experimentation on the analogue model (e.g., hydraulic setup) is used to inform the researcher about what will happen in the analogous setup modeled (e.g., the electrical circuit). The knowledge relied upon in constructing and using the model is the partial differential equation of fluid flow governing the hydraulic flow behavior in the analogue model, the analogous equation for electrical current in the electrical circuit, and that they can be put in the same form so as to permit drawing correspondences between the fluid quantities (e.g., flow velocity, pressure) and electrical quantities (e.g., current, voltage).

This kind of basis has a potential vulnerability: the behavior of analogue model and thing modeled can diverge. A key example of this is due to *Helmholtz' On Discontinuous Movements of Fluids* (1868) in which a "surface of separation" [39.45, p. 59] arises in the case of fluids, but not in the case of electrical flows – even though the partial differential equation takes the same form for both flows [39.45]. Nevertheless, there are cases for which the discontinuity Helmholtz cites is known not to arise (i.e., the pressure does not become negative in the case of interest). Which cases are which is one piece of knowledge relied upon (or, sometimes, explored) when using this kind of basis for the analogue model.

#### Relevant Characteristics in Common

When the basis for the analogy between the analogue model and what is modeled is the fact that the charac-

teristics essential to the behavior of interest are the same in the model as in the thing or system modeled, we will call it a case of *relevant characteristics in common*.

Sometimes the relevant characteristics are derived from equations and/or analyses of mechanisms, so the knowledge relied upon may involve some of the same information that is relied upon in the *equation in common* approach; it depends upon the reasoning the researcher uses to decide which characteristics of the situation are relevant. Sometimes a researcher may employ an approach based on partial knowledge (i.e., less knowledge than that needed to solve the problem) – but the partial knowledge may be enough to identify what the characteristics relevant to producing the behavior of interest are. Or, the partial knowledge may be enough to show that an analogue model and what it models will display the same behavior of interest, so that experimenting on the analogue is informative about how what it models will behave. Some of the scaling arguments used in developing the analogies underwriting models of analogue gravity are examples of this, as in *Rousseaux's* discussions of the differences between shallow and deep water tank analogue models [39.56]. *Rousseaux's* scaling arguments are based upon dimensional analysis, though he does not lay out and solve the problem as one of physically similar systems.

Other examples of this sort are Ernst Mach's work on the Doppler effect and work by *Weinfurter et al.* on identifying the classical features of the Hawking process: E. Mach identified characteristics essential to being a wave that did not depend upon the existence of a medium of transmission (spatial and temporal periodicity, finite velocity, can be algebraically summed) and showed the Doppler effect due to relative motion of source and observer was a consequence of these features, which he then verified by experiments on an analogue model [39.1, 3]. *Weinfurter et al.* identified classical features of the Hawking process that did not depend upon quantum gravity or Planck-scale physics (e.g., wave pair formation), then showed Hawking radiation a consequence of these features of waves, and subsequently verified by experiment that the analogue of Hawking radiation occurred on the analogue model [39.54, 55].

#### Physically Similar Systems

When employing the method known as *physically similar systems*, a (nonunique) set of dimensionless parameters that characterizes the system with respect to a certain kind of behavior is identified using the method of dimensional analysis; similarity of system behavior between  $S$  and  $S'$  is established when these parameters have the same value in  $S$  as in  $S'$  (see Chap. 18, this volume). The knowledge relied upon for this method

is the knowledge as to which quantities are relevant to the behavior of interest. Generally this is *less* information than is required in the *equation in common* method. In a much more fundamental sense of reliance, the researcher is also relying on a basic assumption made

implicitly in much of scientific research: that the behavior is rule-governed, in that it is assumed that there is a relation (possibly unknown) between the quantities relevant to the behavior of interest, and that the relation can be expressed by a physical equation.

## 39.4 Conclusion

The discussion presented here, of a few selected examples of analogue models and the investigation of the different bases for their use, should help put to rest three common misconceptions about the use of analogue models in physics today.

*First, the misconception that analogue models are a thing of the past:* As the examples discussed above indicate, analogue models are not a thing of the past. In fact, there are new areas of application and new kinds of analogue models being developed; the recent surge in development of analogue models in general relativity is one striking example [39.5, 57]. Yet, the use of some of these models in the twenty-first century does have precursors in the nineteenth century.

*Second, the misconception that analogue models serve merely illustrative or pedagogical purposes.* Many of the examples described above are cases of serious investigative research. This is so even when the main benefit of the model is gaining a better qualitative understanding of the mechanisms at work. *Rousseaux* remarks that the investigation of analogue gravity “through the prism of water waves theory has broadened our definition of a horizon” [39.56, p. 106]. In geophysics, experimentation on analogue models has sometimes brought about an appreciation of mechanisms that might be at work and ought to be investigated, which is a kind of discovery [39.43].

*Third, the misconception that numerical methods along with high-speed digital computers can always provide whatever an analogue model could provide.* This is the most pernicious and deep-seated of the three misconceptions. It betrays a fundamental misunderstanding of the logic behind analogue models. Such statements are probably based on assuming that the basis for analogue models is having an equation in common. As Table 39.2 indicates, there are other bases for using analogue models than having an equation, and it is not the case that an analogue model can in principle always be replaced by an equation: the method of dimensional analysis and physically similar systems often requires less information than the method of finding an equation that is instantiated by both the analog model and what it models does. (That is why similarity methods are sometimes called partial information meth-

ods [39.58].) More importantly, even when one does have such an equation, and cites it as the basis for the analogy between analogue model and what is modeled, the role of the equation is to establish a correspondence between items in the analogue model and what it models. Even if one has the means to solve that equation numerically on a digital computer, there is no guarantee that the numerical solution of the equation will reveal every phenomenon that might be observed in the analogue model. It has been the case many times that the use of an analogue model shows a phenomenon that the numerical solution of the equation had not, no matter how many colorful visuals and graphics the computer program is capable of producing. Analogue models do need to be examined to determine when they are and are not appropriate for a certain investigation, but so, of course, do equations and numerical simulations. The fact that the use of analogue models in various fields has been revived (after supposedly being eliminated from those fields), often for new applications and employing new technologies, reflects something that is becoming increasingly clear: sometimes analogue models really are irreplaceable.

*Acknowledgments.* Thanks to Matt Walhout for suggesting, many years ago, that I might find William Unruh’s work on sonic analogues of black holes of interest. Later, in April 2011, I had the good fortune to attend Unruh’s lecture *Measurement of Hawking Radiation in an Analog System* and the discussion afterward, at the University of Pittsburgh.

Thanks also to the organizers of the conference *Philosophy of Scientific Experimentation III* (PSX3), for financial support to present the talk this paper is based upon *Experimentation on Analogs* at PSX3 on October 5, 2012, at the Department of Physics, University of Colorado, Boulder. I benefitted from comments by, and discussion with, the other participants of PSX3 on the many papers related to analogy, including James Mattingly’s talk *Experimental Cosmology*, which also discussed experimentation on an analogue model (using Bose–Einstein condensates).

Another paper on the topic of the same experiments on sonic analogues of black holes, as are discussed in

this paper, was presented at the Philosophy of Science Association Biennial meeting in late 2014. As that presentation by Dardashti (and the subsequent publication of a related paper by Dardashti et al. *Confirmation via Analogue Simulation: What Dumb Holes can tell us*

*About Gravity* (2015)) occurred more than 2 years after I submitted and presented *Experimentation on Analogues* at PSX3 in October 2012, the talk on which my article for this volume is based, their commentary on those experiments is not discussed here.

## References

- 39.1 W.W.G.J. Swoboda: The Thought and Work of the Young Ernst Mach and the Antecedents to His Philosophy, Ph.D. Thesis (Univ. Pittsburgh, Pittsburgh 1973)
- 39.2 J.T. Blackmore: *Ernst Mach: His Work, Life, and Influence* (Univ. California Press, Berkeley 1972)
- 39.3 S.G. Sterrett: Sounds like light: Einstein's special theory of relativity and Mach's work in acoustics and aerodynamics, *Stud. Hist. Philos. Mod. Phys.* **29**, 1–35 (1998)
- 39.4 P.A. Cherenkov: Nobel Lecture: Radiation of Particles Moving at a Velocity Exceeding That of Light, and Some of the Possibilities for Their Use in Experimental Physics, [http://www.nobelprize.org/nobel\\_prizes/physics/laureates/1958/cerenkov-lecture.html](http://www.nobelprize.org/nobel_prizes/physics/laureates/1958/cerenkov-lecture.html) (2014) Nobel Media AB 2014. Web. 2015
- 39.5 C. Barcelo, S. Liberati, M. Visser: Analogue gravity, *Living Rev. Relativ.* **14**(3), 1–179 (2011)
- 39.6 M. Visser, C. Barcelo, S. Liberati: Analogue models of and for gravity, *Gen. Relativ. Gravit.* **34**(10), 1719–1734 (2002)
- 39.7 D. Faccio: Laser pulse analogues for gravity and analogue Hawking radiation, *Contemp. Phys.* **53**(2), 97–112 (2012)
- 39.8 D. Faccio, F. Belgiorno, S. Cacciatori, V. Gorini, S. Liberati, U. Moschella (Eds.): *Analogue Gravity Phenomenology: Analogue Spacetimes and Horizons from Theory to Experiment* (Springer, Cham 2013)
- 39.9 M. Visser: Survey of analogue spacetimes. In: *Analogue Gravity Phenomenology: Analogue Spacetimes and Horizons from Theory to Experiment*, ed. by D. Faccio, F. Belgiorno, S. Cacciatori, V. Gorini, S. Liberati, U. Moschella (Springer, Cham 2013) pp. 31–50
- 39.10 I. Carusotto, G. Rousseaux: The Cerenkov effect revisited: From swimming ducks to zero modes in gravitational analogues. In: *Analogue Gravity Phenomenology: Analogue Spacetimes and Horizons from Theory to Experiment*, ed. by D. Faccio, F. Belgiorno, S. Cacciatori, V. Gorini, S. Liberati, U. Moschella (Springer, Cham 2013) pp. 109–144
- 39.11 B. Ayres: Mechanical models of the electric circuit, *Electr. World* **28**, 276–277 (1896)
- 39.12 C.F. Jenkin: A dynamic model of tuned electrical circuits, *Proc. Inst. Electr. Eng.* **60**, 939–941 (1922)
- 39.13 J.S. Small: *The Analogue Alternative: The Electronic Analogue Computer in Britain and the USA, 1930–1975* (Routledge, New York, London 2001)
- 39.14 V. Bush, S.H. Caldwell: A new type of differential analyzer, *J. Frankl. Inst.* **240**(4), 255–325 (1945)
- 39.15 A.G. MacNee: *An Electronic Differential Analyzer*, Technical Report No. 90, Research Laboratory of Electronics (MIT Press, Cambridge 1948)
- 39.16 B. Randell (Ed.): *The Origins of Digital Computers: Selected Papers*, Springer Monographs in Computer Science (Springer, Berlin, Heidelberg 1982)
- 39.17 C. Isenberg: The soap film: An analogue computer, *Am. Sci.* **64**, 514–518 (1976)
- 39.18 M. Aono, Y. Hirata, M. Hara, K. Aihara: Combinatorial optimization by ameoba-based neurocomputer with chaotic dynamics. In: *Natural Computing*, Vol. 1, ed. by Y. Suzuki, M. Hagiya, H. Umeo, A. Adamatzky (Springer, Tokyo 2009) pp. 1–15
- 39.19 E. Buckingham: Physically similar systems: Illustrations of the use of dimensional equations, *Phys. Rev.* **4**, 345–376 (1914)
- 39.20 H.L. Langhaar: *Dimensional Analysis and Theory of Models* (Wiley, New York 1951)
- 39.21 R.C. Pankhurst: *Dimensional Analysis and Scale Factors* (Chapman Hall, New York 1964)
- 39.22 S.L. Lien, J.A. Hoopes: Wind-driven, steady flows in Lake Superior, *Limnol. Oceanogr.* **23**, 91–103 (1978)
- 39.23 N.P. Wallerstein, C.V. Alonso, S.J. Bennett, C.R. Thorne: Distorted Froude-scaled flume analysis of large woody debris, *Earth Surf. Process. Landf.* **26**, 1265–1283 (2001)
- 39.24 S. De Rosa, F. Franco, V. Meruane: Similarities for the structural response of flexural plates, *Proc. Institution Mech. Eng. J. Mech. Eng. Sci.* (2015), doi:10.1177/0954406215572436
- 39.25 R. Frigg, S. Hartmann: Models in science. In: *The Stanford Encyclopedia of Philosophy*, ed. by E.N. Zalta <http://plato.stanford.edu/archives/fall2012/entries/models-science/> (Fall 2012 Edition)
- 39.26 R.N. Giere: *Science Without Laws* (Univ. Chicago Press, Chicago 1999)
- 39.27 R.N. Giere: How models are used to represent reality, *Philos. Sci.* **71**, 742–752 (2004)
- 39.28 G. Francesco: Models, simulations, and experiments. In: *Model-Based Reasoning: Science, Technology, Values*, ed. by L. Magnani, N. Nersessian (Kluwer, New York 2002) pp. 59–74
- 39.29 E. Winsberg: A tale of two methods, *Synthese* **169**(3), 575–592 (2009)
- 39.30 D. Rothbart (Ed.): *Modeling: Gateway to the Unknown: A Work by Rom Harre* (Elsevier Science, Amsterdam 2004), (Studies in Multidisciplinarity)
- 39.31 P. Kroes: Structural analogies between physical systems, *Br. J. Philos. Sci.* **40**(2), 145–154 (1989)
- 39.32 E.T. Layton: Escape from the jail of shape: Dimensionality and engineering science. In: *Techno-*

- logical Development and Science in the Industrial Age: New Perspectives on the Science–Technology Relationship*, ed. by P. Kroes, M. Bakker (Kluwer, Dordrecht, Boston 1992)
- 39.33 S.G. Sterrett: Dimensional analysis and similarity. In: *Philosophy of Technology and Engineering Sciences*, Vol. 9, ed. by A.W.M. Meijers (Elsevier, Amsterdam 2009) pp. 799–823
- 39.34 S.G. Sterrett: *Wittgenstein Flies a Kite: A Story of Models of Wings and Models of the World* (Penguin, New York 2005)
- 39.35 S.G. Sterrett: Physical models and fundamental laws: Using one piece of the world to tell about another, *Mind Soc.* **5**(3), 51–66 (2002)
- 39.36 J. Mattingly, W. Warwick: Projectible predicates in analogue and simulated systems, *Synthese* **169**, 465–483 (2009)
- 39.37 S.D. Zwart: Scale modelling in engineering: Froude’s case. In: *Philosophy of Technology and Engineering Sciences*, Vol. 9, ed. by A.W.M. Meijers (Elsevier, Amsterdam 2009) pp. 759–798
- 39.38 N. Oreskes: From scaling to simulation: Changing meanings and ambitions of models in geology. In: *Science Without Laws*, ed. by A.N.H. Creager, E. Lunbeck, M. Norton Wise (Duke Univ. Press, Durham, London 2007)
- 39.39 M.K. Hubbert: Theory of scale models as applied to the study of geologic structures, *Geol. Soc. Am. Bull.* **48**, 1459–1520 (1937)
- 39.40 M.K. Hubbert: Strength of the earth, *Bull. Am. Assoc. Petroleum Geol.* **29**, 1630–1653 (1945)
- 39.41 O. Merle, A. Borgia: Scaled experiments of volcanic spreading, *J. Geophys. Res. Solid Earth* **101**, 13805–13817 (1996)
- 39.42 G. Norini, V. Acocella: Analogue modeling of flank instability at Mount Etna: Understanding the driving factors, *J. Geophys. Res.* (2011), doi:[10.1029/2011JB008216](https://doi.org/10.1029/2011JB008216)
- 39.43 V. Gabuchian, A.J. Rosakis, N. Lapusta, D.D. Oglesby: Experimental investigation of strong ground motion due to thrust fault earthquakes, *J. Geophys. Res. Solid Earth* **119**, 1316–1336 (2014)
- 39.44 J.N. Brune: Particle motions in a physical model of shallow angle thrust faulting, *Proc. Indian Acad. Sci.* **105**, 197–206 (1996)
- 39.45 H. von Helmholtz: On discontinuous motions in liquids. In: *Mechanics of the Earth’s Atmosphere: A Collection of Translations*, Smithsonian Miscellaneous Collections, Vol. 843, ed. by C. Abbe (The Smithsonian Inst., Washington DC 1891) pp. 58–66
- 39.46 H. von Helmholtz: On a theorem relative to movements that are geometrically similar in fluid bodies, together with an application to the problem of steering balloons. In: *Mechanics of the Earth’s Atmosphere: A Collection of Translations*, Smithsonian Miscellaneous Collections, Vol. 843, ed. by C. Abbe (The Smithsonian Inst., Washington DC 1891) pp. 67–77
- 39.47 W.G. Unruh: Experimental black-hole evaporation?, *Phys. Rev. Lett.* **46**, 1351–1353 (1981)
- 39.48 R. Courant, K.O. Friedrichs: *Supersonic Flow and Shock Waves* (Springer, New York 2012), reprint of original 1948 edition
- 39.49 A. Bramati, M. Modugno (Eds.): *Physics of Quantum Fluids: New Trends and Hot Topics in Atomic and Polariton Condensates* (Springer, Heidelberg 2013)
- 39.50 M. Born: *Einstein’s Theory of Relativity* (Dover, New York 1962)
- 39.51 R. Schutzhold, W.G. Unruh: Gravity wave analogues of black holes, *Phys. Rev. D* **66**, 044019 (2002)
- 39.52 W.G. Unruh: Dumb holes: Analogues for black holes, *Philos. Trans. R. Soc. A* **366**, 2905–2913 (2008)
- 39.53 M. Visser: Hawking radiation without black hole entropy, *Phys. Rev. Lett.* **80**(16), 3436 (1998)
- 39.54 S. Weinfurtner, E.W. Tedford, M.C.J. Penrice, W.G. Unruh, G.A. Lawrence: Classical aspects of Hawking radiation verified in analogue gravity experiment. In: *Analogue Gravity Phenomenology: Analogue Spacetimes and Horizons from Theory to Experiment*, ed. by D. Faccio, F. Belgiorno, S. Cacciatori, V. Gorini, S. Liberati, U. Moschella (Springer, Cham 2013) pp. 167–180
- 39.55 S. Weinfurtner, E.W. Tedford, M.C.J. Penrice, W.G. Unruh, G.A. Lawrence: Measurement of stimulated Hawking emission in an analogue system, *Phys. Rev. Lett.* **106**, 021302–021305 (2011)
- 39.56 G. Rousseaux: The basics of water waves theory for analogue gravity. In: *Analogue Gravity Phenomenology: Analogue Spacetimes and Horizons from Theory to Experiment*, ed. by D. Faccio, F. Belgiorno, S. Cacciatori, V. Gorini, S. Liberati, U. Moschella (Springer, Cham 2013) pp. 81–107
- 39.57 M. Visser: Essential and inessential features of Hawking radiation, *Int. J. Mod. Phys. D* **12**, 649–661 (2003)
- 39.58 S.J. Kline: *Similitude and Approximation Methods* (Springer, New York 1986)

# Models of Chemical Structure

## 40. Models of Chemical Structure

William Goodwin

Models of chemical structure play dual crucial roles in organic chemistry. First, they allow for the discovery and application of *laws* to the complex phenomena that chemists hope to understand. Second, they are a source of novel concepts that allow for the continuing development of structure theory and theoretical organic chemistry. In chemistry, therefore, the centrality and significance of models to the scientific enterprise is manifest and furthermore chemistry is a relatively clear, useful, and interesting context in which to consider more general philosophical questions about the nature and role of models in science.

40.1	<b>Models, Theory, and Explanations in Structural Organic Chemistry</b> .....	881
40.2	<b>Structures in the Applications of Chemistry</b> .....	883
40.3	<b>The Dynamics of Structure</b> .....	885
40.3.1	Recognizing the Importance of Conformations.....	886
40.3.2	Using Conformations in Organic Chemistry.....	887
40.4	<b>Conclusion</b> .....	889
	<b>References</b> .....	889

One of the most important and influential trends in the philosophy of science over the last 50 years has been the increase in both the attention paid to the concept of a model and the employment of this concept in philosophical reflection on the nature and dynamics of science. This trend has been usefully described and analyzed by many philosophers (see for instance: [40.1–3]). Without trying to be exhaustive, in this paper I plan to identify a few of the most significant philosophical insights that have emerged out of this increased interest in scientific models and then to reflect on these insights in the context of chemistry, which has been relatively neglected in the philosophical literature. I hope to show both that in chemistry the centrality and significance of models to the scientific enterprise is manifest, and that chemistry is a relatively clear, useful, and interesting context in which to consider more general philosophical questions about the nature and role of models in science.

Models have been characterized in many different ways in the philosophical literature, but for the purposes of this paper it will suffice to think of them as instruments for representation that are not primarily linguistic. The important contrast is with the linguistic statements of a theory (in the logical sense). So the double helix model of DNA represents DNA molecules not because it is a statement in a language that describes this molecule, but because it is a physical object with

certain similarities to the objects that it is intended to represent. Likewise, the billiard ball model of a gas is an image of an interacting system (or the abstract idea of such a system) along with, perhaps, a narrative about how to understand this image, which can be used to represent a gas for certain purposes. Many types of objects, other than statements, have been thought of as models (including mathematical structures, abstract objects, and fictional objects). Given the diversity of things that models might be, it seems best to summarize the central insight behind the relatively recent philosophical interest in models as follows: Representational instruments that are not primarily linguistic are crucial to understanding the nature and development of science.

This central insight has been developed in a variety of ways, but I want to focus on two of them. First, models are crucial to the dynamics of science – how scientific representations, theories, experiments and concepts change over time in response to feedback from the world. And second, models are crucial to science’s capacity to confront complexity – its ability to have useful things to say about real world and/or complex systems. Models, being nonlinguistic entities, help philosophers to get a grip on these aspects of science at least in part because they are representationally rich, that is, they are not limited in their representational capacities (as purely linguistic representations would be) by the arbitrary associations between their component

symbols and aspects of the world. In the case of the dynamics of science, the rich representational capacities of models both supply (or allow for) new features that can be exploited in the models' representational role and thereby act as an incubator for novel concepts. Similarly, models can act as intermediaries between theory and complex real world phenomena because their richer resources allow for representation of the more concrete and local features crucial to understanding such phenomena.

Chemistry, like any large and diverse field of scientific inquiry, is replete with models of many different sorts. Much of the modeling in chemistry is of the standard sort discussed in the philosophical literature – that is, a response to the problem of getting abstract mathematical theories to apply to complex real world phenomena. There has been interesting philosophical work on the form that this sort of modeling takes in chemical contexts [40.4–6] and [40.7] for example; however in this paper I want to focus on what is, I think, a distinctively central and important role for modeling in chemistry. Chemistry, at least large parts of it, is concerned with representing the structures of the substances it studies. For the most part, chemists do not use linguistic resources to represent structure; instead, they build models. Sometimes, like Watson and Crick and most students of organic chemistry, they build physical models, but most frequently they use diagrammatic representations like structural formulas. While there are often linguistic components to such formulas (letters for the atoms, for example), the representational resources of these diagrams are not limited to the arbitrary relations between their component signs and predicates or terms in the current language of the theory. Furthermore, much of importance of these models for both the dynamic development of chemistry and for facilitating the application of chemical theories to concrete, real world cases derives from these extra linguistic representational resources.

Structural formulas (Fig. 40.1), which were initially developed over the course of the nineteenth century, are the centerpiece of a research program that has been immensely successful ([40.8] for a summary of the development of structural formulas). The guiding strategy of this research program, articulated by Aleksandr Butlerov in 1861, was to have one structural formula for

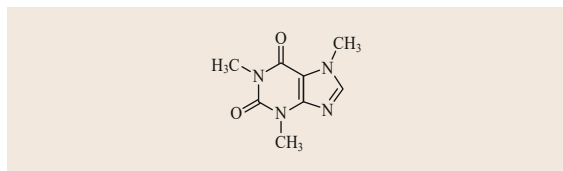


Fig. 40.1 A structural formula for caffeine

each chemical compound, and then to explain the chemical properties of these compounds by determining, “the general laws governing the dependence of chemical properties on chemical structure” [40.9, p. 256]. Structural explanations of chemical properties (and some physical properties as well) have been central to chemistry – particularly organic chemistry, on which I shall focus – ever since. A structural explanation of a chemical property proceeds by identifying the structural features of a compound that are responsible for some (usually contrastive) chemical fact. These structural features are typically general features of the compound as represented in its models, i.e., in its structural formula, which might also be realized by other structural formulas or models as well. In other words, structural patterns identifiable in the models are correlated with chemically significant facts (usually, these days, differences in energy or stability). It is these structural correlations that have ended up playing the role of the *laws* that Butlerov imagined. Thus the *laws* of structural chemistry are formulated in terms of something like chemically significant patterns in the models; in this sense, then, in structural chemistry the laws piggyback on the models. The models don't just interpret or concretize the laws, they make them possible in the first place. Understanding the complex chemical facts of organic chemistry depends, at the most basic level, on building structural models of chemical compounds.

Structural chemistry has not been a static research program. Over the course of its 160 year development, there have been immense changes in chemists' understanding of structure. And not surprisingly, these changes have been reflected in the models used to represent that structure. In the course of this continuous refinement – the back and forth between conceptions of structure and representations of it – models of structure have played a crucial role as a source of new general structural features. In other words, features of the models not originally used in their representational role are reinterpreted as representationally significant in order to explain or account for new theories or experimental facts. Additionally, even after a new aspect of *chemical structure* has been recognized and represented in the models, there still remains the daunting task of making that structural feature experimentally and synthetically relevant (making it useful) and models have also supplied some of the very local structural concepts that allow for the experimental and theoretical development of whole new subfields. Thus, just as models are crucial for structural chemistry to discover and apply laws to the complex phenomena it studies, so too are these models crucial as a source of the novel concepts necessary for the continued development of structure theory and theoretical organic chemistry.

## 40.1 Models, Theory, and Explanations in Structural Organic Chemistry

Structural formulas are the central representational tool used in the explanations and predictions of the theory of organic chemistry. These formulas play many representational roles in organic chemistry, including the role of denoting expressions for chemical kinds. In this role, they serve as descriptive names for these kinds, individuating them according to their composition, bond connectivity and (aspects of their) stereochemistry. Because they can be put (roughly) into one-to-one correspondence with the kinds that they purport to denote, they are also able to act as *stand-ins* for these chemical compounds ([40.10–12] for more on the roles of structural formulas as both names and models in organic chemistry.). When acting as stand-ins for chemical kinds, structural formulas can be *manipulated to teach us things about themselves*, and the things learned about these models can, in many cases, “be transferred to the theory or to the world which the model represents” [40.2, p. 33] ([40.13] for examples of how models acted as paper tools in the early development of organic chemistry.). That is to say that chemists can learn from their models by exploring the implications of (or for) abstract theory in the concrete contexts of particular chemical kinds and often this exploration takes the form of the manipulation of structural models. To bring out some of the interconnections between models of structure, theory, and the explanations in structural chemistry I am going to briefly consider the theory of resonance, which is a modification of structure theory developed in the first half of the twentieth century that is still central to organic chemistry today.

Though it was developed prior to any generally accepted account of the nature of chemical bonding, structure theory has had to evolve in the face of changing theoretical accounts of the nature of the chemical bond. This is to be expected given Butlerov’s aspiration for structural formulas; namely, that they explain the chemical and physical properties of the compounds that they depict. As the theoretical understanding of what a bond was changed, so to did the depictions of chemical bonds in structural formulas, and this was crucial to leveraging the revised understandings of bonding into new structural explanations of chemical or physical properties. The theory of resonance is one important way (and was historically the first broadly accepted way) that the quantum mechanical character of chemical bonding is recognized and applied within structural organic chemistry.

Even before the development of quantum mechanics, in response to the explanatory demands placed on chemical structures, several modifications of structure theory were suggested that [40.14, p. 2]:

“considered it possible for the true state of a molecule to be not identical with that represented by a single classical valence-bond structure, but to be intermediate between those represented by two or more different valence-bond structures.”

These suggested modifications were motivated by cases where a chemical compound did not behave as it would have been expected to behave given its representation using a single structural formula but where, by thinking about the compound as a mixture intermediate between two or more such formulas, the behavior could again be accounted for in structural terms. *Pauling*, in his seminal *Nature of the Chemical Bond* [40.15], laid out the theory of resonance by employing quantum mechanics to rationalize and systematize (but not to deduce), the use of multiple structures to represent individual chemical kinds and then went on to demonstrate the broad usefulness of this theory in organic chemistry.

The theory of resonance is interesting, from the point of view of the role of structural models in chemistry, for two related reasons. First, it was the manipulation of valence bond structures (which are structural formulas that explicitly depict the bonding electrons of the constituent atoms) that first revealed the possibility of explaining recalcitrant chemical and physical phenomena by thinking of chemical kinds as appropriately represented by a combination of distinct structural formulas. Facts in the world of the model – that multiple different valence bond structures were plausible for a given chemical kind – were used to suggest modifications designed to improve the explanatory power of structure theory. Furthermore, these facts about the structural models (multiple possible valence bond structures for a given kind) were systematized and rationalized using the theory of quantum mechanics so that the delocalization of bonding electrons (which is the central implication of quantum mechanics for chemical bonding) could be recognized and exploited in organic chemistry. Manipulations of the model supplied the vehicle for making quantum mechanics first applicable to structural organic chemistry. In this sense, then, the structural models mediated between the theory (quantum mechanics) and the world. Secondly, and similarly, it is the actual exploration of the range of available valence bond structures that often proves crucial to the use of the theory of resonance in generating the structural explanations that are useful in organic chemistry. In a typical case of an explanation invoking resonance, none of the individual structural formulas making up a resonance hybrid allows for the explanation of all of the chemical or physical properties of interest. In-

stead, more standard structural analysis is applied to some of the individual component formulas of the resonance hybrid, and then the behavior of the chemical kind as a whole is understood as some proportional mixture of the structural prediction based on its component structures. In other words, without the more complex depiction of a chemical kind allowed by resonance theory, it would not be possible to successfully supply structural explanations of its chemical or physical behavior. Getting the structural *laws* to apply to some chemical kinds requires a more complex model of their structure – their depiction as a resonance hybrid. Explorations in the world of the model uncovering potential significant resonance structures mediate between the theory (structure theory) and the world by allowing for the successful application of this theory to complex cases to which it would otherwise not be useful.

In order to see how the theory of resonance allows structural formulas to mediate between more general structure theory and experimental facts, it will be useful to consider some of Pauling's work on the structure of proteins. Proteins are polymers of amino acids formed when the carboxyl group of one amino acid reacts with the amino group of another forming an amide linkage called a peptide bond. As a result, there is a recurring structural pattern in proteins: tetrahedral carbon atoms (bonded to the R-groups of the amino acids) joined by amide groups (-NH-CO-). Amide groups are therefore a fundamental structural component of proteins; however, in order to predict the behavior of these groups, and thus to outline the basic structural features of proteins, it is not sufficient to consider only one of the structural formulas that can be used to represent them. Instead amide groups, at least in Pauling's treatment, were thought of as a resonance hybrid of two component structural formulas, and their structural behavior was anticipated to be a sort of weighted average of the behavior predicted by the structure theory for these individual structural formulas.

The most important feature of the amide linkage from the point of view of predicting the structure of proteins is that the carbon and nitrogen of the amide linkage lie in a single plane with the two tetrahedral carbons that they connect. *Pauling* regarded the planarity of the amide group as, "a sound structural principle" concluding that a "structure in which the atoms of the amide group are not approximately coplanar should be regarded with skepticism" [40.16, p.19]. Though he provided substantial experimental confirmation of the planarity of the peptide bond, it was the theoretical arguments for this principle that invoked the theory of resonance. An amide linkage is typically represented by a structural formula in which there is a single bond between the nitrogen and the carbonyl carbon, which

is itself double bonded to oxygen. However, another possible structural formula for the amide linkage has a double bond between the carbon and the nitrogen while there are three unshared electron pairs around the oxygen (resulting in a net formal charge of  $-1$  on the oxygen) and no unshared pairs around the nitrogen (resulting in a formal charge of  $+1$ ). The theory of resonance indicates that the first, and more typical, structure should be the most significant contributor to the overall structure of the amide linkage, but that the second structure might also be important to consider (Fig. 40.2).

The most salient difference between these two structures is where the double bond is located, either between carbon and oxygen or between carbon and nitrogen. Pauling argued based on experimental measurements of the bond lengths in some simple peptides (by x-ray crystallography), that in the actual peptide linkage (on average) the relative contribution of these two structures was 60% for the typical structure and 40% for the secondary structure. These numbers were based on the fact that the measured C-O bond length in the peptide bonds was longer than typical double bonds between carbon and oxygen (in cases where no alternative resonance structures were available) but also shorter than typical single bonds between C and O. Similarly, the measured C-N bond length was shorter than typical single bonds, but longer than typical double bonds between these atoms. If he supposed that the relative contributions of the two structures were 60% and 40% respectively, and thus that the C-O bond was 60% double and 40% single, while the C-N bond was 40% double and 60% single, then the predicted length of the bonds closely matched the measured values.

Once Pauling had argued that the second resonance structure with a double bond between carbon and nitrogen was an important contributor to the overall structure of the linkage, it followed from standard structural theory that the peptide linkage should be essentially planar. Double bonds do not allow free rotation; that is, you have to break the bond (costing a lot of energy) in order to rotate around the axis of the bond. The energetic cost of rotation around the double bond is the reason that double bonds lead to stereoisomerism (there are distinct chemical compounds, with different structural formulas, that reflect different arrangements of substituents around a double bond). Since the C-N bond in



Fig. 40.2 Resonance structures of the peptide bond



the peptide linkage was 40% of a double bond, Pauling was able to estimate that the energetic cost of rotating around this bond would be about 40% of the bond energy of a typical double bond between these atoms. Furthermore, he was able to estimate the strain energy of deviations from planarity, concluding: “we can calculate strain energies of about 0.9 kcal/mole for 10° distortion of the amide group” [40.16, p. 14]. This effectively meant that large deviations from planarity in peptide linkages would be very energetically expensive and would therefore constitute a “highly unusual steric relationship” [40.16, p. 16].

Pauling’s analysis of the structure of peptide linkages shows both what structural accounts of chemical phenomena look like, and how resonance theory allowed structural analysis to be applied to a broader range of chemical facts. The measured bond lengths of the amide linkages in peptides do not correspond to the average bond length that one would expect based on the typical representation of the amide linkage using structural formulas (with typical CO double bonds and typical CN single bonds). In order to provide a structural explanation for this deviation from expected values, Pauling needed to find some recurring structural feature of the amide linkage that could explain it. What he found, by manipulation of the formulas, is that such amide linkages could be represented by another structural formula, one obtained by redistributing the valence electrons (and thus moving around the bonds). According to the principles of resonance

theory, this alternative structure was energetically plausible and should be regarded as a potential contributor to the overall structure of the amide linkage. However, a structural analysis of this second structure by itself would also not explain the experimental bond lengths. Instead, only by regarding the actual structure as intermediate between the two resonance structures, did an explanation of the observed bond lengths become possible. Exploration in the world of the model, then, was crucial to providing a structural account of the observed bond distances. Furthermore, the success of this explanation gave Pauling confidence that his resonance structures provided a reasonable representation of the peptide bond, and thus that he could apply a structural analysis to this representation in order to make a significant structural prediction about all proteins. Again this prediction (the planarity of the peptide bond) was based on giving a weighted analysis of the various accessible resonance structures. Without the detour through the range of plausible resonance structures, explored by the chemist through manipulations of structural formulas, neither the original explanation, nor the extremely significant prediction that Pauling made about the structure of proteins would have been possible. (Now there are other ways of taking into account the implications of quantum mechanics on chemical bonding, and these can also presumably support the same explanations and predictions; but the historical fact is that Pauling, who basically initiated the study of protein structure, used resonance theory.)

## 40.2 Structures in the Applications of Chemistry

In addition to their role supporting the explanations and predictions of structural organic chemistry, structural formulas and/or models also play a crucial role in applying the theory of organic chemistry to the solution of synthesis problems. Synthesis problems are the guiding application of the theory of organic chemistry. Of course not all organic chemists are working to synthesize compounds, but this is the characteristic goal around which the field developed, and it is possible to understand the theoretical structure of the field as reflecting this goal. That is, an important reason that the explanations of organic chemistry take the form that they do – looking for structural accounts of chemical phenomena, for example – is because this approach facilitates the solution of synthesis problems [40.17, 18]. Synthesis problems have a common form, and the intellectual challenges that they present derive from this form. By understanding the basic form of synthesis problems, and the basic strategies developed for solving

these problems, it is possible to appreciate the central importance of the sorts of structural analysis undertaken in the typical explanations and predictions of organic chemistry. As seen in the last section, explorations in the world of the model, and thus the role of structural formulas as models, can be crucial to providing the structural analyses of organic chemistry. However, the role of structural formulas and models in solving synthesis problems is not limited to their support of structural explanations or predictions. Instead they play additional roles in delimiting the array of possible synthetic approaches and evaluating the plausibility of those approaches.

A synthesis problem begins with a target molecule. The chemist’s goal is to come up with a method for making this target molecule by a sequence of chemical reactions that begins with compounds that chemists already know how to make. Often, no chemist has ever synthesized the target molecule before, though it may

be a natural product that is synthesized by some biological systems. The first step in solving such a problem is to get a clear idea about the structure of the target molecule. Since organic molecules are individuated by their structures, this amounts to insisting that the synthesis problem is well defined – it has a clear goal. Once the structure of the target is settled, the synthetic chemist must come up with some way to leverage knowledge about the outcomes of lots of chemical reactions run on different (typically simpler) compounds into a strategy for making the target compound, which, presumably, no chemist has ever experimented with before. It is crucial, therefore, for the synthetic chemist to exploit some notion of structural similarity. Structural patterns identified in the target indicate which known reactions might be plausibly employed in its synthesis. Furthermore, because the structural patterns identified in the target are in a novel context, the chemist must have some way of accounting for, or anticipating, the way that the structural context influences the behavior of known reactions (characterized and understood in simpler structural contexts). This is what the structural accounts (explanations and predictions) of theoretical organic chemistry do.

One way to think about the process of coming up with a synthesis for a target compound is through the process of retrosynthetic analysis [40.19]. Retrosynthetic analysis works backwards from the target molecule, systematically investigating all of the ways that one might produce the target molecule by a known chemical reaction (these are characterized by the structural patterns on which they operate, for instance, by the functional groups that they begin with and that they produce). All of the reactants that might produce the target molecule by one of these known reactions are then subjected to the same process, generating their own lists of possible second-order reactants. This process is repeated until it generates a path terminating in compounds that can already be synthesized. Given that there are thousands of known reactions, many of which might apply to a complex target molecule, the branching array of possibilities generated by such a process – the retrosynthetic tree – is immense and must be systematically pruned into a plausible synthetic plan. (I have described this process in significantly more detail, with concrete examples, in [40.20, 21].)

The pruning of the retrosynthetic tree, following Corey's conception, takes place in stages. In the first stage, strategic pruning, the synthetic chemist analyzes the target compound in order to identify the sources of synthetic complexity in it. By identifying these sources of complexity, the chemist can focus on paths in the retrosynthetic tree that reduce synthetic complexity and that, therefore, are more likely to terminate in com-

pounds that have already been synthesized or are easy to make. Assessing the sources of synthetic complexity in a target molecule amounts to using a set of heuristic principles, grounded in both the collective experience of synthetic chemists and the theory of organic chemistry, to identify particular bonds or atoms whose structural environment will make them particularly difficult to create. The relative difficulty of dealing with these sources of complexity can also often be estimated, giving the synthetic chemist, in the end, a clear focus on branches of the retrosynthetic tree that eliminate the largest source of complexity. Though this can result in a drastic narrowing of the possible synthetic paths that need to be explored, strategic pruning must be followed up by plausibility assessment, where the possible paths removing the largest source of complexity are evaluated for their relative structural plausibility. As I characterized the retrosynthetic tree, the possible reactions that might produce a structure were characterized based on the presence of some structural feature in the target. Any reactions that might produce that product were part of the tree. However, not all of these reactions are actually plausible because, for example, the target has other structural features that would interfere with the success of that particular reaction. And even among those that are plausible, the synthetic chemists will want to decide which path or paths are most likely to work and to generate the fewest complications downstream. These assessments again depend on analyzing how a reaction, understood and characterized in some other, simpler structural context, would perform in the complex local circumstances of the target molecule. After plausibility assessment comes the final stage of synthetic design, which is optimization, where precise ordering of synthetic steps is worked through and control steps are added. These control steps are added in order to eliminate complicating factors identified by a careful structural analysis of the synthetic route. They work by adding chemical groups to synthetic intermediates in the proposed path that either eliminate the influence of complicating structural factors or promote the formation of desired products. These control groups can then be removed after they have done their job. Often the precise ordering of the synthetic path can influence which control groups are needed, and vice-versa, so the overall optimization of the synthetic route must involve both of these considerations.

This brief sketch of the process of designing a chemical synthesis has made it clear, I hope, that close structural analysis of the both the target molecule and the potential intermediates is crucial to the process. The possible reactions resulting in the target molecule (or some intermediate) guided by structural similarity to the products of known reactions is what generates

the array of potential precursors at each stage in the generation of the retrosynthetic tree. This exploration of possible reactants and reactions, and the array that it generates, all take place in the world of the model – in fact, the molecules depicted by these structural formulas may never have existed. The possible reactants must be deduced based on the reaction being considered, and this can be done by exploring what reacting structures would, upon application of the considered reaction, result in the target molecule; the reactions must be worked through backwards in the world of the model to generate potential precursors. Similarly, the strategic pruning of the array of possible reactions depends on investigating the detailed local environments of particular atoms or bonds in the model of the target molecule. Rules of thumb about the relative difficulties of producing these atomic arrangements or bonds (based in part on what the reactant structures would have to be if they were generated by certain procedures) guide the synthetic chemist to particular routes in the retrosynthetic tree. Recurring structural features identified in the local environment of the model of the target molecule provide the basis for the application of these rules of thumb, and thus for the large-scale decisions

about synthetic strategy. Both plausibility analysis and optimization depend on determining how generically characterized reactions would be likely to perform in the complex local environment of the target molecule or intermediates. Often the typical explanations (and/or predictions) of organic chemistry can be used to figure out how individual structural features would affect the reaction. But in complex environments there are often multiple relevant, and potentially competing, structural features at play. To make sensible choices about strategy in these cases, synthetic chemists can either attempt to theoretically discriminate the plausibility of potential pathways, or to modify the structure to make its behavior more predictable using control steps. All of this takes place in the world of the model, using whatever theoretical principles are applicable in that local environment, to analyze and make sensible decisions about what synthetic pathways might work in the lab. Synthetic design is thus a process that, from beginning to end, involves manipulating, exploring, deducing possible precursors and analyzing structural models. Theory can be brought to bear on the problem only through its application to, and analysis in the context of, particular structural models.

### 40.3 The Dynamics of Structure

In this last section, I want to describe two ways that structural models have contributed to the development of the research program of structural chemistry by supplying new structural concepts. In the first case, models of structure had features that were not initially recognized to be representationally significant but which, when interpreted as significant, could be used to explain anomalous results. Chemists did not abandon the structural research program when they encountered unexpected experimental results; instead, they modified their models of structure, taking features readily available in the model and attributing new representational significance to them. General features carried around in the models were appropriated in order to modify the conception of chemical structure in the face of new experimental results. In the second case, particular structures supplied foothold concepts that allowed for experimental results to be brought to bear on these newly representationally significant features of structural formulas. Particular structures are cognitively richer than general types of models or abstract theories; they have all sorts of features that might turn out to support important inferences about the target system. In this example, chemists isolated particular cases where the significance of this new structural element was clear,

used very local concepts to explain and predict in those cases, and then generalized from there. Thus features identified in particular structures were appropriated to develop and articulate the experimental consequences of this new aspect of chemical structure. Models of structure do play an important role in the dynamics of science by supplying concepts or features that can be appropriated to modify or develop a research program. Visual representations of structure, and models of structure more generally, act as incubators for the concepts essential to modifying and teasing out the experimental consequences of chemical structure.

One of the most dramatic changes in chemists' conception of structure occurred during the middle third of the twentieth century with the gradual realization that the *conformations* of molecules, and not just their bond connectivity, had a crucial role to play in understanding their physical and chemical behavior. A conformation is, roughly, any of the three-dimensional arrangements of atoms in space resulting from rotations around single bonds in a molecule. The development of the theory of conformations (typically called conformational analysis) occurred when features of structural formulas that had originally not been thought to have any representational significance, the 3-D arrangement of bonds or

its 2-D depiction, was recognized to represent something about the compounds that the formulas denote. With these new features available, new concepts were crafted to organize the phenomena and then articulated throughout the domain. Though it is somewhat artificial, in order to relate the development of conformational analysis to the themes of this paper, it can be understood to have occurred in two phases. First, the prior understanding of structural formulas had to be found to be insufficient, and the three-dimensional arrangement of bonds recognized to accommodate those insufficiencies. Second, once the three-dimensional orientation of bonds had been seen to be significant, the consequences of the newly enhanced conception of structure had to be developed and articulated.

### 40.3.1 Recognizing the Importance of Conformations

At the end of the nineteenth century, structural formulas (roughly) allowed for the generation of one distinct formula for each known, distinguishable chemical compound. The formulas used at this time included not just single bonds between adjacent atoms, but also occasional double bonds. Double bonds allow for *geometrical isomers* in which the same groups are connected to the four available positions in a double bond in two different ways. Similarly, there are two distinct ways of orienting four distinct groups around a carbon. As a result, given the number of *asymmetric double bonds* and the number of *centers of asymmetry* one could compute, using a formula due to Van't Hoff, the number of stereoisomers to be expected. This formula worked because: "It was based on the concept of *restricted* rotation about double bonds and of *free* rotation about single bonds" [40.22, p. 299]. Rotation around single bonds had to be free because otherwise one would have expected many more distinguishable isomers. In order for structural formulas to accurately map onto the results of isomer counting experiments, certain features of the models of organic molecules had to be regarded as representationally significant. For example, structural formulas had to distinguish the two distinct ways that groups can be oriented about a double bond because these represented two distinguishable compounds. At the same time, however, the experimental facts demanded that other features of the formulas not be taken to be representationally significant. The fact that there were lots of ways to produce formulas with the same bonding and orientation (differing by what we would now think of as rotations around single bonds) was explicitly not taken to be significant in the resulting structural formulas. When it came to individuating chemical compounds, the various physical models

or structural formulas that could be generated by rotations about single bonds were distinct without being different. The possibility of rotational variants was an incidental feature of the symbol system that needed to be ignored when deducing the experimental facts from the models. They were not taken to reflect significant features of the target system.

Chemical structures are not frozen in time, however, and chemists aspired to add to the array of chemical and physical features that could be explained in terms of them. Chemists knew (or thought they did) that the many distinct models producible by rotations about single bonds weren't important for the individuation of chemical compounds, and thus for isomer counting experiments, but that left it open whether these differences might be employed to explain other sorts of experimental results. In fact, given the rich array of distinctions available in the models as yet uncorrelated with differences in the compounds they depicted, these distinctions would seem to have been ripe for exploration should new experimental results force modifications of the models.

Eventually, new experimental results did force such modifications. There are at least two distinct sorts of evidence that put pressure on the idea of free rotation about single bonds. First were failed isomer counting experiments, beginning in 1922, in which chemists were able to distinguish optically active forms of (unusual) compounds where, if all rotation about single bonds had been free, there should not be any such forms. More precisely, so long as all rotational variants around single bonds were regarded as indistinguishable, structural formulas did not predict the existence of distinct optically active forms, but optically distinct forms there were. The second sort of evidence came from discrepancies between the observed and measured entropy of ethane. These discrepancies "could only be explained by a barrier to free rotation about the two methyl groups" [40.22, p. 299]. These new experimental results were accommodated, eventually, by changing the representational significance of the models. Most fundamentally, the fact that a model has lots of rotational variants was now regarded as an explanatorily significant fact. Many of those differences between models of structure that had been irrelevant became differences that could correspond to differences in the energy or stability of the represented compound. For example, the distances between the *atoms* in the model (or suggested by the structural formula) became a feature used to connect differences in structure to differences in the energy or stability of the represented compound. It is because of differences in the distance relationships between the depicted atoms that rotational variants have different energies.

By imagining the atoms of a molecular *model* or *structural formula* to be interacting (either by attraction or repulsion) in a manner that varied according to the distance between them, the chemists looking to revise earlier interpretations of *chemical structure* could explain both why the rotations of ethane would be restricted and why there might be optically distinct forms of some strategically bulky organic molecules. This required the idea of nonbonding interactions between the atoms in the compound and the addition of this idea was nontrivial, depending for its plausibility on the dawning awareness of the nature of the chemical bond. But once this idea was in place, not only could the new experimental results be explained, but the success of Van't Hoff's formula could be preserved as approximately true. Most of the time, the newly postulated nonbonding interactions would be insufficient to allow for distinct forms of chemical compounds to be isolated. Sometimes, however, such distinctions would show up in physical properties that Van't Hoff hadn't been concerned to explain (like the entropy of ethane). And occasionally, in structurally rationalizable exceptional cases, these distinctions would result in failed isomer counting experiments. Instead of there being free rotation about single bonds, now the rotation about single bonds was just substantially freer than rotation about double bonds, except in certain special circumstances.

The significance of nonbonded interactions in ethane and in dramatically rotation-restricted organic molecules suggested that such interactions would also be significant to the physical properties of organic molecules in general. Thinking in terms of such nonbonding interactions required interpreting chemists' representations of structure, including structural formulas, to be significant in new ways. However, perhaps because "there was no technique available to demonstrate the phenomenon experimentally" [40.22, p. 299] this more general significance was not systematically explored until after the Second World War. Still, by this point, the rotational variants of structural formulas or physical models had demonstrated their usefulness by explaining several different sorts of novel experimental results (entropy measurements and failed isomer counting experiments) and had therefore earned their place as representationally significant.

### 40.3.2 Using Conformations in Organic Chemistry

Though this newly significant feature of chemists' structural models had been used to explain unexpected experimental results, it had not yet been integrated into the mainstream practice of organic chemists and used to generate results of its own. This began to change when

Odd Hassel published his systematic investigations of the conformations of cyclohexane and its derivatives. Cyclohexane is an ideal experimental system for investigating the significance of conformations because, as investigation of a model will quickly show, there are only three conformations of this system (what are now called the *chair*, *boat*, and *twist-boat*), out of the infinite number that are theoretically possible, that have no angle strain (Fig. 40.3). In an earlier application of structural formulas as models ([40.10, 20], for a description of this work), chemists had shown that angle strain (or deviations from the standard tetrahedral bonding angles) was an important factor in the stability of rings. This meant that when trying to understand the behavior of cyclohexane, it was principally these three conformations that needed to be considered because all others would be energetically unfavorable. Hassel not only showed that the *chair* conformation was the most stable, but was also able to establish that the relevant nonbonded interactions were repulsive, because the chair form maximizes the distances between atoms in the ring.

Exploring a careful drawing of a chair conformation, or better yet a physical model of it, quickly reveals that there are two distinct types of bonds emanating from the carbon ring. In an obvious case of using models to introduce new conceptual distinctions, these are now called axial and equatorial bonds, according to whether they are parallel to the axis of symmetry of the molecule or in an equatorial belt around it. Furthermore, it is also clear that substituents attached axially are closer to the other atoms in cyclohexane than are substituents attached equatorially. As a result, substituted cyclohexanes generally prefer to have their substituents equatorial since this minimizes the nonbonding repulsive interactions. Hassel's work showed how the conformational preferences of cyclohexane derivatives could be rationalized using repulsive nonbonding inter-

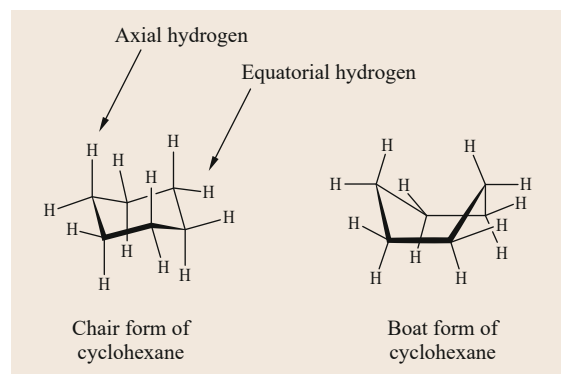


Fig. 40.3 (a) Chair form of cyclohexane. (b) Boat form of cyclohexane

actions in a way that had strong experimental support. Furthermore, he not only isolated a structural type in which the energetic implications of conformational differences were clear, but he also provided structural concepts (axial versus equatorial positions on the ring) useful in explaining the relative energies of structures of this type.

It was Barton who established the importance of conformational analysis in explaining and predicting the chemical behavior of synthetically important organic molecules. He did this by recognizing that steroids are instances of the structural type carefully studied by Hassel. The steroid nucleus consists of three cyclohexane rings fused to a five-membered ring. Because “the ring fusions of the steroid nucleus fix the conformation of the whole molecule” [40.22, p. 302] there are basically two different significant conformations of the steroid nucleus. In both of these conformations, all three of the cyclohexane rings are fixed in the *chair* form. So just as with cyclohexane itself, the significance of conformations for the behavior of steroids can be understood by considering just a few of the infinitely many possible conformations. Better still, following Hassel, the axial and equatorial substituents in the steroid nucleus can be distinguished and their relative stability rationalized in terms of repulsive non-bonded interactions.

Barton next showed how differences in the relative stability of steroids based on the conformational location of their substituents could be used to explain the chemical behavior of these molecules. Most simply, in a chemical reaction known for mechanistic reasons to result in the most stable product, one can often predict which of several candidate steroids will be preferred. Similarly, if one knows something about either the steric requirements or the geometry of the transition state, then one can often deduce which steroid will react more quickly or which product will be preferred in a reaction under kinetic control. What Barton did (originally in [40.23]) was to show that “an enormous literature of stereochemical fact” [40.22, p. 302] about steroids could be systematically and consistently interpreted using the conformational analysis of the steroid nucleus. He went through a variety of different results previously reported in the steroid literature and showed that the differences in rates or product distribution were what would be expected based on the conformational analysis of the steroid nucleus. This established by a sort of consilience of inductions that, at least in the case of steroid chemistry, conformations had an important role to play in understanding chemical behavior.

Between Hassel and Barton, not only had conformations proved themselves to be useful in explaining

significant chemical behavior, but also a set of structural circumstances (and concepts) had been articulated that allowed chemists to clearly discern the implications of conformation. With these resources in place, chemists were able to begin to apply these concepts in synthesis and experimental design. For example, once chemists understood why certain substitution patterns of the steroid nucleus were more stable than others, they could begin to exploit this knowledge in designing synthetic reactions. Barton describes how the tendency for adjacent diaxial substituents to rearrange into the more stable diequatorial form led to “a convenient route for shifting an oxygen function from one carbon atom to the adjacent carbon” [40.22, p. 304]. Similarly, because the conformation of the steroid nucleus was well known and restricted, it could be used to investigate the mechanisms of chemical reactions by effectively locking the substrate in a reaction into a particular geometry. For example, steroids were useful in establishing that “the phenomenon of neighboring group participation demands a conformational interpretation (diaxial participation)” [40.22, p. 304]. These, and other cases of application, depend on being able to recognize a set of structural circumstances in which conformational analysis is straightforward because it can be directly related back to cases that have already been successfully analyzed.

Of course, chemists were not content to apply conformational analysis just to cyclohexane and steroids. Instead, conformational analysis was articulated, from this base, along several different avenues. In the first place, it was applied to other molecules containing cyclohexane subunits, such as triterpenoids and oleanolic acid [40.22, p. 305]. Quantitative approaches were developed and this allowed for precise predictions of energy differences between conformations in these sorts of systems. Eventually, the structural limits of this approach were probed by identifying situations in which molecules with cyclohexane subunits did not behave as expected. New concepts, such as *conformational transmission* were then introduced to account for these deviations from expectation. These were refinements in the application of conformational analysis to the same basic type of system in which its clear consequences were originally discerned. Additionally, attempts were made to extrapolate the same basic approach used in analyzing cyclohexane to unsaturated six-membered rings and heterocyclic compounds. This is a case of pushing the approach into new territory. It required adapting the concepts used in the cyclohexane case to these structurally similar but importantly different new cases. New issues had to be confronted, such as how to account for the conformational implications of electron pairs. Eventually, conformations became one of the central tools

used to understand the behavior of biologically relevant molecules.

Once the rotational variants of structural models were recognized to be significant, chemists still faced the daunting task of organizing and sorting these infinite structural variations into categories that could be inferentially connected with experimental results, and eventually lead to new experiments. This was not done in a top-down way, by somehow deducing the implications of nonbonded interactions and conformations for chemical reactions. Instead, successfully doing this depended on finding a particular case where the conformational implications were clear and then generalizing and articulating from there. The concepts used to connect conformations with experiment came, initially, from considering cyclohexane. Models of cyclohexane played a crucial role in both the recognition of these concepts and their connection to experiment.

Cyclohexane was focused on because chemists already knew, from manipulation of models, that it had just a small number of strain-free conformations. This feature of cyclohexane is not shared with most other organic compounds, but it was crucial to its role in revealing the power of conformational analysis. Examination of these conformations showed that the chair

form maximized interatomic distances, which led to the conclusion that the relevant nonbonded interactions were repulsive. Additionally, inspection of the chair form led to the important distinction between axial and equatorial positions about the ring, which was subsequently linked with important energy differences between structural variants of cyclohexane (including, ultimately, steroids). These conformational features of cyclohexane are also not features shared by most molecules. The concept of an axial or an equatorial substituent simply doesn't apply in most molecules, but this concept turned out to be crucial is deducing the chemical consequences of conformations. The distinctions between conformations that were actually used in order to connect this new aspect of chemical structure with experiment were available only in concrete representations of a particular structure. Models of cyclohexane are rich with discernible differences not previously identified as significant in chemical explanations. These previously neutral features supplied the concepts that eventually got connected with experimental results. It was then by generalizing, adapting, and articulating these foothold concepts that the broad applicability and novel applications of conformations were developed.

## 40.4 Conclusion

I hope to have established that models of structure, typically in the form of structural formulas, are essential tools for chemists. They mediate between theory and phenomena, providing the platform on which theoretical principles are both recognized and applied. They also facilitate application, as seen in the use of the theory of organic chemistry in solving synthesis problems by – in addition to its role in explanation and prediction – providing for the possible reaction pathways, strategic evaluations, plausibility assessments,

and optimization that are crucial to synthetic design. Furthermore, structural models have also played important roles as sources of the concepts that chemists use to adapt their models to both theoretical and experimental developments. In sum, structural models are the keystone of the success of structural chemistry, not only because they are crucial to its theoretical content and application at any particular time, but also because of their contribution to its continued viability as a research program.

## References

- 40.1 D.M. Bailer-Jones: Tracing the development of models in the philosophy of science. In: *Model-Based Reasoning in Scientific Discovery*, ed. by L. Magnani, N. Nersessian, P. Thagard (Kluwer, Dordrecht 1999) pp. 23–40
- 40.2 M. Morgan, M. Morrison: *Models as Mediators. Perspectives on Natural and Social Science* (Cambridge Univ. Press, Cambridge 1999)
- 40.3 R. Frigg, S. Hartmann: Models in science. In: *The Stanford Encyclopedia of Philosophy*, ed. by E. N. Zalta <http://plato.stanford.edu/archives/fall2012/entries/models-science/> (Fall 2012 Edition)
- 40.4 M. Weisberg, P. Needham, R. Hendry: Philosophy of chemistry. In: *The Stanford Encyclopedia of Philosophy*, ed. by E. N. Zalta, <http://plato.stanford.edu/archives/win2011/entries/chemistry/> (Winter 2011 Edition)
- 40.5 M. Weisberg: Qualitative theory and chemical explanation, *Philos. Sci.* **71**, 1071–1081 (2004)
- 40.6 A. Woody: Putting quantum mechanics to work in chemistry: The power of diagrammatic representation, *Philos. Sci.* **67**(Supp.), S612–S627 (2000)

- 40.7 E.F. Caldin: Theories and the development of chemistry, *Br. J. Philos. Sci.* **10**, 209–222 (1959)
- 40.8 O.T. Benfey: *From Vital Force to Structural formulas* (Houghton Mifflin, Boston 1964)
- 40.9 W.H. Brock: *The Chemical Tree: A History of Chemistry* (W. W. Norton, New York 2000)
- 40.10 W. Goodwin: Structural formulas and explanation in organic chemistry, *Found. Chem.* **10**, 117–127 (2008)
- 40.11 W. Goodwin: Visual representations in science, *Philos. Sci.* **76**, 372–390 (2009)
- 40.12 W. Goodwin: How do structural formulas embody the theory of organic chemistry?, *Br. J. Philos. Sci.* **61**(3), 621–633 (2010)
- 40.13 U. Klein: *Experiments, Models, Paper Tools* (Stanford Univ. Press, Stanford 2003)
- 40.14 G.W. Wheland: *The Theory of Resonance* (Wiley, New York 1944)
- 40.15 L. Pauling: *Nature of the Chemical Bond* (Cornell Univ. Press, Ithaca 1948)
- 40.16 L. Pauling, R.B. Corey: Fundamental dimensions of polypeptide chains, *Proc. R. Soc. London. Ser. B Biol. Sci.* **141**(902), 10–20 (1953)
- 40.17 W. Goodwin: Experiments and theory in the preparative sciences, *Philos. Sci.* **79**(4), 429–447 (2012)
- 40.18 W. Goodwin: Quantum chemistry and organic theory, *Philos. Sci.* **80**(5), 1159–1169 (2013)
- 40.19 E.J. Corey, X.M. Cheng: *The Logic of Chemical Synthesis* (Wiley, New York 1989)
- 40.20 W. Goodwin: Implementation and innovation in total synthesis, *Found. Chem.* **10**, 177–186 (2008)
- 40.21 W. Goodwin: Scientific understanding and synthetic design, *Br. J. Philos. Sci.* **60**, 271–301 (2009)
- 40.22 D.H.R. Barton: *The Principles of Conformational Analysis* (1969), Nobel Lecture, [http://nobelprize.org/nobel\\_prizes/chemistry/laureates/1969/barton-lecture.html](http://nobelprize.org/nobel_prizes/chemistry/laureates/1969/barton-lecture.html)
- 40.23 D. Barton: The conformation of the steroid nucleus, *Experientia* **6**(8), 316–320 (1950)



# Models in Ge

## 41. Models in Geosciences

Alisa Bokulich, Naomi Oreskes

The geosciences include a wide spectrum of disciplines ranging from paleontology to climate science, and involve studies of a vast range of spatial and temporal scales, from the deep-time history of microbial life to the future of a system no less immense and complex than the entire Earth. Modeling is thus a central and indispensable tool across the geosciences. Here, we review both the history and current state of model-based inquiry in the geosciences. Research in these fields makes use of a wide variety of models, such as conceptual, physical, and numerical models, and more specifically cellular automata, artificial neural networks, agent-based models, coupled models, and hierarchical models. We note the increasing demands to incorporate biological and social systems into geoscience modeling, challenging the traditional boundaries of these fields. Understanding and articulating the many different sources of scientific uncertainty – and finding tools and methods to address them – has been at the forefront of most research in geoscience modeling. We discuss not only structural model uncertainties, parameter uncertainties, and solution uncertainties, but also the diverse sources of uncertainty arising from the complex nature of geoscience systems themselves. Without an examination of the geosciences, our philosophies of science and our understanding of the nature of model-based science are incomplete.

41.1	<b>What Are Geosciences?</b> .....	891
41.2	<b>Conceptual Models in the Geosciences</b> .....	892
41.3	<b>Physical Models in the Geosciences</b> ...	893
41.4	<b>Numerical Models in the Geosciences</b>	895
41.5	<b>Bringing the Social Sciences Into Geoscience Modeling</b> .....	897
41.6	<b>Testing Models: From Calibration to Validation</b> .....	898
41.6.1	Data and Models .....	898
41.6.2	Parametrization, Calibration, and Validation .....	899
41.6.3	Sensitivity Analysis and Other Model Tests .....	901
41.7	<b>Inverse Problem Modeling</b> .....	902
41.8	<b>Uncertainty in Geoscience Modeling</b> .	903
41.9	<b>Multimodel Approaches in Geosciences</b> .....	907
41.10	<b>Conclusions</b> .....	908
	<b>References</b> .....	908

### 41.1 What Are Geosciences?

The geosciences (sometimes also referred to as the Earth sciences) cover a very broad spectrum of disciplines including geology, paleontology, hydrology (the distribution and movement of water, on the surface and underground), glaciology (the study of ice and glaciers), climate science, oceanography, geophysics (the internal structure of the Earth, its gravitational and magnetic fields, plate tectonics, and volcanology), and geomorphology (how surface landscapes change over time). There is significant overlap between these different sub-

fields because the various subsystems of the Earth are not isolated from one another and are often interacting in complex ways. Usually, the geosciences are understood as ending where biological systems begin, but given, for example, the great relevance of plants for the hydrological cycle (e.g., ecohydrology) and erosion phenomena (e.g., biogeomorphology), as well as the great relevance of human activity in altering the climate, landscapes, and oceans, this division is becoming increasingly difficult to maintain [41.1].

Although the geosciences have traditionally focused on the Earth, the conceptual and disciplinary divides between studies of the Earth and studies of other planets are also breaking down. For example, the wealth of new information coming from the space program (e.g., the Mars Rovers, HiRISE images from the Mars Reconnaissance Orbiter, and images of various planets and moons from the Cassini-Huygens spacecraft and the New Horizons Pluto mission) has helped to generate the field of *planetary* geomorphology in addition to terrestrial (Earth) geomorphology. Planetary geomorphology includes the study of landscapes on not only planets, but also on moons (such as Saturn’s moon Titan, which has the largest dune field in our Solar System) and other large celestial bodies (such as the Comet 67P which was determined by the Rosetta-Philae lander module to have water).

The phenomena that geoscientists investigate are extremely complex and can span a vast range of spatial and temporal scales. Hence, idealized models play a central role in all of the geosciences. These models are used for a variety of purposes, including both prediction and explanation. They are used not only for basic scientific research (theoretical tools for advancing insight and understanding) but also for planning purposes, policy, and hazard mitigation. Models are used to fore-

cast a wide range of phenomena of human interest, such as earthquakes, volcanic eruptions, landslides, flooding, the movement of groundwater and spread of contaminants, and coastal erosion.

The geosciences are one of the most rapidly growing areas of interest in scientific modeling. This is led, in large part, by the tremendous amount of attention and resources that have been invested recently in climate modeling. Climate science is not unique, however, and many of the methodological issues found there are in fact widespread among the Earth sciences. Although, traditionally, philosophers of science have largely neglected the geosciences, leaving it a philosophical terra incognita [41.2], it is increasingly being recognized that our picture of the nature of science is inadequate if we do not take this research in the geosciences into account.

A complete review of all the relevant work in the diverse domains of the geosciences – and all the conceptual and methodological issues in modeling that arise within these different fields – is not possible in a single chapter. We provide here an overview of philosophical perspectives on this research that we hope will encourage more scholars to explore these topics further. The sections are organized primarily by the relevant conceptual and methodological issues.

## 41.2 Conceptual Models in the Geosciences

Conceptual models are the first step one takes before creating a more formal model (i. e., either a physical or numerical model). It is conceptualization of the key processes operating in the system of interest and the interactions between the components in the system. A conceptual model can simply take a narrative form or it can be an elaborate diagram. Typically, however, conceptual models can yield only qualitative predictions.

Some of the earliest models in geomorphology were conceptual models. Two historically important examples of conceptual models are Grove Karl Gilbert’s (1843–1918) *balance of forces* conceptual model and William Morris Davis’s (1850–1934) *cycle of erosion* conceptual model. In 1877, Gilbert introduced a conceptual model of a stream that appealed to physical concepts such as equilibrium, balance of forces, and work to explain the tendency of a stream to produce a uniform-grade bed. *Gilbert* describes his conceptual model as follows [41.3, p. 112]:

“Let us suppose that a stream endowed with a constant volume of water is at some point continuously supplied with as great a load as it is capable of car-

rying. For so great a distance as its velocity remains the same, it will neither corrade (downward) nor deposit, but will leave the grade of its bed unchanged. But if in its progress it reaches a place where a less declivity of bed gives a diminished velocity, its capacity for transportation will become less than the load and part of the load will be deposited. Or if in its progress it reaches a place where a greater declivity of bed gives an increased velocity, the capacity for transportation will become greater than the load and there will be corrasion of the bed. In this way a stream which has a supply of *débris* equal to its capacity, tends to build up the gentler slopes of its bed and cut away the steeper. It tends to establish a single, uniform grade.”

As *Grant et al.* note [41.4, p. 9]:

“Gilbert’s greatest and most enduring contribution to conceptual models in geomorphology [...] was the application of basic principles of energy and thermodynamics to the behavior of rivers. He did so with clarity of expression and an absence of math-

ematics that appeals directly to intuition, logic, and analog reasoning.”

Note that Gilbert’s model provides the conceptual foundation on which a numerical model, giving an equation to describe the balance of these forces, could be constructed, though he himself does not take this further step.

Another seminal conceptual model in the history of geomorphology is Davis’s cycle of erosion [41.5]. Davis was a professor of geology at Harvard University; in an 1899 article entitled *The geographical cycle*, he established a framework for thinking about modeling in geomorphology [41.6, p. 481]:

“All the varied forms of the lands are dependent upon – or, as the mathematician would say, are functions of – three variable quantities, which may be called structure, process, and time.”

The evolution of a landscape may be understood as a cycle, which begins with a *peneplain* (a low relief plain) near a base (e.g., sea) level, is followed by rapid uplift leading to a youthful stage of rugged topograph, in

which streams become established, and then a mature stage of tectonic stability in which those streams widen and gradually erode the landscape back down toward the base level. Finally, there will be an *old age* stage, involving low-relief landscapes with hills where mountains used to be. This then becomes eroded back to the peneplain stage until tectonic activity resumes and the cycle begins again. He used this idea to explain, for example, the features of the Appalachian mountains. It was a qualitative and explanatory conceptual model: it sought to explain and provide qualitative predictions for various features of a landscape.

For Davis, the conceptual model was the end point of his research; in recent years, most geoscientists have sought to quantify these sorts of processes. Thus, conceptual models can be seen as either the final product (an end in itself), or as a preliminary step in the process of creating a physical or mathematical model. In the case of mathematical models, there are two levels of modeling at which questions can be raised: Is the fundamental conceptual model adequate? And has that conceptual model been adequately represented or captured by that particular choice of mathematical equations?

## 41.3 Physical Models in the Geosciences

Until the mid-twentieth century, most conceptual models in the geosciences were realized as physical models. Physical models, also sometimes referred to as *hardware* or *table top models*, are a (usually, but not always) scaled-down version of the physical system of interest. In the geosciences, the systems of interest are typically large-scale, complex, open systems that are not amenable to experimental manipulation. A physical model allows a geoscientist to bring a version of the landscape into the laboratory, manipulate various variables in a controlled way, and explore hypothetical scenarios.

One of the central questions for geologists in the late nineteenth century was the origin of mountains (a subject known as orogenesis, from the Greek word *oros* meaning mountain). A popular orogenic theory in the nineteenth century was that mountains resulted from an overall contraction of the Earth, which was thought to be a consequence of the nebular hypothesis, first proposed by *Immanuel Kant* [41.7] and *Pierre-Simone Laplace* [41.8]. To explore this hypothesis, the Swiss geologist Alphonse Favre (1815–1890) built a physical model involving layers of clay on a piece of stretched rubber, which was then released and the resulting structures were observed. The ability of this

model to successfully reproduce some of the features of mountains led Favre to conclude that it supported the *plausibility* of the hypothesis [41.9, p. 96]. It was, what we would today call, a “how-possibly model explanation” (see Chap. 4, Sect. 4.4).

One of the great challenges for physical modeling in the geosciences, however, is that the relevant pressures, temperatures, durations, etc., of geological processes are largely beyond our reach. This limitation was recognized in the nineteenth century by the French geologist and director of the *École Nationale des Mines*, Auguste Daubrée (1814–1896), who notes (*Daubrée* [41.10, p. 5], [41.11] quoted in *Oreskes* [41.9, p. 99]),

“[T]he equipment and forces that we can set to work are always circumscribed, and they can only imitate geological phenomena at the scale [...] of our own actions.”

In order to make further advances in physical modeling in the geosciences, it was realized that the relevant forces and processes would have to be appropriately scaled in the model. The quantitative mathematical theory by which such scaling could be achieved, however, would not be developed until the work of M. King Hub-

bert (1903–1989), an American oil company geologist, in the late 1930s and 1940s.

Hubbert's work provided [41.9, p. 110]:

“the first fully quantitative treatment of the question how to choose the physical properties of materials in a model to account for the much smaller scale and time frame as compared with nature.”

Hubbert's 1945 paper begins by noting the paradox that has long perplexed the geologic sciences: How could an Earth whose surface is composed of hard, rigid rock have undergone repeated deformations as if it were composed of a plastic material, as field observations of mountains and strata suggest? He notes that this paradox is a result of failing to adequately consider the concept of physical similarity, which like geometric similarity in a map, requires that *all* the relevant physical quantities (not just lengths, but densities, forces, stresses, strengths, viscosities, etc.) bear *constant ratios* to one another [41.12, p. 1638]. He notes that when the strengths are appropriately scaled, the resulting strength of the rock on a human scale is “comparable with that of very soft mud or pancake batter” [41.12, p. 1651]. So, for example, since the elastic properties of solids depend on the strain rate, scale models that operate orders of magnitude faster than terrestrial processes need to use materials that are orders of magnitude weaker than terrestrial rocks [41.9, p. 113]. Hubbert's work on scaling not only helped explain the puzzling field observations, but also provided the key to more adequate physical modeling.

Physical models can be classified by how they do or do not scale down. At one extreme there are *life size* (1 : 1) replica models of the system of interest. Sometimes such 1 : 1 physical models are a localized study of a particular process, such as *Bagnold's* [41.13] use of a wind tunnel to study how grains of sand saltate form ripples. However, a full-scale physical model can also be an entire complex system, such as the Outdoor Streamlab at the University of Minnesota. In this full-scale model of a river segment, water and sediment flow down an artificial river system where the sediment is collected, measured, and recirculated to a sediment feeder. Although such replica models are able to avoid some of the problems arising from scaling issues (discussed below), they still involve simplifications and *laboratory effects* that can affect the reliability of the conclusions drawn for their real-world counterparts. More generally, however, many of the systems that geoscientists are interested in (e.g., mountain ranges and coastlines) are simply too large to be recreated on a 1 : 1 scale; hence, this type of physical model is typically not feasible.

Scale models are physical models that have been shrunk down according to some scale ratio (scale mod-

els can in principle be enlarged versions of their real-world counterparts, though this is not typical in the geosciences). For example, a 500 m-wide real river may be represented by a 5 m-wide scaled physical model, in which case the scale is 1 : 100. As Hubbert realized, simply shrinking a system down by some factor, however, will rarely preserve the necessary dynamical relations [41.14, p. 4]:

“A true scaled model requires perfect geometric, kinematic, and dynamic similitude, something that cannot be achieved when using the same fluid as in the real world system due to equivalent gravitational and fluid motion forces.”

Further complicating accurate scale modeling is the fact that different hydrodynamic processes are occurring at different spatial scales, and different physical effects can become dominant at those different scales too. For example, when scaling down one might substitute a fine sand for a pebbly gravel, but then cohesive forces can become dominant in the model when they are negligible in the target. These are examples of what are known as *scale effects*, when the force ratios are incomparable between the model and target. In such cases, one might need to substitute a liquid with a different viscosity or a different bed material into the model to try to overcome these scaling limitations – an example of how modelers sometimes deliberately get things more wrong in the model in order to get the conclusions to come out more right.

More often, the physical models are *distorted scale models*, where not all factors are scaled by the same ratio. The San Francisco Bay model, which is a table-top working hydraulic model of the San Francisco bay and Sacramento–San Joaquin River Delta system built by the US Army Corps of engineers, is an example of a geometrically distorted scale model, with the horizontal scale ratio being 1 : 1000, while the vertical scale ratio is only 1 : 100, and the time scale being 15 min to one day (for a philosophical discussion of this model see *Weisberg* [41.15]). Relaxing scale requirements further get what are sometimes referred to as *analog physical models*, where one reproduces certain features of a target system without satisfying the scale requirements. These are typically seen as physical systems to be investigated in their own right for what they can teach us about certain physical processes, rather than miniature versions of some specific real system [41.14, p. 5].

Physical models have their own strengths and weaknesses. The strengths, as mentioned, involve bringing a version of the system of interest into the laboratory as a closed system that is amenable to experimental manipulation and control. One does not need to have a mathematical representation of the system in order

to explore its behavior. The weaknesses, or limitations, of physical models predominantly fall into two classes: *laboratory effects* and *scale effects*. Laboratory effects are those that occur in the laboratory system but not in the real-world counterpart. These can be related to model boundary conditions (sometimes literally the wall or edge of the table) where the behavior can drastically change, unrealistic forcing conditions, or the omission of causally relevant factors in the model. Scale effects refer to problems in maintaining the correct re-

lations between variables when they are scaled down. This can lead to certain forces (e.g., cohesive forces) becoming dominant in the model that are not dominant in nature. More generally, these laboratory and scale effects are yet another example of the problem of external validity: Does the model accurately reflect the behavior of the system in the real world? This problem is pervasive among the sciences, and physical models are no more immune to it, despite dealing with the same physical stuff as their target.

## 41.4 Numerical Models in the Geosciences

Numerical models are mathematical models that represent natural systems and their interactions by means of a system of equations. These equations are typically so complex that they cannot be solved analytically, and so they have to be solved by numerical methods (such as finite difference or finite volume methods) that provide an approximate solution, or the equations need to be substituted with alternative algorithms, such as cellular automaton models. Numerical models are often implemented on a computer in a simulation that shows how the model will behave over an extended period of time, with some sort of graphical output to visualize that behavior (for a review of some of the philosophical issues in computer simulations see *Winsberg* [41.16]). This has enabled geoscientists to do something that they were generally unable (and often unwilling) to do in the past: to expand the goals of the geosciences to include forecasting and prediction as well as explanation.

In the context of the geosciences, there are many different kinds of numerical models, which can be categorized in different ways. The British geomorphologist *Kirkby* et al. [41.17], for example, distinguish the following four broad types of numerical models:

1. Black-box models
2. Process models
3. Mass–energy balance models
4. Stochastic models.

As *Kirkby* et al. explain, *black-box models* are models where “the system is treated as a single unit without any attempt to unravel its internal structure” [41.17, p. 16]. *Tucker* [41.18] gives as an example of a black-box model what is known as *Horton’s laws* of river network topology. The law predicts the average number of branching stream segments of a certain order (roughly size or width). It was discovered by Robert Horton in 1945 from purely empirical analyses of stream basins, but gives no insight into why this so-called law would

hold (it is not a law in the traditional sense, in that it does not hold universally). Black-box models are phenomenological models that involve a brute fitting to the empirical data. Although such models give no insight or understanding of the internal processes, they can be useful for making predictions.

At the other extreme of numerical modeling are *process models*, which try to describe the internal mechanisms giving rise to the empirical relations. *Tucker* explains, while [41.18, p. 687]:

“a black-box model of soil erosion would be based on regression equations obtained directly from data [...] a process model would attempt to represent the mechanics of overland flow and particle detachment.”

In between these two extremes are what *Kirkby* et al. [41.17] have called *grey-box models*, where some mechanisms may be known and included, but the rest is filled by empirical relations.

An important class of process models are landscape evolution models (LEMs). LEMs are numerical models in which the evolution of the landscape is related to the key underlying physical processes. These include, for example, the physical and chemical processes of rock weathering leading to rock disintegration and regolith production (*regolith* is a generic term referring to loose rock material, such as dust, soil, and broken rock, that covers solid rock), gravity-driven mass movement/landsliding, and water flow/run off processes (e.g., represented by the St. Venant shallow-water equations, which are a vertically integrated form of the Navier–Stokes equations). Each of these processes is represented mathematically by a *geomorphic transport function* (GTF), which get linked together to form the LEM. LEMs are often constructed as a software framework within which a variety of different component processes (represented by a particular choice of GTFs or equations), arranged in a partic-

ular configuration, can be implemented. Examples of such LEMs include the channel-hillslope integrated landscape development (CHILD) model, developed by Tucker et al. [41.19], and the cellular automaton evolutionary slope and river (CAESAR) model developed by Coulthard et al. [41.20]. These LEMs can simulate the evolution of landscapes on scales ranging from 1 to 500 km<sup>2</sup> and temporal scales ranging from days to millennia.

Often a component of LEMs, but sometimes presented as a model on their own, are *mass-balance models* (or *energy-balance models*). Mass-balance models use the fact that mass–energy is conserved to develop a continuity equation to describe the movement of mass (or energy) between different *stores*. A store could be anything ranging from water in lake, the population of a species in ecosystem, the energy stored as latent heat in an atmospheric column, the carbon mass in a tree, to the depth of soil at a point on a hillslope [41.18, p. 688]. An example of a mass-balance numerical model is a glacier model that describes the relation between ice accumulation and ablation (by melting and sublimation) at a given point of time under certain climate conditions [41.21]. Similarly, an energy-balance model in glaciology would be one that calculates the energy (heat) fluxes at the surface of the glacier that control melting and affect mass balance.

Climate science is a field of the geosciences in which both energy-balance and process numerical models have been developed to a high level of sophistication. Energy-balance models represent the climate of the Earth as a whole, without detailed information about processes or geographical variation. General circulation models (GCM) go a step further in explicitly representing atmospheric and oceanic processes. The most recent generation of climate models are Earth system models (ESM), which additionally include information about the carbon cycle and relevant biogeochemical processes. More specifically, ESMs are a composite of a large number of coupled models or *modules*, including an atmospheric general circulation model, an oceanic general circulation model, an ice dynamics model, biogeochemistry modules for land and ocean (e.g., for tracking the carbon cycle), and a software architecture or framework in which all these modules are integrated and able to communicate with each other. Developing and running GCMs and ESMs require a large number of collaborating scientists (scores to hundreds), significant supercomputing time, and millions of dollars. Because of the resource-intensive nature of such modeling projects, there are currently only a few dozen of them, and their outputs are periodically compared in intercomparison projects (e.g., coupled model intercomparison project

(CMIP5) [41.22]). (For more on coupled models and intermodel comparison projects, see Sect. 41.9 below.) At present, GCMs and ESMs typically have a spatial resolution of 100–300 km; to fill this gap at the finer level of resolution, regional climate models (RCMs) have been developed for various locations.

While the trend in climate modeling has been toward increasing the complexity of these models with ever more process modules being added, there has recently been an interesting debate about whether a fundamentally new approach to climate modeling is required (for an excellent review and assessment of the leading proposals see Katzav and Parker [41.23]). More generally the trend toward ever more complex models in the geosciences has led to what Naomi Oreskes calls the *model-complexity paradox* [41.24, p. 13]:

“The attempt to make models capture the complexities of natural systems leads to a paradox: the more we strive for realism by incorporating as many as possible of the different processes and parameters that we believe to be operating in the system, the more difficult it is for us to know if our tests of the model are meaningful.”

In opposition to this trend, many geoscience modelers have started developing what are known as *reduced complexity models*, which are motivated by the idea that complex phenomena do not always need complex models, and simpler models may be easier to understand and test. A simpler model may also be run more often, and with more different parameters, making it more amenable to sensitivity analysis (see Sect. 41.6.3 below).

In the context of geomorphology, reduced complexity modeling is often defined in contrast with what is termed *simulation* modeling (*simulation* here refers not to models that are run as a computer simulation, but rather models that try to simulate or mimic all the details of nature as closely as possible). While simulation models try to remain grounded in the fundamental laws of classical mechanics and try to represent as many of the processes operating, and in as much detail, as is computationally feasible, reduced complexity models represent a complex system with just a few simple rules formulated at a higher level of description. As physical geographers Nicholas and Quine note, emphasis added [41.25, p. 319]:

“In one sense, the classification of a model as a *reduced complexity* approach appears unnecessary since, by definition, all models represent simplifications of reality. However, in the context of fluvial geomorphology, such terminology says much about

both the central position of classical mechanics within theoretical and numerical modeling, and the role of the individual modeler in defining what constitutes an acceptable representation of the natural environment.”

One of the first successful reduced complexity models in geomorphology was a cellular automata-type model of a braided river (i. e., a river with a number of interwoven channels that shift over time) that used just two key rules [41.26]. This model was heralded as a paradigm shift in geomorphic modeling [41.27,

p. 194]. As Brad Murray, one of the proponents of this approach, argues, knowing how the many small-scale processes give rise to the large-scale variables in the phenomenon of interest is a *separate* scientific endeavor from modeling that large-scale phenomenon (Murray [41.28]; see also Werner [41.29]). Although reduced complexity models may seem like caricatures of their target systems, they can be surprisingly successful in generating realistic behaviors and providing explanatory insight (for further philosophical discussion of reduced complexity models and this case, see Bokulich [41.30] and Murray [41.31]).

## 41.5 Bringing the Social Sciences Into Geoscience Modeling

The geosciences are considered a branch of the physical sciences, being concerned with the chemistry and physics of the Earth, its history, and (more recently) its future. As such, the geosciences are typically thought of as excluding the domains of both the biological sciences and social sciences. Maintaining these artificial divisions, however, has increasingly become difficult. As Oreskes argues [41.1, p. 247]:

“Many, perhaps, most, significant topics in Earth science research today address matters that involve not only the functioning of physical systems, but the interaction of physical and social systems. Information and assumptions about human behavior, human institutions, and infrastructures, and human reactions and responses are now built into various domains of Earth scientific research, including hydrology, climate research, seismology and volcanology.”

For example, hydrological models that attempt to predict groundwater levels on the basis of physical considerations alone, can be inadequate for failing to include possible changes in human groundwater pumping activity, an external forcing function that can have dramatic effects on the physical system.

Climate science is another domain of the geosciences in which the need to incorporate the social sciences (specifically patterns and projections of human behavior involving, e.g., emission scenarios and deforestation practices) is evident. The Intergovernmental Panel on Climate Change (IPCC) has attempted to incorporate these social factors by three separate working groups, the first on the physical basis and the others on the social and policy dimensions, each issuing separate reports, released at different times. But, as Oreskes notes, the social variables are not just relevant to the

social–policy questions, but to “the work that provides the (allegedly) physical science basis as well” [41.1, p. 253].

Increasingly geoscientists are being called upon to not only use their models to predict geoscience phenomena, but also to perform risk assessments and to communicate those risks to the public. Given that geoscientists are typically not trained in risk assessment, risk policy, or public communication, the results can be troubling. Oreskes recounts the high-profile case of the 2009 earthquake in central Italy that killed 309 people, and for which six geophysicists were sentenced to six years in prison for involuntary manslaughter in connection with those deaths. Although the international scientific community expressed outrage that these seismologists were being charged with failing to predict the unpredictable, the prosecutor, as reported in *Nature* painted a different picture (Hall [41.32, p. 266]; quoted in Oreskes [41.1, p. 257]):

“‘I’m not crazy’, Picuti says. ‘I know they can’t predict earthquakes. The basis of the charges is not that they didn’t predict the earthquake. As functionaries of the state, they had certain duties imposed by law: to evaluate and characterize the risks that were present in L’Aquila.’ Part of that risk assessment, he says, should have included the density of the urban population and the known fragility of many ancient buildings in the city centre. ‘They were obligated to evaluate the degree of risk given all these factors’, he says, and they did not.”

Oreskes concludes from this case [41.1, p. 257]:

“[s]eismology in the twenty-first century, it would seem, is not just a matter of learning about earthquakes, it is also about adequately communicating what we have (and have not) learned.”

Whether it is communicating the risks revealed by geoscience models or incorporating social variables directly into geoscience models, geoscientists are under increasing pressure to find ways to model these hybrid geosocial systems.

In some areas, such as geomorphology, agent-based models (ABMs) (which are common in fields such as economics) are starting to be used. ABMs consist of a set of agents with certain characteristics, following certain rules of self-directed behavior, a set of relationships describing how agents can interact with each other, and an environment both within which, and on which, the agents can act. As *Wainwright and Millington* note [41.33, p. 842]:

“Despite an increasing recognition that human activity is currently the dominant force modifying landscapes, and that this activity has been increasing through the Holocene, there has been little integrative work to evaluate human interactions with geomorphic processes. We argue that ABMs are a useful tool for overcoming limitations of existing [...] approaches.”

These ABM models, with their simplistic representation of human behavior, however, face many challenges, including not only difficulties in integrating the different disciplinary perspectives required to model these hybrid geosocial systems, but also issues of model evaluation.

## 41.6 Testing Models: From Calibration to Validation

### 41.6.1 Data and Models

Empirical data was long assumed to be the objective and unimpeachable ground against which theories or theoretical models are judged; when theory and data clashed, it was the theory or model that was expected to bend. Beginning in the early 1960s, however, philosophers of science including *Kuhn* [41.34, pp. 133–134], *Suppes* [41.35], and *Lakatos* [41.36, pp. 128–130] began to realize that this is not always the case: sometimes it is reasonable to view the theory as correct and use it to interpret data as either reliable or faulty. In a 1962 paper called *Models of Data*, *Suppes* argued that theories or theoretical models are not compared with raw empirical data, but rather with models of the data, which are a cleaned up, organized, and processed version of the data of experience. The production of a data model can involve, among other things, *data reduction* (any data points that are due to error or noise, or what are otherwise artifacts of the experimental conditions are eliminated from consideration) and *curve fitting* (a decision about which of several possible curves compatible with the data will be drawn).

This same insight has been recognized by scientists as well. The ecological modeler *Rykiel*, for example, writes, “Data are not an infallible standard for judging model performance. Rather the model and data are two moving targets that we try to overlay one upon the other” [41.37, p. 235]. Similarly *Wainwright and Mulligan* argue that the data of measurements are an abstraction from reality depending on timing, technique, spatial distribution, scale, and density of sampling. They continue [41.33, p. 13]:

“If a model under-performs in terms of predictive or explanatory power, this can be the result of inap-

propriate sampling for parametrization or validation as much as model performance itself. It is often assumed implicitly that data represents reality better than a model does (or indeed that data is reality). Both are models and it is important to be critical of both.”

A similar point has been made by the historian *Paul Edwards* [41.38] in his book on the development of climate modeling. There he traces in detail the changing meaning of *data* in meteorology and atmospheric science, noting how the existing incomplete, inconsistent, and heterogeneous data had to be transformed into a complete and coherent global dataset, with large numbers of missing gridpoint values interpolated from computer models in a process known as “objective analysis” [41.38, p. 252]. *Edwards* further argues that even the data obtained from measuring instruments is model-laden. He notes, for example, that [41.38, pp. 282–283]:

“meteorology’s arsenal of instrumentation grew to include devices, from Doppler radar to satellites, whose raw signals could not be understood as meteorological information. Until converted – through modeling – into quantities such as temperature, pressure, and precipitation, these signals did not count as data at all.”

The importance of recognizing this model-ladenness of data is vividly illustrated in *Elizabeth Lloyd’s* [41.39] recounting of the high-profile case in which it was claimed in a US congressional hearing that data from satellites and weather balloons contradicted climate model evidence that greenhouse warming was occurring. In the end, the climate models were vindicated as more reliable than the data. *Lloyd* concludes



from this case that we need to move towards a more complex empiricist understanding of the nature of data.

The data from measurements can, for example, be skewed by the fact that measurements are local, and yet the model might require a more global value (especially when there is significant heterogeneity), or more generally that measurements can only be made at one scale, and yet have to be extrapolated to another scale. Hence, when using data to parameterize, calibrate, or validate a model (see below) it is important to be aware of the limitations of the *data model* as well, and pay attention to any biases or errors that may have been introduced during the collection and processing of that data.

In some areas of the geosciences, such as paleontology, models have even been used to correct biases in available data. For example, one aim of paleontology is to gather information about the deep-time history of biodiversity (ranging from the Cambrian explosion to the various mass extinctions) on the basis of the observed fossil record. The conditions under which fossils are formed, preserved, and revealed are not only rare, but highly contingent and uneven with respect to space, time, and type of organism. Hence, there is arguably a strong detection (or sampling) bias in the observations. While some have taken the paleodiversity curves constructed from these fossil observations as a literal description of ancient biodiversity, others have argued that observed paleodiversity is a composite pattern, representing a biological signal that is overprinted by variation in sampling effort and geological drivers that have created a nonuniform fossil record [41.40, 41]. Before any evolutionary theories can be tested against the data of the fossil record, these data need to be corrected to extract the relevant biological signal from other confounding factors. Thus, for example, “many vertebrate paleodiversity studies have relied on modeling approaches (e.g., multivariate regression models) to ‘correct’ data for uneven sampling” [41.40, p. 127]. Of course, how the data are to be properly corrected, including which models of possible drivers and sources of bias are included in the multivariate analysis yielding the corrected data, involves substantial theoretical assumptions. As *Kuhn* noted years ago, observations are not “given of experience”, but are “collected with difficulty” [41.34, p. 126].

The model-ladenness of data has led philosophers such as *Giere* to claim that “it is models almost all the way down” [41.42, p. 55] – a conclusion *Edwards* [41.38] argues is strongly supported by his historical analysis of the nature of data in meteorology and atmospheric science. Others, such as *Norton* and *Suppe*, have taken this conclusion even further, arguing that it is models *all* the way down. They write [41.43, p. 73]:

“Whether physically or computationally realized, all data collection from instruments involves modeling. Thus raw data also are models of data. Therefore, there is no important epistemological difference between raw and reduced data. The distinction is relative.”

However, saying that all data is model-laden to some degree does not imply that there is no epistemological difference, nor that all models are epistemically on par [41.44, pp. 103–104]. One of the most underdeveloped issues in this literature on data models is an analysis of what makes some data models better than others, and under what sorts of conditions data models should – or should not – be taken as more reliable than more theoretical models.

#### 41.6.2 Parametrization, Calibration, and Validation

In mathematical modeling, one can distinguish variables, which are quantities that can vary and are to be calculated as part of the modeling solution, and parameters, which are quantities used to represent intrinsic characteristics of the system and are specified external to the model by the modeler. Also specified external to the model are the boundary conditions and the initial conditions (the latter describe the values of the variables at the beginning of a model run). Whether something is a variable or parameter depends on how it is treated in a particular model. Parameters need not be constant and can also vary across space, for example, but how they vary is specified external to the model. One can further distinguish two general types of parameters: those related to characteristics of the dynamics of a process and those related to the characteristics of a specific system or location where the model is being applied [41.45, p. 7].

Sometimes parameters can be universal constants (e.g., gravitational acceleration or the latent heat of water), in which case specifying their values is relatively unproblematic (though the process by which the values of constants are initially determined is nontrivial, and as *Chang* [41.46] cogently argues, challenges arise even in so-called basic measurements, such as temperature). More typically in the geosciences, however, the value of a parameter has to be determined on the basis of complex measurements, and even an idealization or averaging of those measurements (such as in the case of the parameter for bed roughness of a stream bed). The process by which input parameters are initially chosen has not been well studied, and is greatly in need of a better understanding. What has been the subject of considerable attention is the problem of calibration: the

adjustment of model parameters in response to inadequate model performance.

In an ideal world, modelers would build a model based on physical principles and the equations that represent them, and then, with the use of appropriate input parameters for physical variables (like temperature, pressure, permeability, equilibrium constants, etc.), build a numerical simulation that accurately reflects the system under analysis. But most models do not do this: for a variety of reasons the match between the model output and available empirical information is often quite poor [41.47]. Therefore, modelers *calibrate* their models: they adjust the input parameters until the fit of the model to available information is improved to a level that they consider acceptable.

There are several concerns that can be raised about this process. One is that parameterized models are nonunique, and there is no way to know which particular set of parameterizations (if any) is the so-called right one; many different parameterizations may produce a given output. (This may be understood as a variation on the theme of underdetermination, discussed further below.) As hydrologist *Beven* notes [41.45, p. 7]:

“parameters are usually calibrated on the basis of very limited measurements, by extrapolation from applications at other sites, or by inference from a comparison of model outputs and observed responses at the site of interest.”

Moreover, because of the variability and uniqueness of many complex systems, parameter values extrapolated from one site may not be appropriate for another. Even if one restricts oneself to a given site, a model calibrated for one purpose (e.g., predicting peak runoff) may be predictively useless for another purpose (e.g., predicting total runoff) [41.33, p. 15]. Indeed, if the chosen parameterization is not an accurate representation of the physical system under consideration, it is likely that the model will not perform reliably when used for other purposes. This helps to explain the observation that many calibrated models fail, not only when used for purposes other than that for which they were calibrated, but sometimes even when used for their intended purposes [41.48].

Once a model has been built and calibrated, many modelers engage in an activity they call model validation, by which they normally mean the testing of the model against available data to determine whether the model is adequate for the purpose in question. Many geoscientists acknowledge that the use of the term *validation* should not be taken to imply that the model is true or correct, but rather only that “a model is acceptable for its intended use because it meets specified

performance requirements” [41.37, p. 229]. *Rykiel* thus argues that before validation can be undertaken, the following must be specified:

- a) The purpose of the model
- b) The performance criteria
- c) The context of the model.

However, many so-called validated models have failed even in their intended use. For example, in a 2001 study, *Oreskes* and *Belitz* showed that many hydrological models fail because of unanticipated changes in the forcing functions of the systems they represent. More broadly, validated models may fail for the following reasons [41.9, p. 119]:

1. Systems may have emergent properties not evident on smaller scales.
2. Small errors that do not impact the fit of the model with the observed data may nonetheless accumulate over time and space to compromise the fit of the model in the long run.
3. Models that predict long-term behavior may not anticipate changes in boundary conditions or forcing functions that can radically alter the system’s behavior.

The idea that a model can be validated has been critiqued on both semantic and epistemic grounds. Semantically, *Oreskes* et al. have noted that the terminology of *validation* implies that the model is *valid* – and thus serves as a claim about the legitimacy or accuracy – a claim that, as already suggested above, cannot be sustained philosophically and is often disproved in practice [41.47, 49]. Hence, a better term than model validation might be *model evaluation*. Even with this change in terminology, however, epistemological challenges remain. In many cases, the available empirical data (e.g., historic temperature records) have already been used to build the model, and therefore cannot also be used to test it without invoking circular reasoning. Some modelers attempt to avoid this circularity by calibrating and validating the model against different historical time periods, with respect to different variables, or even different entities and organisms.

Paleontologists, for example, use biomechanical models to try to answer functional questions about extinct animals based on the structures found in the fossil record (which is a subtle and difficult process, see e.g., [41.50]). These biomechanical models, which are used to make predictions about paleospecies, are validated or tested against data for present-day species. More specifically, *Hutchinson* et al. have used such models to determine how fast large theropod dinosaurs, such as *Tyrannosaurus rex*, could run. They write [41.51, p. 1018]:

“The model’s predictions are validated for living alligators and chickens [...]. [m]odels show that in order to run quickly, an adult Tyrannosaurus would have needed an unreasonably large mass of extensor muscle.”

Such an approach may work in cases where very large amounts of data are available, or where there are clearly distinct domains that may be enlisted. In many areas of the geosciences, however, data is scant and all available data need to be used in the initial construction of the model.

### 41.6.3 Sensitivity Analysis and Other Model Tests

Irrespective of the difficulties of model construction and calibration, models can be highly effective in helping to identify the *relative* importance of variables, through techniques such as sensitivity analyses. Sensitivity analysis – also known (inversely) as robustness analysis – is the process of determining how changes in model input parameters affect the magnitude of changes in the output (for philosophical discussions of robustness analyses see, e.g., *Weisberg* [41.52] or *Calcott* [41.53]; for a comprehensive, technical introduction to sensitivity analysis in a variety of domains see *Saltelli et al.* [41.54]). For example, in the context of the paleontology research on models of *Tyrannosaurus rex* introduced above, *Hutchinson* writes [41.55, p. 116]:

“Because any model incorporates assumptions about unknown parameters, those assumptions need to be explicitly stated and their influences on model predictions need to be quantified by sensitivity analysis [...]. In many models this can be determined by varying one parameter at a time between minimal and maximal values (e.g., crouched and columnar limb poses) and evaluating the changes in model output (e.g., the required leg muscle mass).”

Varying one parameter at a time is known as a *local* sensitivity analysis. However, for some sorts of systems (especially systems in which nonlinearities and thresholds operate), a complicating factor is that model sensitivity to a parameter can also depend on the values of the other model parameters [41.56, p. 141] and [41.33, p. 18]. Hence, in these latter cases, one needs to perform what is known as a *global* sensitivity analysis, where all the parameters are varied simultaneously to assess how their interactions might affect model output [41.57].

Sensitivity analysis is used in nearly all domains of modeling, and it can be an important guide to data

collection: alerting the scientific community to where additional or better empirical information is most likely to make a difference. That is to say, sensitivity analyses can reveal which parameters are most important in a model (and hence should be targeted for additional data collection) and which parameters are relatively unimportant or even negligible. It may thus suggest parameters that should be omitted, which can save on computational time. Sensitivity analyses can also help determine whether a model might be overparameterized, which involves a kind of overfitting to the data that occurs when too many parameters are included and fixed.

Model testing can involve a wide spectrum of different techniques, ranging from subjective expert judgments to sophisticated statistical techniques. *Rykiel* [41.37] has assembled a list of 13 different procedures, which he calls *validation procedures*. However, given the concerns raised above about the term validation and the heterogeneity of the procedures collected in his list, the broader rubric of *model tests* is arguably more appropriate. *Rykiel*’s list is as follows [41.37, pp. 235–237]:

1. Face validity, where experts are asked if the model and its behavior are reasonable.
2. Turing-test validity, where experts assess whether they can distinguish between system and model outputs.
3. Visual validation, where visual outputs of model are (subjectively) assessed for visual goodness of fit.
4. Inter-model comparisons.
5. Internal validity of model.
6. Qualitative validation: the ability to produce proper relationships among model variables and their dynamic behavior (not quantitative values).
7. Historical data validation, where a part of the historical data is used to build the model and a part is used to validate it.
8. Extreme conditions tests, where model behavior is checked for unlikely conditions.
9. Traces: the behavior of certain variables is traced through the model to see if it remains reasonable at intermediate stages.
10. Sensitivity analyses: the parameters to which the model is sensitive are assessed against the parameters to which the system is or is not sensitive.
11. Multistage validation: validation at certain critical stages throughout the model-building process.
12. Predictive validation: model predictions are compared to system behavior.
13. Statistical validation: statistical properties of model output are evaluated and errors are statistically analyzed.

Although, as noted before, the term validation is inappropriate and this heterogeneous list could be usefully organized into different categories, it nonetheless provides a good sense of the broad spectrum of techniques

that modelers deploy in testing and evaluating their models. Each of the procedure on this list can play an important role in the modeling process and is arguably worthy of further philosophical and methodological reflection.

## 41.7 Inverse Problem Modeling

One of the central tasks of geophysics is to determine the properties of the interior structure of the Earth on the basis of measurements made at the surface. The primary method by which this is done is known as *inverse problem modeling*. Most broadly, an inverse problem is defined as that of reconstructing the parameters of a system or model based on the data it produces; in other words, one starts with a set of observational data and then tries to reason back to the causal structure that might have produced it. The inverse problem is contrasted with the *forward problem*, which involves starting with a known model and then calculating what observations or data that model structure will produce. Inverse problems are found across the sciences, such as in finding the quantum potential in the Schrödinger equation on the basis of scattering experiments, diagnostic imaging in medicine using X-ray computer assisted tomography, or, most relevantly here, determining information about the interior structure of the Earth on the basis of travel-time data of waves (e.g., earthquakes). Indeed, the first methods for solving inverse problems were developed in the context of seismology by a German mathematical physicist Gustav Herglotz (1881–1953) and the geophysicist Emil Wiechert (1861–1928).

A fundamental challenge for inverse modeling methods is the problem of underdetermination [41.58, p. 120]:

“[T]he model one aims to determine is a continuous function of the space variables. This means the model has infinitely many degrees of freedom. However, in a realistic experiment the amount of data that can be used for the determination of the model is usually finite. A simple count of variables shows that the data cannot carry sufficient information to determine the model uniquely.”

In other words, the solution to the inverse problem is not unique: there are many different models that can account for any given set of data equally well. This is true for both linear and nonlinear inverse problems [41.59].

One method for trying to constrain this underdetermination is known as the model-based inversion

approach, which involves introducing a second, intermediary model known as the *estimated* or *assumed* model [41.60, p. 626]. The estimated model is used in the forward direction to generate *synthetic* data, which is then compared with the observational data. On the basis of the discrepancy between the two datasets, the estimated model is modified and the synthetic data it produces is again compared in an iterative optimization process. As *Snieder* and *Trampert* note, however [41.58, p. 121]:

“There are two reasons why the estimated model differs from the true model. The first reason is the nonuniqueness of the inverse problem that causes several (usually infinitely many) models to fit the data [...] The second reason is that real data [...] are always contaminated with errors and the estimated model is therefore affected by these errors as well.”

In other words, one must also be aware of errors arising from the data model (as discussed earlier). Different modeling approaches for dealing with inverse problems in geophysics have been developed, such as the use of artificial neural network (ANN) models (see, e.g., *Sandham* and *Hamilton* [41.61] for a brief review).

Recently, a number of philosophers of science have highlighted the philosophical implications of the underdetermination one finds in geophysical inverse problems. *Belot* [41.62], for example, argues that this “down to earth underdetermination” shifts the burden of proof in the realism–antirealism debate by showing that a radical underdetermination of theory by (all possible) data is not just possible, but actual, and likely widespread in the geosciences (and elsewhere). *Miyake* similarly calls attention to the problem of underdetermination in these Earth models and notes that there are additional sources of uncertainty that are not even considered in the setting up of the inverse problem [41.63]. He argues that thinking of these Earth models as a case of what philosophers [41.64] call *model-based measurement* is important for understanding the epistemology of seismology.

## 41.8 Uncertainty in Geoscience Modeling

Geoscientists have paid considerable attention to the problem of model uncertainty and sources of error, but many (if not all) of the sources of uncertainty they identify are not unique to the geosciences. There are different ways in which one can construct a taxonomy of the sources of uncertainty in modeling. One can, for example, organize the sources of uncertainty by the relevant stage in the modeling process. Here, one can group the various uncertainties into the following three categories:

1. Structural model uncertainties
2. Parameter uncertainties
3. Solution uncertainties.

Alternatively, one can also organize the sources of uncertainty in modeling on the basis of various complexities that arise for the sort of systems one is trying to model. This latter approach is taken by geomorphologist *Stanley Schumm* [41.65], who organizes the sources of uncertainty into the following three categories:

1. Problems of scale and place
2. Problems of cause and process
3. Problems of system response.

Each of these ways of thinking about sources of uncertainty in modeling serves to highlight a different set of philosophical and methodological issues.

Uncertainties can be identified at each step of the modeling process. During the construction phase of the model there are a number of uncertainties that can be grouped together under the broad rubric of *structural model uncertainties*. In this category, there are what are termed *closure uncertainties*, which involve uncertainties about which processes are to be included or not included in the model [41.66, p. 291]. There can be uncertainties regarding both which processes are in fact operating in the target system (some processes might be unknown) and which of the processes known to be operating are in fact important to include (we may know that a process is operating, but not think it is relevant). Sometimes whether a process is important, however, depends on what other processes are included in the model, as well as on other factors, such as the relevant spatiotemporal scale over which the model will be applied. As an example of this type of structural model (closure) uncertainty, *O'Reilly et al.* [41.67] discuss the case of early attempts to model stratospheric ozone depletion (that resulted in the unexpected *ozone hole* in the Antarctic, which was discovered in 1985). They write [41.67, p. 731]:

“[B]efore the ozone hole discovery led scientists to rethink their conceptual models, ozone assessments had not considered such multiphase reactions [i. e., heterogeneous chemical reactions] to be important. At the time, gas-phase atmospheric chemistry was much better understood than multiphase chemistry, and heterogeneous reactions were seen as arcane and generally unimportant in atmospheric processes.”

Because these chemical processes were not well understood scientifically and were not recognized as important to this phenomenon, they were left out of the model, leading to a drastic underprediction of the rate at which ozone depletion would take place. More generally, as *Oreskes* and *Belitz* have noted, when modelers lack reliable information about known or suspected processes, they may simply leave out those processes entirely, which effectively amounts to assigning them a value of zero [41.48, 67]. Such closure uncertainties in modeling can thus lead to significant errors.

Second, there are *process uncertainties*, which are concerned with how those processes should be represented mathematically in the model. For many processes in the geosciences, there is no consensus on the right way to represent a given process mathematically, and different representations may be more or less appropriate for different applications. For example, there are different ways that turbulence can be represented in models of river flow, from the greatly simplified to the highly complex [41.66, p. 291].

Third, there are what are more narrowly called *structural uncertainties*; these are uncertainties in the various ways the processes can be linked together and represented in the model. Included in this category are uncertainties associated with whether a component is taken to be active (allowed to evolve as dictated by the model) or passive (e. g., treated as a fixed boundary condition). *Lane* [41.66, p. 291] gives the example of the different ways the ocean can be treated in global climate models: because of water's high specific heat capacity, the ocean responds slowly to atmospheric changes; hence, if used on short enough time scales, the modeler can represent the ocean as a passive contributor to atmospheric processes (as a source of heat and moisture, but not one that in turn responds to atmospheric processes). *Parker* [41.68] also discusses structural uncertainty in climate modeling, with regard to the choice of model equations.

Structural model uncertainties can give rise to *structural model error*, which *Frigg et al.* [41.69] define broadly as a discrepancy between the model dynamics

and target system dynamics. They demonstrate that in a nonlinear model, even a small structural model error can lead to divergent outcomes more drastic than those due to the sensitive dependence on initial conditions characteristic of chaotic systems. In analogy with the well-known butterfly effect, they (following Thompson [41.70]) call this the *hawkmoth effect*. They conclude that the structural model error in a nonlinear model “is a poison pill . . . operational probability forecasts are therefore unreliable as a guide to rational action if interpreted as providing the probability of various outcomes” [41.69, p. 57]. Nonetheless, they note that such models may still be useful for generating insight and understanding.

In addition to these three types of structural model uncertainty (closure, process, and structure uncertainties), another significant source of uncertainty is *parameter uncertainty*. As discussed earlier, models contain both variables (whose values are determined by the model itself) and parameters (whose values must be specified externally by the modeler). In the global circulation or ESMs of climate science, parameters are used, for example, in representations of unresolved processes (such as cloud systems or ocean eddies) that are on a finer-grained scale than that on which the model operates. Ideally, the value of a parameter is determined directly by field measurements, but often this is not possible. In many cases, the parameter is either prohibitively difficult to measure or has no simple field equivalent. The parameters then need to be estimated or calculated on the basis of other models (e.g., as detailed by Edwards [41.38] in his discussion of parameters in meteorology and atmospheric science). Beven [41.45, p. 8] gives the example of the parameter representing soil hydraulic conductivity in hydrology. Measurements of soil hydraulic conductivity are typically made on soil samples in a small area, but are known to exhibit order of magnitude variability over even short distances. Often, however, the model will require a value of hydraulic conductivity over a much larger spatial scale (e.g., the whole catchment area). Hence, substantial uncertainties can arise as one tries to determine an *effective* value for the parameter.

Parameters can also take on different values than their real-world counterparts during the process of calibration or optimization. An example is the bed roughness parameter, which is used to represent the grain size of a river bed affecting the friction and turbulence of the flow. As Odoni and Lane note [41.71, p. 169]:

“it is common to have to increase this quite significantly at tributary junctions, to values much greater than might be suggested by [...] the bed grain size. In this case there is a good justifica-

tion for it, as one-dimensional models represent not only bed roughness effects but also two- and three-dimensional flow processes and turbulence.”

In other words, the bed roughness parameter in the model is used to capture not just the bed roughness, but other effects that act like bed roughness on the behavior of the flow. This is another example of what was earlier called *getting things more wrong in order to get them more right*. More generally, parameter values determined for one model may be calibrated for that particular model structure, and hence not be independent of that model structure or even different discretizations or numerical algorithms of that model structure, and therefore are not transferable to other models without additional error [41.45, p. 8]. Hence, one must be aware of the problem of *parameter incommensurability*, where parameters that share the same name might in fact “mean” different things [41.45, p. 8].

Although they are not strictly speaking parameters, one can also include under this umbrella category uncertainties in the initial conditions and the boundary conditions, which also need to be specified externally by the modeler in order to operate the model. Examples include [41.71, p. 169]:

“the geometry of the problem (e.g., the morphology of the river and floodplain system that is being used to drive the model) or boundary conditions (e.g., the flux of nutrients to a lake in a eutrophication model).”

In order to integrate a model forward in time, one needs to first input the current state of the system as initial conditions. Not only can there be uncertainties in the current state of the system, but also some chaotic models will be very sensitive to such errors in the initial conditions.

The final category of model uncertainties is *solution uncertainties*. Once the model equations are set up, the parameters fixed, and the initial and boundary conditions are specified, the next step is to solve or run the model. Often in geoscience modeling, the governing equations are nonlinear partial differential equations that do not have general analytic solutions. In such cases, one must resort to various discretization or numerical approximation algorithms (e.g., finite difference methods, finite element methods, boundary element methods, etc.) to obtain solutions, which will not be exact (though they can often be benchmarked against analytic solutions). There can also be uncertainties introduced by the way the algorithm is implemented on a computer for a simulation model. Beven notes [41.45, p. 6]:

“[D]ifferent implementations will, of course, give different predictions depending on the coding and degree of approximation. [The] computer code [...] represents a further level of approximation to the processes of the real system.”

In implementing a model on a computer, decisions must be made about the appropriate choice of time steps and spatial discretizations, and these and other solution uncertainties can lead to further sources of error.

In his book *To interpret the Earth: Ten ways to be wrong*, Schumm identifies 10 sources of uncertainty, which he organizes into the three categories of problems of scale and place, problems of cause and process, and problems of system response [41.65]. The first source of uncertainty concerns *time*. Compared to the long-time history over which Earth’s landscapes evolve, the time scale of human observation is extremely short. There can be short-term patterns in geoscience phenomena that are very different from the long-term pattern one is trying to predict or explain; hence, extrapolations from short-term observations may not be reliable (e.g., the short-term wind direction you observe may not be indicative of the prevailing long-term wind direction that predominantly shapes the landscape). Also, different features of a landscape (and the corresponding different processes) can become salient as different time scales are considered. The processes that are most relevant on a short-time scale (such as storm events) may be insignificant on a long-time scale, as well as the reverse (e.g., uplift phenomena are negligible over the short term, but are such stuff as the Himalayas are made of over the long term). Hence, inadequate attention to these issues of time, both in the construction and application of the model, can be a significant source of uncertainty. The second source of uncertainty, *space*, is analogous to these problems of time. For example, to understand how water moves through the ground on a small spatial scale, the type of soil or rock (e.g., its porousness) might be most relevant to model, while on a large scale, the overall topology of the landscape (e.g., whether it is on a steep slope) and whether it has large-scale rills (cracks or channels) might be more relevant. The third source of uncertainty Schumm calls *location*, which relates to the uniqueness of geomorphic systems (e.g., there is a sense in which no two rivers are exactly the same, and hence models developed for one location, might not be applicable to other locations).

In the next cluster, Schumm identifies *convergence* as a fourth source of uncertainty. Convergence is the idea that different processes or causes can produce similar effects. For example, sinuous rills on the Moon look like dried river beds formed by flowing water, but were later concluded to be the result of collapsed lava

tubes [41.65, p. 59]. Hence, one needs to be careful in inferring cause from effect, and in drawing an analogy from the causes of an effect at one location to the causes of a very similar effect at another location. The fifth source of uncertainty, *divergence*, is the opposite of convergence: the same cause can produce different effects. Schumm gives the example of glacio-eustasy, or the change of sea levels due to the melting of glaciers and ice sheets. He explains [41.65, p. 64]:

“With the melting of the Pleistocene continental ice sheets the assumption is that a global sea-level rise will submerge all coastlines. However, the results are quite variable [...] [a]s a result of isostatic uplift following melting of the continental ice sheets.”

Isostatic uplift refers to the rebounding or rise of land masses that were depressed under the massive weight of the ice sheets (this rebound is still ongoing and averages at the rate of a centimeter per year: see, e.g., *Sella et al.* [41.72]). In other words, the melting of glaciers and icesheets can cause sea levels both to rise and to fall (depending on the location): one cause, two different (and opposite) effects.

The sixth source of uncertainty Schumm identifies is what he calls *efficiency*, which he identifies with the assumption that the more energy expended, the greater the response or work done. He notes that this will not generally be the case [41.65, p. 66]:

“When more than one variable is acting or when a change of the independent variable, such as precipitation, has two different effects, for example, increased runoff and increased vegetation density, there may be a peak of efficiency at an intermediate condition.”

He gives as an example the rate of abrasion of a rock by blown sand, which has a maximum abrasion efficiency at relatively low rates of sand feed (presumably due to an interference of rebounding particles with incoming particles).

The seventh source of uncertainty he identifies is *multiplicity*, which is the idea that there are often multiple causes operating in coordination to produce a phenomenon, and hence one should adopt a *multiple explanation approach*. This concept originated in the work of the American geologist Thomas C. Chamberlin (1843–1928), specifically in his method of multiple working hypotheses, a method which he urged was beneficial not only to scientific investigation, but also to education and citizenship. In his 1890 article introducing this method he considers the example of explaining the origin of the Great Lake Basins. *Chamberlin* writes [41.73, p. 94]:

“It is practically demonstrable that these basins were river-valleys antecedent to the glacial incursion, and that they owe their origin in part to the pre-existence of those valleys and to the blocking-up of their outlets [. . .]. So, again, it is demonstrable that they were occupied by great lobes of ice, which excavated them to a marked degree, and therefore the theory of glacial excavation finds support [. . .]. I think it is furthermore demonstrable that the earth’s crust beneath these basins was flexed downward, and that they owe [. . .] their origin to crustal deformation.”

What might initially appear to be a scientific controversy involving rival hypotheses or competing explanations, in fact turns out to be a case where each hypothesis correctly has part of the story. Chamberlin concludes that one benefit of considering diverse explanations for observed phenomena is that it forces the geologist to move beyond hasty or simplistic explanations, and instead to consider the possibility that more than one relevant process has been involved. (For a philosophical discussion of the method of multiple hypotheses in the case of plate tectonics, see *Rachel Laudan* [41.74].)

An example of this from paleontology is the long-standing debate about the cause of the Cretaceous (K–T) mass extinction (in which 70% of all species, including all the (nonavian) dinosaurs, went extinct). The favored explanation of this extinction event is the impact hypothesis: that the extinction was caused by a large comet or asteroid that hit Earth near present-day Chicxulub, Mexico. While the fact that this impact occurred is not in doubt, some scientists question whether the impact hypothesis can explain the gradual and step-wise extinction pattern that is observed in the fossil record. They favor instead an explanation that appeals to massive volcanism and climate change, which was already underway. While often viewed as rivals, these two explanations might be complementary [41.75]. *Schumm* concludes, “if there is more than one cause of a phenomenon, unless all are comprehended, extrapolation will be weak and composite explanations are needed” [41.65, pp. 74–75]. (For a more general philosophical discussion of explanation in the Earth sciences, including a discussion of the explanation of the K–T extinction, see *Cleland* [41.76].)

The final three sources of uncertainty *Schumm* identifies are *singularity*, the idea that landforms, though also having many commonalities, have features that make them unique, and hence respond to changes in slightly different ways or at different rates; *sensitivity*, the idea that small perturbations to a system can have significant effects, especially when a system involves

either internal or external thresholds; and the *complexity* of geomorphic systems, which means they have numerous interconnected parts interacting in typically nonlinear ways. An example of an important threshold in the geosciences is the velocity at which a sediment particle of a given size is set in motion by a particular fluid (e.g., water or wind). This is an example of an extrinsic threshold involving changes in an external variable. There can, however, also be intrinsic thresholds in which there is an abrupt change in a system without there being a corresponding change in an external variable. For example, under constant weathering conditions the strength of materials can be weakened until there is an abrupt adjustment of the system (such as a landslide). Another example of an intrinsic threshold is when a bend or loop in a meandering river will suddenly be cut off by the formation of a new channel. More generally, geomorphic systems often exhibit what are called *autogenic behaviors*, in which there can be a sudden and pronounced change in the system’s behavior or characteristics, not due to an external cause, but rather due to internal feedbacks in the system, in which gradual changes can result in sudden, threshold-like responses (for a discussion see *Murray et al.* [41.77]; for an example of an autogenic behavior discovered in the St. Anthony’s Falls physical model discussed earlier, see *Paola et al.* [41.78]). *Schumm* concludes [41.65, p. 84]:

“The recognition of sensitive threshold conditions appears to be essential in order that reasonable explanations and extrapolations can be made in geomorphology, soil science, sedimentology and stratigraphy, and many environmental and ecosystem areas.”

So far we have reviewed five sources of uncertainty arising during stages of the modeling process and 10 sources of uncertainty arising from the complexity of geoscience systems. A further complication arises from the fact that even models with these sorts of errors can generate predictions that agree reasonably well with observations – a case of getting the right answer for the wrong reason. Hence, on pain of committing the fallacy of affirming the consequent, one cannot deductively conclude that one’s model is right, just because it produces predictions that match observations. More generally, this is related to the fact that more than one model or theory can account for a given set of observations: the data underdetermine the model or theory choice. In the philosophical literature this is known as the problem of underdetermination (e.g., see *Duhem* [41.79], or for contemporary discussion, see *Stanford* [41.80]; for a philosophical discussion of underdetermination in the Earth sciences see *Kleinhaus*



et al. [41.2]). In the geoscience literature the problem of underdetermination is sometimes referred to as the problem of nonuniqueness, or *equifinality* [41.81]. *Beven* and *Freer* write [41.82, p. 11]:

“It may be endemic to mechanistic modeling of complex environmental systems that there are many different model structures and many different parameter sets within a chosen model structure that may be [...] acceptable in reproducing the observed behavior of that system. This has been called the equifinality concept.”

In other words, the data are not sufficient to uniquely pick out a model structure or parameter set. (A similar sort of equifinality was seen in the nonuniqueness of inverse problems discussed earlier.) Moreover, the acceptable parameter sets may be scattered throughout parameter space (i. e., not localized around some optimum parameter set). This problem of equifinality is not

just hypothesized, but has been demonstrated in computer simulations, which are now cheap and efficient enough to allow explorations of the parameter space of models of a variety of geoscience systems.

The problem of equifinality has led *Beven* et al. to develop a method to deal with uncertainty that they call the generalized likelihood uncertainty estimation (GLUE) methodology [41.83]. GLUE involves a kind of Monte Carlo method with a random sampling of the space of possible model–parameter combinations, in which each possible set of parameters is assigned a likelihood function (assessing the fit between model predictions and observations). The idea is not to pick one *best* model–parameter set, but rather to take into account the predictions of all acceptable models (models not ruled out by current data or knowledge), weighted by their relative likelihood or acceptability, in something like a Bayesian averaging of models and predictions. (For a recent review and discussion of objections to the GLUE methodology see *Beven* and *Binley* [41.84].)

## 41.9 Multimodel Approaches in Geosciences

The GLUE methodology is just one of several different approaches that try to use multiple models in concert to reduce uncertainty. The GLUE methodology requires a large number of runs to adequately explore the parameter space. However, this is not typically feasible in computationally intensive models. An alternative approach that can be used with more complex models is the *metamodel* approach (for a review see *Kleijnen* [41.85]). A metamodel is a simplified surrogate model that is abstracted from the primary model and used to aid in the exploration of the primary model and its parameter space. While metamodels have long been used in engineering research, they have only recently started to be applied to models in the geosciences.

*Odoni* [41.86], for example, has applied the metamodel approach to the study of a landscape evolution model (LEM) developed by *Slingerland* and *Tucker* [41.87] known as GOLEM (where GO stands for geomorphic-orogenic). GOLEM has been used, for example, to model the evolution of a catchment landscape of the Oregon Coast Range around the headwaters of the Smith River over a period of 100 000 years. In order to understand how equifinality manifests itself in GOLEM, *Odoni* selected 10 parameters (related to mass movement, channel formation, fluvial erosion, and weathering processes) to vary over a range of values that was determined to be consistent with the location based on published data and calibration. The model outputs used to describe the landscape at 100 000

years include sediment yield, drainage density, sediment delivery ratio, and a topographic metric. Rather than trying to solve the full GOLEM model for the immense number of possible parameter value combinations, *Odoni* derived a metamodel, or set of regression equations, that described each model output as a function of the GOLEM parameters. As he explains, “The parameter space is then sampled rapidly and densely ( $> 1 \times 10^6$  times), using each metamodel to predict GOLEM’s output at each sample point” [41.86, p. i]. In this way metamodels yield a clearer picture of what drives model output (leading to a possible further simplification of the model) and an understanding of where equifinality may be lurking. It is important to note that this equifinality is not just an abstract cooked-up possibility, but a genuine, wide-spread practical problem, making it yet another example of what *Belot* termed down-to-earth *underdetermination*.

More common than both the GLUE and metamodel approaches are classic *intermodel comparison* projects. The most well known here are the large-scale, multi-phase intercomparison projects used by the IPCC in their assessments. The most recent coupled model intercomparison project (CMIP5), for example, compares the predictions of dozens of climate models running the same set of scenarios. The aim of such multimodel ensembles is to “sample uncertainties in emission scenarios, model uncertainty and initial condition uncertainty, and provide a basis to estimate projection uncertain-

ties” [41.88, p. 369]. *Lloyd* has emphasized the strength of such multimodel approaches, arguing that it is “a version of reasoning from variety of evidence, enabling this robustness to be a confirmatory virtue” [41.89, p. 971].

The proper assessment of such intermodel comparisons for robustness and uncertainty reduction involves some subtleties, however (see, e.g., *Parker* [41.90, 91]; *Lenhard* and *Winsberg* [41.92]). Models can, for example, agree because they share some common model structure, rather than indicating model accuracy. As *Masson* and *Knutti* explain [41.93, p. 1]:

“All models of course contain common elements (e.g., the equations of motion) because they describe the same system, and they produce similar results. But if they make the same simplifications in parameterizing unresolved process, use numerical schemes with similar problems, or even share components or parts thereof (e.g., a land surface model),

then their deviations from the true system or other models will be similar.”

In such cases an agreement among climate models does not indicate that modelers are on the right track. It remains unclear how best to conceptualize and assess model independence [41.23, p. 485]. More generally, the spread of an ensemble of models is often taken to approximate the uncertainty in our predictions; however, as *Knutti* et al. [41.94] have argued, these are *ensembles of opportunity*, not systematic explorations of model or parameter space. They suggest a number of ways forward, including having a larger diversity of models to help find constraints valid across structurally different models, and developing new statistical methods for incorporating structural model uncertainty [41.94, p. 2755]. There are many other multimodel approaches used in the geosciences, including *coupled models* and *hierarchical modeling*.

## 41.10 Conclusions

The geosciences provide a rich and fruitful context in which to explore methodological issues in scientific modeling. The problem of understanding and articulating scientific uncertainty has particularly come to the fore in these fields. The complex and multiscale nature of geological and geophysical phenomena require that a wide variety of kinds of models be deployed and a broad spectrum of sources of uncertainty be confronted. Most modelers do not expect their models to give specific, quantitative predictions of the detailed behavior of the systems under investigation. Rather, they are understood as providing a tool by which scientists can test hypotheses (including causal ones), evaluate the relative importance of different elements of the system, develop model-based explanations [41.95, 96], and generate qualitatively accurate projections of future conditions. Indeed, it is precisely by grappling with

these many sources of uncertainty that geoscientists gain insight and understanding into the various processes that shape the Earth, their relative importance and patterns of dependence, and the emergent structures that they produce.

The geosciences, as we have seen, constitute a significant portion of scientific research today. Our philosophies of science and our understanding of the nature of model-based inquiry are inadequate if we do not take this research into account. As we hope this review has made clear [41.44, p. 100]:

“the earth sciences are profoundly important, not only because they challenge conventional philosophical portraits of how scientific knowledge is produced, tested, and stabilized, but also because they matter for the future of the *world*.”

## References

- 41.1 N. Oreskes: How earth science has become a social science. In: *Special Issue: Climate and Beyond: The Production of Knowledge about the Earth as a Signpost of Social Change*, ed. by A. Westermann, C. Rohr, Historical Soc. Res. **40** (2015) 246–270
- 41.2 M. Kleinhans, C. Buskes, H. de Regt: Terra incognita: Explanation and reduction in earth science, *Int. Stud. Phil. Sci.* **19**(3), 289–317 (2005)
- 41.3 G.K. Gilbert: *Report on the Geology of the Henry Mountains* (Government Printing Office, Washington 1877)
- 41.4 G.E. Grant, J.E. O’Connor, M.G. Wolman: A river runs through it: Conceptual models in fluvial geomorphology. In: *Treatise on Geomorphology*, Vol. 9, ed. by J.F. Shroder (Academic, San Diego 2013) pp. 6–21
- 41.5 W.M. Davis: The systematic description of land forms, *Geogr. J.* **34**, 300–318 (1909)
- 41.6 W.M. Davis: The geographical cycle, *Geogr. J.* **14**, 481–504 (1899)

- 41.7 I. Kant: Universal natural history and theory of the heavens or essay on the constitution and the mechanical origin of the whole universe according to Newtonian principles. In: *Kant: Natural Science*, ed. by E. Watkins (Cambridge Univ. Press, Cambridge 2012), transl. by O. Reinhardt, originally published in 1755
- 41.8 P.-S. Laplace: *Exposition du Système du Monde* (Cambridge Univ. Press, Cambridge 2009), originally published in 1796
- 41.9 N. Oreskes: From scaling to simulation: Changing meanings and ambitions of models in the earth sciences. In: *Science without Laws: Model Systems, Cases, and Exemplary Narratives*, ed. by A. Creager, E. Lunbeck, M.N. Wise (Duke Univ. Press, Durham 2007) pp. 93–124
- 41.10 A. Daubrée: *Études Synthétiques de Géologie Expérimentale* (Dunod, Paris 1879), in French
- 41.11 A. Bokulich: How the tiger bush got its stripes: How possibly versus how actually model explanations, *Monist* **97**(3), 321–338 (2014)
- 41.12 M.K. Hubbert: Strength of the earth, *Bull. Am. Assoc. Petroleum Geol.* **29**(11), 1630–1653 (1945)
- 41.13 R. Bagnold: *The Physics of Blown Sand and Desert Dunes* (Dover, Mineola 2005), originally published in 1941
- 41.14 D. Green: Modelling geomorphic systems: Scaled physical models. In: *Geomorphological Techniques (Online Edition)*, ed. by S.J. Cook, L.E. Clarke, J.M. Nield (British Society for Geomorphology, London 2014), Chap. 5, Sect. 3
- 41.15 M. Weisberg: *Simulation and Similarity* (Oxford Univ. Press, Oxford 2013)
- 41.16 E. Winsberg: Computer simulations in science. In: *The Stanford Encyclopedia of Philosophy*, ed. by E. Zalta <http://plato.stanford.edu/archives/sum2015/entries/simulations-science> (Summer 2015 Edition)
- 41.17 M. Kirkby, P. Naden, T. Burt, D. Butcher: *Computer Simulation in Physical Geography* (Wiley, New York 1987)
- 41.18 G. Tucker: Models. In: *Encyclopedia of Geomorphology*, Vol. 2, ed. by A. Goudie (Routledge, London 2004) pp. 687–691
- 41.19 G. Tucker, S. Lancaster, N. Gasparini, R. Bras: The channel–hillslope integrated landscape development model (CHILD). In: *Landscape Erosion and Evolution Modeling*, ed. by H. Doe (Kluwer Academic/Plenum, New York 2001)
- 41.20 T. Coulthard, M. Macklin, M. Kirkby: A cellular model of holocene upland river basin and alluvial fan evolution, *Earth Surf. Process. Landf.* **27**(3), 268–288 (2002)
- 41.21 A. Rowan: Modeling geomorphic systems: Glacial. In: *Geomorphological Techniques*, ed. by L.E. Clark, J.M. Nield (British Society for Geomorphology, London 2011), Sect. 5, Chap. 5.6.5 (Online Version)
- 41.22 CMIP5: World Climate Research Programme’s Coupled Model Intercomparison Project, Phase 5 Multi-Model Dataset, <http://cmip-pcmdi.llnl.gov/cmip5/> (2011)
- 41.23 J. Katzav, W. Parker: The future of climate modeling, *Clim. Change* **132**, 475–487 (2015)
- 41.24 N. Oreskes: The role of quantitative models in science. In: *Models in Ecosystem Science*, ed. by C. Canham, J. Cole, W. Lauenroth (Princeton UP, Princeton 2003)
- 41.25 A. Nicholas, T. Quine: Crossing the divide: Representation of channels and processes in reduced-complexity river models at reach and landscape scales, *Geomorphology* **90**, 318–339 (2007)
- 41.26 A.B. Murray, C. Paola: A cellular model of braided rivers, *Nature* **371**, 54–57 (1994)
- 41.27 T. Coulthard, D. Hicks, M. Van De Wiel: Cellular modeling of river catchments and reaches: Advantages, limitations, and prospects, *Geomorphology* **90**, 192–207 (2007)
- 41.28 A.B. Murray: Contrasting the goals, strategies, and predictions associated with simplified numerical models and detailed simulations. In: *Prediction in Geomorphology*, ed. by P. Wilcock, R. Iverson (American Geophysical Union, Washington 2003) pp. 151–165
- 41.29 B.T. Werner: Complexity in natural landform patterns, *Science* **284**, 102–104 (1999)
- 41.30 A. Bokulich: Explanatory models versus predictive models: Reduced complexity modeling in geomorphology, *Proc. Eur. Philos. Sci. Assoc.: EPSA11 Perspect. Found. Probl. Philos. Sci.*, ed. by V. Karakostas, D. Dieks (Springer, Cham 2013)
- 41.31 A.B. Murray: Reducing model complexity for explanation and prediction, *Geomorphology* **90**, 178–191 (2007)
- 41.32 S. Hall: At fault?, *Nature* **477**, 264–269 (2011)
- 41.33 J. Wainwright, M. Mulligan: Mind, the gap in landscape evolution modelling, *Earth Surf. Process. Landf.* **35**, 842–855 (2010)
- 41.34 T. Kuhn: *The Structure of Scientific Revolutions* (Univ. Chicago Press, Chicago 2012), [1962]
- 41.35 P. Suppes: Models of data, *Proc. Int. Congr. Logic, Methodol. Philos. Sci.*, ed. by E. Nagel, P. Suppes, A. Tarski (Stanford Univ. Press, Stanford 1962) pp. 251–261
- 41.36 I. Lakatos: Falsification and the methodology of scientific research programmes, *Proc. Int. Colloquium Phil. Sci.: Crit. Growth Knowl.*, Vol. 4, ed. by I. Lakatos, A. Musgrave (Cambridge Univ. Press, Cambridge 1970), London, 1965
- 41.37 E. Rykiel: Testing ecological models: The meaning of validation, *Ecol. Model.* **90**, 229–244 (1996)
- 41.38 P. Edwards: *Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming* (MIT Press, Cambridge 2010)
- 41.39 E. Lloyd: The role of complex empiricism in the debates about satellite data and climate models, *Stud. Hist. Philos. Sci.* **43**, 390–401 (2012)
- 41.40 R. Benson, P. Mannion: Multi-variate models are essential for understanding vertebrate diversification in deep time, *Biol. Lett.* **8**(1), 127–130 (2012)
- 41.41 A. Mc Gowan, A. Smith (Eds.): *Comparing the Geological and Fossil Records: Implications for Biodiversity Studies* (Geological Society, London 2011), No. 358. The Geological Society Special Publication

- 41.42 R. Giere: Using models to represent reality. In: *Model-Based Reasoning in Scientific Discovery*, ed. by L. Magnani, N. Nersessian, P. Hagard (Springer, New York 1999)
- 41.43 S. Norton, F. Suppe: Why atmospheric modeling is good science. In: *Changing the Atmosphere: Expert Knowledge and Environmental Governance*, ed. by C. Miller, P. Edwards (MIT Press, Cambridge 2001) pp. 67–106
- 41.44 N. Oreskes: Models all the way down (review of Edwards *A Vast Machine*), *Metascience* **21**, 99–104 (2012)
- 41.45 K. Beven: *Environmental Modelling: An Uncertain Future? An Introduction to Techniques for Uncertainty Estimation in Environmental Prediction* (Routledge, New York 2009)
- 41.46 H. Chang: *Inventing Temperature: Measurement and Scientific Progress* (Oxford Univ. Press, Oxford 2004)
- 41.47 N.K.S. Oreskes: Frechette, K. Belitz: Verification, validation, and confirmation of numerical models in the earth sciences, *Science* **263**, 641–646 (1994)
- 41.48 N. Oreskes, K. Belitz: Philosophical issues in model assessment. In: *Model Validation: Perspectives in Hydrological Science*, ed. by M. Anderson, P. Bates (Wiley, West Sussex 2001) pp. 23–42
- 41.49 N. Oreskes: Evaluation (not validation) of quantitative models, *Environ. Health Perspect.* **106**(supp. 6), 1453–1460 (1998)
- 41.50 G. Lauder: On the inference of function from structure. In: *Functional Morphology in Vertebrate Paleontology*, ed. by J. Thomason (Cambridge Univ. Press, Cambridge 1995) pp. 1–18
- 41.51 J. Hutchinson, M. Garcia: Tyrannosaurus was not a fast runner, *Nature* **415**, 1018–1021 (2002)
- 41.52 M. Weisberg: Robustness analysis, *Phil. Sci.* **73**, 730–742 (2006)
- 41.53 B. Calcott: Wimsatt and the robustness family: Review of Wimsatt's re-engineering philosophy for limited beings, *Biol. Phil.* **26**, 281–293 (2011)
- 41.54 A. Saltelli, K. Chan, M. Scott: *Sensitivity Analysis* (Wiley, West Sussex 2009)
- 41.55 J. Hutchinson: On the inference of structure using biomechanical modelling and simulation of extinct organisms, *Biol. Lett.* **8**(1), 115–118 (2012)
- 41.56 D. Hamby: A review of techniques for parameter sensitivity analysis of environmental models, *Environ. Monit. Assess.* **32**, 135–154 (1994)
- 41.57 A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, S. Tarantola: *Global Sensitivity Analysis: The Primer* (Wiley, West Sussex 2008)
- 41.58 R. Snieder, J. Trampert: Inverse problems in geophysics. In: *Wavefield Inversion*, ed. by A. Wirgin (Springer, New York 1999) pp. 119–190
- 41.59 G. Backus, J. Gilbert: Numerical applications of a formalism for geophysical inverse problems, *Geophys. J. R. Astron. Soc.* **13**, 247–276 (1967)
- 41.60 M. Sen, P. Stoffa: Inverse theory, global optimization. In: *Encyclopedia of Solid Earth Geophysics*, Vol. 1, ed. by H. Gupta (Springer, Dordrecht 2011)
- 41.61 W. Sandham, D. Hamilton: Inverse theory, artificial neural networks. In: *Encyclopedia of Solid Earth Geophysics*, ed. by H. Gupta (Springer, Dordrecht 2011) pp. 618–625
- 41.62 G. Belot: Down to earth underdetermination, *Phil. Phenomenol. Res.* **XCI** **2**, 456–464 (2015)
- 41.63 T. Miyake: Uncertainty and modeling in seismology. In: *Reasoning in Measurement*, ed. by N. Mössner, A. Nordmann (Taylor Francis, London 2017)
- 41.64 E. Tal: The Epistemology of Measurement: A Model-Based Account, Ph.D. Thesis (Univ. Toronto, London 2012)
- 41.65 S. Schumm: *To Interpret the Earth: Ten Ways to be Wrong* (Cambridge UP, Cambridge 1998)
- 41.66 S. Lane: Numerical modelling: Understanding explanation and prediction in physical geography. In: *Key Methods in Geography*, 2nd edn., ed. by N. Clifford, S. French, G. Valentine (Sage, Los Angeles 2010) pp. 274–298, 2003
- 41.67 J. O'Reilly, K. Brysse, M. Oppenheimer, N. Oreskes: Characterizing uncertainty in expert assessments: Ozone depletion and the west antarctic ice sheet, *WIREs Clim. Change* **2**(5), 728–743 (2011)
- 41.68 W. Parker: Predicting weather and climate: Uncertainty, ensembles, and climate, *Stud. Hist. Phil. Mod. Phys.* **41**, 263–272 (2010)
- 41.69 R. Frigg, S. Bradley, H. Du, L. Smith: Laplace's demon and the adventures of his apprentices, *Phil. Sci.* **81**, 31–59 (2014)
- 41.70 E.L. Thompson: Modelling North Atlantic Storms in a Changing Climate, Ph.D. Thesis (Imperial College, London 2013)
- 41.71 N. Odoni, S. Lane: The significance of models in geomorphology: From concepts to experiments. In: *The SAGE Handbook of Geomorphology*, ed. by K. Gregory, A. Goudie (SAGE, London 2011)
- 41.72 G. Sella, S. Stein, T. Dixon, M. Craymer, T. James, S. Mazzotti, R. Dokka: Observation of glacial isostatic adjustment in stable North America with GPS, *Geophys. Res. Lett.* **34**(2), 1–6 (2007), L02306
- 41.73 T. Chamberlin: The method of multiple working hypotheses, *Science* **15**(366), 92–96 (1890)
- 41.74 R. Laudan: The method of multiple working hypotheses and the development of plate tectonic theory. In: *Scientific Discovery: Case Studies*, Boston Studies in the Philosophy of Science, Vol. 60, ed. by T. Nickles (Springer, Dordrecht 1980) pp. 331–343
- 41.75 M. Richards: The cretaceous-tertiary mass extinction: What really killed the dinosaurs?, <http://hmn.harvard.edu/file/366291> (2015) Lecture given on February 3rd, 2015 at the Harvard Museum of Natural History
- 41.76 C. Cleland: Prediction and explanation in historical natural science, *Br. J. Phil. Sci.* **62**, 551–582 (2011)
- 41.77 A.B. Murray: Cause and effect in geomorphic systems: Complex systems perspectives, *Geomorphology* **214**, 1–9 (2014)
- 41.78 C. Paola, K. Straub, D. Mohrig, L. Reinhardt: The unreasonable effectiveness of stratigraphic and geomorphic experiments, *Earth Sci. Rev.* **97**(1–4), 1–43 (2009)

- 41.79 P. Duhem: *The Aim and Structure of Physical Theory* (Princeton Univ. Press, Princeton 1954), trans. P. Wiener, 1906
- 41.80 K. Stanford: Underdetermination of scientific theory. In: *Stanford Encyclopedia of Philosophy*, ed. by N. Edward, E. Zalta <http://plato.stanford.edu/archives/win2013/entries/scientific-underdetermination> (Winter 2013 Edition)
- 41.81 K. Beven: Prophecy, reality and uncertainty in distributed hydrological modelling, *Adv. Water Resour.* **16**(1), 41–51 (1993)
- 41.82 K. Beven, J. Freer: Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, *J. Hydrol.* **249**(1–4), 11–29 (2001)
- 41.83 K. Beven: Equifinality and uncertainty in geomorphological modeling, *Proc. 27th Binghampton symp. geomorphol.: Sci. Nat. Geomorphol.*, ed. by B. Rhoads, C. Thorn (Wiley, Hoboken 1996) pp. 289–313
- 41.84 K. Beven, A. Binley: GLUE: 20 years on, *Hydrol. Process.* **28**(24), 5897–5918 (2014)
- 41.85 J.P.C. Kleijnen: Experimental design for sensitivity analysis, optimization, and validation of simulation models. In: *Handbook of Simulation: Principles, Methodology, Advances, Applications, and Practice*, ed. by J. Banks (Wiley, New York 1998) pp. 173–223
- 41.86 N. Odoni: Exploring Equifinality in a Landscape Evolution Model, Ph.D. Thesis (Univ. Southampton, School of Geography, Southampton 2007)
- 41.87 R.L. Slingerland, G. Tucker: Erosional dynamics, flexural isostasy, and long-lived escarpments, *J. Geophys. Res.* **99**, 229–243 (1994)
- 41.88 R. Knutti, J. Sedláček: Robustness and uncertainties in the new CMIP5 climate model projections, *Nat. Clim. Change* **3**, 369–373 (2013)
- 41.89 E. Lloyd: Confirmation and robustness of climate models, *Phil. Sci.* **77**, 971–984 (2010)
- 41.90 W. Parker: When climate models agree: The significance of robust model predictions, *Phil. Sci.* **78**, 579–600 (2011)
- 41.91 W. Parker: Ensemble modeling, uncertainty, and robust predictions, *WIREs Clim. Change* **4**, 213–223 (2013)
- 41.92 J. Lenhard, E. Winsberg: Holism, entrenchment, and the future of climate model pluralism, *Stud. Hist. Phil. Mod. Phys.* **41**, 253–262 (2010)
- 41.93 D. Masson, R. Knutti: Climate model genealogy, *Geophys. Res. Lett.* **38**, L08703 (2011)
- 41.94 R. Knutti, R. Furrer, C. Tebaldi, J. Cermak, G. Meehl: Challenges in combining projections from multiple climate models, *J. Clim.* **23**(10), 2739–2758 (2010)
- 41.95 A. Bokulich: How scientific models can explain, *Synthese* **180**(1), 33–45 (2011)
- 41.96 A. Bokulich: Models and explanation. In: *Handbook of Model-Based Science*, ed. by L. Magnani, T. Bertolotti (Springer, Dordrecht 2016)

# Models in the

## 42. Models in the Biological Sciences

Elisabeth A. Lloyd

Evolutionary theory may be understood as a set of overlapping model types, the most prominent of which is the natural selection model, introduced by Charles Darwin and Alfred Russel Wallace. Many of the most prominent models today are represented through mathematical population genetics, in which genetical representations of populations evolve over time to produce evolutionary change. I review the variety of evolutionary models – from genic to group to species selection models – and how they are confirmed through evidence today. I discuss both applications to cases where we do not know the genetics, and to animal behavior and evolution.

<b>42.1 Evolutionary Theory</b> .....	913
42.1.1 The Structure of Darwinian Evolutionary Models.....	913
42.1.2 The Structure of Population Genetic Evolutionary Models .....	914
42.1.3 Representational Adequacy of Models .	918
42.1.4 Expansions and Alternative Views of the Structure of Evolutionary Theory	920
<b>42.2 Confirmation in Evolutionary Biology</b> .	922
42.2.1 Confirming and Testing Models .....	922
<b>42.3 Models in Behavioral Evolution and Ecology</b> .....	925
42.3.1 The Phenotypic Gambit.....	925
42.3.2 Evolutionary Stable Strategies, Animal Signalling .....	925
42.3.3 Physiological/Evolution Models, Cognitive Ethology .....	926
42.3.4 Optimality Models Including Agent Models .....	927
<b>References</b> .....	927

### 42.1 Evolutionary Theory

Charles Darwin proposed a general type of natural selection model that could explain a variety of particular cases of adaptation to local environments, once details of organismic traits and selection pressure were inserted. Much of evolutionary theory today, though not all, is represented through mathematical models, especially through the models of population genetics, of the evolution of states of a given system, both in isolation and interaction through time. This chapter discusses in detail various ways to describe the evolutionary models that make up evolutionary theory. The main items needed for this description are the model types of natural selection, drift, and so forth, most often described through the definition of a state space, state variables, parameters, and a set of laws of succession and coexistence for the system. Choosing a *state space* (and thereby, a set of state variables) for the representation of genetic states and changes in a population is a crucial part of population genetics theory. Claims about evo-

lutionary models may be confirmed in three different ways:

1. Through fit of the outcome of the model to a natural system
2. Through independent testing of assumptions of the model, including parameters and parameter ranges
3. Through a range of instances of fit over a variety of the natural systems to be accounted for by the model, through a variety of assumptions tested, and including both instances of fit and some independent support for aspects of the model.

#### 42.1.1 The Structure of Darwinian Evolutionary Models

The basic structure of Darwinian evolutionary models is deceptively simple and elegant; such simplicity can yield powerful change over the proposed time spans

encountered by the biological systems involved. For evolution by natural selection, for example, we start with a basic very general model structure that includes a set of assumptions or components [42.1, 2]. *Van Fraassen* has defined *model type* as the description of a structure in which certain parameters are left unspecified, like this [42.3, p. 44].

#### Definition 42.1 Model Type

*Model type* = a structure in which certain parameters are left unspecified [42.3, p. 44].

The classic selection model type concerns a population of varied organisms, and these variations are assumed to be at least partially inheritable. In this population's environment (which is always defined relative to the organism's needs and sensitivities), there are demands and stresses affecting some variants more than others, resulting in differential reproduction. Thus, the basic natural selection model type leads to the change in the population variants' composition and structure over evolutionary time. (For discussion about multilevel selection model types, see *Lloyd* [42.4]).

In 1859, Charles Darwin, in *On the Origin of Species*, proposed his basic abstract selection model type, and other evolutionary model types, and also filled in their details in various cases in various ways in the book and his correspondence; in this manner, he illustrated how to use and apply the models.

#### Definition 42.2 Natural Selection Model Type

Population (with details filled in) \_\_\_\_\_  
 Traits \_\_\_\_\_  
 Genetic basis \_\_\_\_\_  
 Correlation with fitness \_\_\_\_\_  
 Selection pressure, environment \_\_\_\_\_

More specifically, each of the model types serve as formats for explanation; the particular terms or factors in a model type vary in each application, depending on the outcome of the model and various assumed conditions [42.1, 2]. For example, in the *Origin*, Darwin needed evidence for his assumption that wild organisms spontaneously developed heritable variations some of which would be advantageous to their survival and reproduction. Such a general assumption of the existence of useful variations was necessary for a variable in the natural selection model type, which would then be instantiated in specific models by individual cases of variation. Darwin found empirical evidence for this assumption in the animal breeders' information, as he

presents in the *Origin*. (Darwin kept up an avid and long-term correspondence with a number of pigeon breeders, from which he learned detailed information about the vast variety of feather form, pattern, and color, as well as behaviors, which spontaneously arose in their pigeons.) Elsewhere in his correspondence, Darwin also offers evidence to support an empirical assumption of a specific model constructed using the selection model type. In looking at wingless insects that appear on oceanic islands at a higher concentration than elsewhere, Darwin proposed to explain their frequency by a selection pressure of high winds blowing them off the islands, thus favoring insects without wings. He tested this model assumption by comparing an even smaller island to the other islands he was examining earlier: he found, as he predicted, even a higher percentage of wingless insects. His earlier assumption of a selection pressure was thus indirectly confirmed in his proposed model [42.1, pp. 121–122], [42.5, p. 401], [42.6, pp. 226–227]. See Sect. 42.2 for more detailed discussion about the confirmation of evolutionary models. Note that the evolutionary selection models can range from the very abstract, general, model structure with its high level assumptions, through ever more specified models as we identify and fill in the necessary assumptions of selection pressures, types of variation, and evidence for heritability. Ultimately, our ever more detailed specifications of the assumptions result in a fully specified selection model, which anchors the most concrete end of the model-continuum, from most abstract model type to most concrete model [42.7], [42.2, pp. 106–107], [42.1, pp. 118–119].

#### 42.1.2 The Structure of Population Genetic Evolutionary Models

While Darwin's evolutionary models and explanations were nonquantitative, much of evolutionary theory at the beginning of the twenty-first century is represented through mathematical models or equations, especially through the models of population genetics, of the evolution of states of a given system, both in isolation and interaction, through time. This is done by conceiving of the evolutionary model as capable of a certain set of states – these states are represented by elements of a certain mathematical space, the state space [42.5, 6, 8, 9].

#### Definition 42.3 (State Space)

A *state space* is a mathematical space specified by variables capable of a certain set of states; the collection of all possible configurations or states of these variables.

**Definition 42.4 (Data Models)**

*Data models* are simplified structures representing the natural world made up of measurements, observations, or the results of experiments, that is, numbers or values extracted from the real world, many of which are arranged in relations suggested by the theoretical models.

Generally speaking, *models* and *systems* always refer to ideal systems, described theoretically. When the actual biological systems are being discussed, they are called *empirical* or *natural* systems. The variables used in each model represent distinct measurable or potentially quantifiable, physical, or biological magnitudes. Classically, any particular configuration of values for these variables is a *state* of the system, the *state space* being the collection of all possible configurations of the variables.

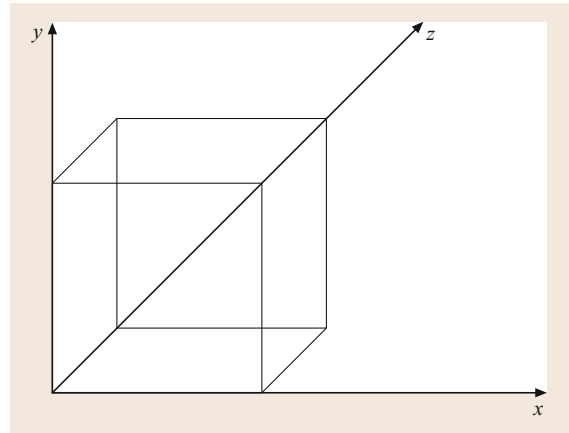
The theory itself represents the behavior of the system in terms of its states: the rules or laws of the theory (i. e., laws of coexistence, succession, or interaction) can delineate various configurations and trajectories on the state space. A description of the structure of the theory itself therefore only involves the description of the set of models, which make up the theory. In an application of the theory or models to the real world, we consider both *data models* and the idealized models describing idealized systems and their resemblance. *Data models* are simplified structures representing the natural world made up of measurements, observations, or the results of experiments, that is, numbers or values extracted from the real world, many of which are arranged in relations suggested by the theoretical models [42.10].

The ideal systems are usually gradually specified, as described in Sect. 42.1.1, to show a good match or similarity with the data models, to which they are compared during an application of the theory or model type [42.3, 11]. Construction of a model within the theory involves assignment of a location in the state space of the theory to a system of the kind defined by the theory. Potentially, there are many kinds of systems that a given theory can be used to describe – limitations come from the dynamical sufficiency (whether it can be used to describe the system accurately and completely) and the accuracy and effectiveness of the laws used to describe the system and its changes.

**Definition 42.5 (Dynamical Sufficiency)**

The concept of *dynamical sufficiency* concerns what state space and variables are sufficient to describe the evolution of a system given the parameters being used in the specific system.

Thus, there are two main aspects to defining a model. First, the state space must be defined – this involves



**Fig. 42.1** A three-dimensional Cartesian coordinate system, with axes  $x$ ,  $y$ , and  $z$  (after [42.12])

choosing the variables and parameters with which the system will be described, e.g. in Fig. 42.1; second, coexistence laws, which describe the structure of the system, and laws of succession, which describe changes in its structure, must be defined.

**Definition 42.6 (Coexistence Laws)**

Laws representing compatible states of the state variables.

**Definition 42.7 (Laws of Succession)**

Laws representing progressive changes of states of a system.

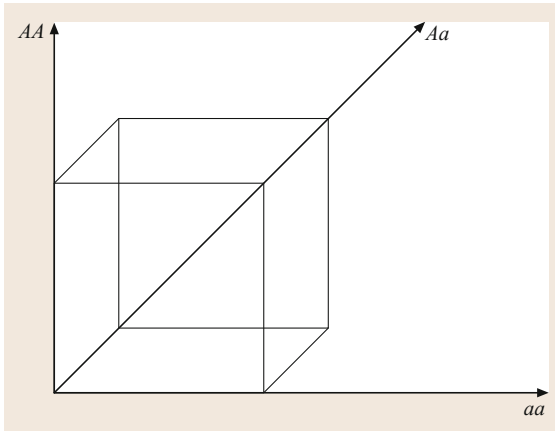
**Definition 42.8 (Parameters)**

Values that are not themselves a function of the state of the system. Unvarying values.

Defining the state space involves defining the set of all the states the system could possibly exhibit. Certain mathematical entities – in the case of many evolutionary models, these are vectors – are chosen to represent these states. The collection of all the possible values for each variable assigned a place in the vector is the state space of the system. The system and its states can have a geometrical interpretation: the variables used in the state description (i. e., state variables) can be conceived as the axes of a Cartesian space.

The state of the system at any time may be represented as a point in that space, located by projection on the various axes. The family of measurable physical magnitudes, in terms of which a given system is defined, also includes a set of parameters. *Parameters* are values that are not themselves a function of the state





**Fig. 42.2** Genotype space with genotypic coordinates  $AA$ ,  $Aa$ , and  $aa$  (after [42.13])

of the system. Thus, a parameter can be understood as a *fixed value* of a variable in the state space – topologically, setting a parameter amounts to limiting the number of possible structures in the state space by reducing the dimensionality of the model.

Laws, used to describe the behavior of the system in question, must also be defined in a description of a model or set of models. Laws have various forms: in general, coexistence laws describe the possible states of the system, while laws of succession describe changes in the state of the system.

Let us discuss in, more detail, the description of the evolutionary models that make up current evolutionary theory. The main items needed for this description are the definition of a state space, state variables, parameters, and a set of laws of succession and coexistence for the system. Choosing a *state space* (and thereby, a set of state variables) for the representation of genetic states and changes in a population is a crucial part of population genetics theory. (Population genetics models are a most technical, but not necessarily a *core*, part of evolutionary theory itself (but see *Thompson* [42.6]).

*Paul Thompson* suggests that the state space for population genetics would include the physically possible states of populations in terms of genotype frequencies. The state space would be “a Cartesian  $n$ -space where  $n$  is a function of the number of possible pairs of alleles in the population” [42.6, p. 223]. We can picture this geometrically as  $n$  axes, the values of which are frequencies of the genotype, as in Fig. 42.2.

The state variables are the frequencies for each genotype. Note that this is a one-locus system, that is, we take only a single gene locus and determine the dimensionality of the model as a function of the number of alleles at that single locus.

Another type of single locus system, used less commonly than the one described by *Thompson*, involves using single gene frequencies, rather than genotype frequencies, as state variables. Some of the debates about *genic selectionism* center around the descriptive, dynamical, and parametric adequacy of this state space and its parameters for representing evolutionary phenomena (*Dawkins* [42.14], *Lloyd et al.* [42.15], see later discussion of genic selection and dynamical insufficiency).

With both genotype and gene frequency state spaces, treating the genetic system of an organism as being able to be isolated into single loci involves a number of assumptions about the system as a whole. For instance, if the relative fitnesses of the genotypes at a locus are dependent on other loci, then the frequencies of a single locus observed in isolation will not be sufficient to determine the actual genotype frequencies. Assumptions about the structure of the system as a whole can thus be incorporated into the state space in order to reduce its dimensionality. *Lewontin* [42.16] offers a detailed analysis of the quantitative effects of dimensionality of various assumptions about the biological systems being modeled. Interactions between genotypes, and between one locus and another cannot be represented in a single locus model (with the exception of frequency-dependent selection), for the simple reason that they involve more than one locus [42.16, pp. 273–81].

As far as the structure of the theory goes, although all single locus models should, in some sense, be grouped together, they are not all exactly the same model – each particular model has a different number of state variables, depending upon the number of alleles at that locus. Because a model type is simply an abstraction of a model, constructed by abstracting one or more of the models parameters or variables, a single model can be an instance of more than one model type. This is an extremely important aspect of the flexibility of this approach to theory structure.

Along similar lines, we may say that each model type be associated with a distinctive state space type. In the preceding example, the single locus model is to be taken as an instance of a general state space type for all single locus models, i. e., the different single locus model types are conceived as utilizing the same state space type. Alternatives, such as two-locus models, must be taken as instances of a different state space type.

*Parameters* are the values that appear in the succession and coexistence laws of a system that are the same for all possible states of the defined system. For instance, in the modification of the Hardy–Weinberg

equation that predicts the frequencies of the genotypes after selection, the selection coefficient,  $s$ , or multiplier, appears as a parameter in the formula  $p^2 AA + 2pqAa + (1-s)q^2 aa = 1 - sq^2$  [42.17, p. 102], [42.18, p. 81].

There are a variety of methods of establishing the value at which a parameter should be fixed or set in the construction of models for a given real system, for example, simulation techniques can be used to obtain estimates of biologically important parameters. In some contexts, maximum likelihood estimations may be possible. Parameters can also be set arbitrarily or ignored. This is equivalent to incorporating certain assumptions into the model for purposes of simplification [42.19, pp. 8,89].

One expects the values of parameters to have an impact on the system being represented; but variations in parameter values can make a larger or smaller amount of difference to the system. For instance, take a deterministic model that incorporates a parameter for mutation,  $\mu$ . The rates of change of this model can be virtually insensitive to realistic variations in the value of the mutation parameter,  $\mu$ , because they are several orders of magnitude smaller. (However, mutation rates do sometimes play important roles with significant consequences, for example, in *Kimura and Ohta* [42.20] (nearly neutral theory), and in *Kondrashov's hatchet* (in the evolution of sex, [42.21]).

Going back to Hardy–Weinberg models, the selection parameters play a crucial role in these models. A very small amount of selection in favor of an allele will have a cumulative effect strong enough to replace other alleles [42.16, p. 267].

Population size is another case in which the value assigned to the parameter has a large impact on the model results. The parameter for the effective population size,  $N$ , can play a crucial role in some models, because selection results can be quite different with a restricted gene pool size [42.22, pp. 48–50]. In many of the stochastic models involved in calculating rates of evolutionary change, the resulting distributions and their moments can depend completely on the ratio of the mean deterministic force to the variance arising from random processes [42.16, p. 268]. This variance is usually proportional to  $1/N$  and is related to the finiteness of population size. Thus, change in the value of the single parameter,  $N$ , can completely alter the structure represented by the theory.

The choice of parameters can also make a major difference to the model outcome. Theoreticians have choices about how to express certain aspects of the system or environment. The choice of parameters used to represent the various aspects can have a profound effect on the structure, even to the point of rendering the model useless for representing the real world system

in question. Group selection models provide a case in which choice of parameters not only alters the results of the models, but also can lead to the near disappearance of the phenomenon being modeled [42.23, 24].

Some authors, when discussing genetic changes in populations, speak of the system in terms of a type of state space involving phenotypes. This makes sense, because the phenotype determines the breeding system and the action of natural selection, the results of which are reflected in the genetic changes in the population; quantitative genetics is the set of models that concentrate on phenotypic state spaces. In his analysis of the present structure of population genetics theory, *Lewontin* traces a single calculation of a change in genetic state through both genotypic and phenotypic descriptions of the population [42.16].

But in addition, mating patterns, and probabilities of survival and reproduction, although influenced by genes, are a consequence of developmental events that are contingent on the environment of the developing organism, and involve more than any simple phenotype. In the most general case, environment includes influences of the phenotypes of previous generations by means of cytoplasmic inheritance through the egg. A complete general representation of genetic evolutionary processes then requires not two, as is usually done, but six spaces with sequential transitions within them and mappings from one to the other [42.15]. Haploid spaces represent only the single chromosome or set of one sex's genes, while (sexually reproducing) diploid state spaces represent both sets. The complete set of six spaces,  $S$ , is thus:

- $S_1$  A diploid phenotypic space
- $S_2$  A diploid genotypic space
- $S_3$  A diploid pair phenotypic space
- $S_4$  A diploid pair genotypic space
- $S_5$  A haploid phenotypic space and
- $S_6$  A haploid genic space.

(Assuming that organisms reproduce in discrete generations, with no overlap between generations.) While it is true that the system's transition from one generation to the next may be represented in any one of the six spaces, we need to know the entire loop – and all the parameter values in each of the model stages in that loop – in order to obtain a fully accurate representation of any chosen state space in the next generation, i. e., to get the state transformation equation between generations *within any one of the spaces* (e.g., if we want to move from ( $S_6$ ) to ( $S_6$  in the new generation)). For instance, for the transition in allelic space, we must move out of that space, into genotypic space to define the fitness parameters, and back into allelic space in order to

characterize the next generation [42.15, pp. 143–144]. This case illustrates the issue of the representational adequacy of models, to which we shall now turn.

### 42.1.3 Representational Adequacy of Models

The most common approach to comparing population genetic models has emphasized prediction of allele frequency changes: If two models both predict the same changes in allele frequencies, it is thought, then the models are *equivalent*. But this is an inadequate approach to understanding and confirming models. We advance the notion of *representational adequacy*, which we define as parametric and dynamical sufficiency. Why introduce representational adequacy? What is wrong with straightforwardly checking whether the model fits the allele frequency data? There are a variety of ways that models can be tested against data, and fitting the outcome or prediction of the model – in these cases, the predicted allelic or genotypic frequencies – is only one of them [42.25].

Other significant components of the empirical evaluation of any mathematical model include: testing the values of its parameters against the system independently (e.g., measuring or estimating the mutation parameter value in the model); evaluating the appropriateness of the state space and parameters used; and testing the model against a range of values in the variety of systems to which it is supposed to apply (variety of fit) (Sect. 42.2). In addition, a model is taken to be better confirmed when it has more of its parameter values – that is, a variety of them – estimated or confirmed independently [42.9, pp. 145–159]. Our notion of representational adequacy combines the traditional standards of predictive accuracy and goodness of fit with the broader requirements of confirming that the state space, parameters, and laws being used in the models are appropriate and sufficient to the task [42.26–28]. We take it as foundational to any notion of adequate representation that the standards of parametric sufficiency in model building be weighed in judging overall model adequacy. Parametric sufficiency is dependent upon choice of space and parameter set, in any particular case.

The concept of dynamical sufficiency is precisely defined in terms of a set of objects and their frequencies, and another set of objects and their frequencies. Dynamical and parametric sufficiency together provide a much more adequate measure of a model's empirical worth than the vague notion of empirical adequacy, or the overly simplistic idea that if a model's prediction of allelic or genotypic frequency is correct, then the model is empirically substantiated.

#### Parametric Sufficiency

Parameters are properties of objects, and may be properties of more than one object at a time. (Parameters are represented in the models as values that are not variables.) Two models may look as though they should be dynamically equivalent because of the similar appearance and *names* of their parameters, such as *fitness*, but real differences in their parameter measures may result in dynamics that are not equivalent. Two models should be considered *parametrically equivalent* if the parameters that apply in one model have a natural representation in the parameters that apply in another. One case in which similar-looking parameters yield very different dynamics concerns allelic and genotypic fitnesses, which has led to much confusion in current controversies.

#### Dynamical Sufficiency

The concept of dynamical sufficiency concerns what state space and variables are sufficient to describe the evolution of a system given the parameters being used in the specific system. What happens to the frequency of the variable over time? In a simple allelic model, this question becomes: Can we describe the changes in the frequency of allele A over time, with the information that we have, which includes the state space (variables) and the parameters (fitnesses, population size, etc.)?

*Problematic Claims.* One widely repeated set of claims has revolved around [42.29] assessment of the mathematical equivalence of a wide variety of population genetic models [42.30, p. 577], [42.31, p. 57, pp. 98–99], [42.32, pp. 168–169, p. 172], [42.33, p. 479, p. 508], [42.34, p. 312]

*Dugatkin and Reeve's Formulation of Allelic Equivalence.* Dugatkin and Reeve claim [42.29, p. 108]:

“A number of theoretical investigations [42.35–43] have shown that the mathematics of the gene-, individual-, kin-, and new group-selection approaches are equivalent [...] We will show [...] that this must be the case.”

They do this with a pair of inequalities [42.29, p. 109]. Dugatkin and Reeve claim that their inequality (2) “encompasses both broad-sense individual selection and any form of trait-group selection that one may care to envision.” Moreover, they conclude, “If broad-sense individual selection, genic selection, and trait-group selection all can be represented by a single condition based only on allele frequencies, then they cannot fundamentally differ from one another” [42.29, p. 109].

But there are serious problems here. The first is that their inequalities (1) and (2) hide completely the causes of why the numbers of alleles change, is misleading and illustrates the importance of both dynamic and parametric sufficiency.

For a true haploid dynamic, we would usually write for one locus, but this also necessarily involves genotypic parameters for the fitnesses, as well as the allelic frequencies. In other words, the apparently purely-allelic parameters depend crucially on genotypic fitnesses – but these values are usually *completely hidden*. The assumption of Mendelian transmission is also made in deriving (3) (see also [42.44]). For this two-allele diploid case, if meiotic drives were introduced, one more parameter,  $k$ , giving the probability that  $A$  is produced by  $Aa$  heterozygotes, would be necessary. Thus, we note that in describing the algebra, many discussions of allelic models and their algebra obscure the origins of that algebra and all the information it contains and represents. A common move is to infer that the genic state space, and its basic entity, the allele, have a metaphysically fundamental and autonomous character [42.30]. This is both biologically and mathematically problematic. Biologically, the changes represented are dependent on all changes in the entire generational cycle reviewed above, represented in the variety of spaces and parameters. Mathematically, because genic space is dynamically insufficient for representing many system changes, and because in diploids there are no allelic transition laws within allelic space with allelic parameters, there is nothing autonomous about it; thus, it cannot support the metaphysical inferences based on its supposed autonomy [42.15, 45]. But there is another account available which might be thought to avoid some of the above problems.

In this crucial case, from Dugatkin and Reeve, the allele frequencies  $G$  in the allelic state space are dynamically sufficient to study the evolution of genotype frequencies  $k$ . These equivalences depend, however, on the parameters,  $w_{11}$ ,  $w_{12}$ ,  $w_{22}$ , that could only be determined in the genotypic space.

Increasing attention has recently been paid to the phenomenon of epigenetics, which includes a variety of biological processes that act on genes and may be transmitted between generations, but not according to any of the rules of genetic inheritance. Formal models for the evolution of epigenetic objects or properties (Feldman and Cavalli-Sforza [42.45, 46]; see also Jablonka and Lamb [42.47]) utilize an additional state space  $S_7$ , the phenogentotype, in which changes may occur during organismal development.

Now consider a few particular aspects and forms of the *laws* used in population genetic models. Coexistence laws describe the possible states of the system

in terms of the state space. In the case of evolutionary theory, these laws would consist of conditions delineating a subset of the state space that contains only the biologically possible states. Laws of succession describe changes in the state of the system. In the case of evolutionary theory, dynamic laws concern changes in the genetic composition of populations. The laws of succession select the biologically possible trajectories in the state space, with states at particular times being represented by points in the state space. The law of succession is the equation of which the biologically possible trajectories are the solutions. The Hardy–Weinberg equation is the fundamental law of both coexistence and succession in population genetics theory. Even the dynamic laws of the theory are usually used to assess the properties of the equilibrium states and steady state distributions. The Hardy–Weinberg law is a very simple, deterministic succession law that is used in a very simple state space. As parameters are added to the equation, we get different laws, technically speaking. For example, compare the laws used to calculate the frequency  $p'$  of the  $A$  allele in the next generation. Plugging only the selection coefficient against a recessive into the basic Hardy–Weinberg law, we get the recursion for the dominant allele frequency as  $p' = p/(1 - sq^2)$ .

Addition of parameters for the mutation rates yields a completely different law,  $p'' = 1 - \mu p' + \nu q'$ , where  $\mu$  is the rate of mutation away from the dominant allele, and where  $\nu$  is the rate of mutation toward it. We could consider these laws to be of a single type – variations on the basic Hardy–Weinberg law – that are usually used in a certain state space type. The actual state space used in each instance depends on the genetic characteristics of the natural system, and not usually on the parameters. For instance, the succession of a system at Hardy–Weinberg equilibrium and one that is not at equilibrium but is under selection pressure, could both be modeled in the same state space, using different laws.

A theory can have either deterministic or statistical laws for its state transitions. Furthermore, the states themselves can be either statistical or nonstatistical.

#### Definition 42.9 Statistical Laws

*Statistical laws* are constructed by specifying a probability measure on the state space. For example, we could assign probabilities (frequencies) to each distinct possible value of gene frequency. Thus, the probability measure is constructed by taking a certain value for the gene frequency, obtaining the joint distribution (for example, through simulation) and making a new state space of probabilities on the old state space of gene frequencies.

In population genetics models, gene frequencies often appear in the set of state variables, thus the states themselves are statistical entities. In general, a law is deterministic if, when all of the parameters and variables are specified, the succeeding states are uniquely determined. In population genetics, this means that the initial population and parameters are all that is needed to get an exact prediction of the new population state.

In sum, the structure of evolutionary theory can be understood by examining the families of models it presents. In the case of population genetics theory, the set of model types – stochastic and deterministic, single locus or multilocus – can be understood as related families of models. The question then becomes defining the exact nature of the relationships among them, and how they relate to the rest of the evolutionary models, such as phenotypic or evolutionary developmental models.

#### 42.1.4 Expansions and Alternative Views of the Structure of Evolutionary Theory

In philosophy of science, the above analysis instantiates an approach to theory structure known as the *semantic view*, that focuses on models [42.3, 10, 11, 48–50].

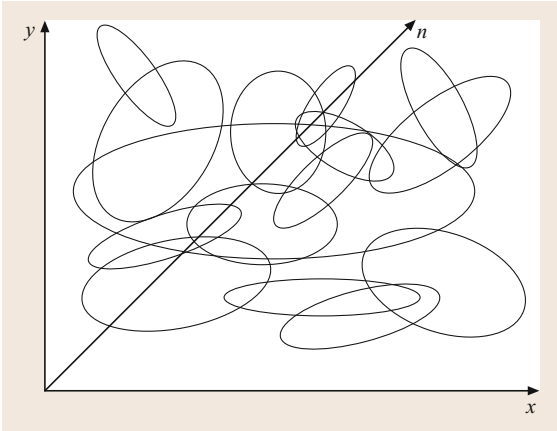
##### Definition 42.10

The *semantic view* states that scientific theories are understood as being interpretable as families of models (viewed in terms of set theory, not necessarily algebraically). More recently it has gone under the name, the *model-based approach*. The *syntactic* or *received view* comes from the *logical positivist* tradition, where scientific theories are viewed as axiomatizable sets of sentences. The axioms are the universal laws of the theory, and all that can be derived from those axioms are the regularities of the theory. Evolutionary theory had been criticized for not being easily formulated in terms of the received view.

Some authors have misrepresented the semantic approach, while some also support what they tend to call a *model-based* approach, whose differences from the semantic view are unclear [42.51, 52]. For instance, Downes [42.51, p. 143], Godfrey Smith [42.52, p. 731], and Love [42.53, p. 7] all claim that the semantic view is committed to “mathematized theories,” which is not useful for some biological contexts. But both Beatty [42.5] and Lloyd [42.1] analyzed Darwin’s (non-mathematical) selection theory and models as their first presentations of the semantic view of evolutionary theory. So clearly the semantic view can be used to analyze *nonmathematized* theories and biological contexts. Especially in the hands of Suppes, the semantic view

can be used to analyze any structures into set theoretic terms, and thus into *mathematical* terms (mathematics concerns structures, and all structures can be seen as *mathematical objects*); but the whole purpose is to represent the system in the most accurate, most useful, or most beneficial, etc. way. This does *not* mean that the analyzed structures are themselves represented mathematically *before* the analysis, and also does not mean that they are represented in equations *after* the analysis. Some authors also neglect the various hierarchies of evolutionary models offered using the semantic view, as outlined in Sect. 42.1.1, with misleading results (e.g., Godfrey Smith [42.52, pp. 732–739], Lloyd [42.1, pp. 118–119]).

In contrast, biologist Samuel Scheiner has recently offered an expanded account of the structure of evolutionary theory that takes the semantic view as its touchstone, and goes on to include other areas of biology [42.54]. Alan Love has also provided a very interesting expansion of the analysis of evolutionary theory through the inclusion of developmental biology models and model types. The analysis of the structure of the theory offered above for population genetics can thus be advanced and expanded (using a variety of detailed methods of describing models, if needed), to incorporate developmental biology, by extending the linkages and overlaps of the model families understood to make up evolutionary theory. Love’s overall approach is extremely useful, as it focuses on which research problems in evolutionary biology subfields compose the theory as a whole [42.54, p. 428], [42.55, 56]. Just as on the earlier analysis of Darwinian models in Sect. 42.1.1, on Love’s view, the entire hierarchy of models is explored, from the very abstract or general, to the extremely concrete. He emphasizes that different researchers use mid-level models in different ways and toward different ends; the resulting fully filled-out and concrete models can even be incompatible with one another. Love offers lists of model constraints from cellular and developmental biology that add additional model structure that may be needed to complement evolutionary models in order to fulfill the requirements of an extended evolutionary synthesis [42.53, p. 8]. By combining different empirical generalizations, some developmental biologists create mechanisms for the *evolution of form* that can then be tested. (The component principles of these generalizations are not, themselves, best considered mechanistic models, according to Love [42.53, p. 8]). The relevant causal mechanisms operate on the level of individual ontogeny of organisms, not on population-level mechanisms such as natural selection or genetic drift. Thus, we get the concrete-level models just mentioned. Having a flexible and robust view of theory structure that



**Fig. 42.3** The semantic view structure of evolutionary theory, with overlapping substructures in an  $n$ -dimensional space. Each oval represents a subtheory in evolutionary theory, such as genotypic selection or species selection

can account for this wide range of models is imperative, for Love. He contrasts the complex and flexible view he derives from *Griesemer's* [42.58] semantic view account, as seen in Fig. 42.3, with that of biologists *Pigliucci and Müller* [42.57].

*Pigliucci and Müller* discuss the possibility of an expanded evolutionary synthesis, which is intended to incorporate evolutionary developmental biology (evo-devo) in significant and integrated ways. They see the need for such an expansion, based on new concepts and findings in evo-devo that cannot be forced into their view of the traditional evolutionary synthesis from the 1930s and 1940s [42.57].

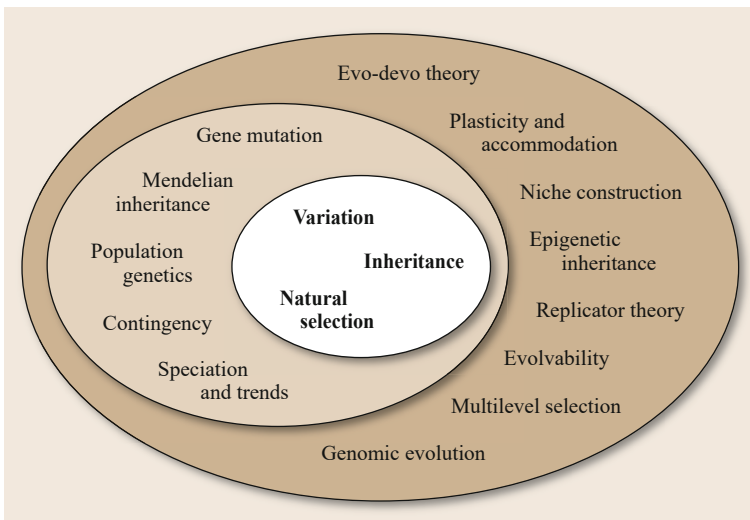
*Pigliucci and Müller* offer an onion-type diagram to show the structure of evolutionary theory: They place

*Darwinism* as a circle at the very center of the onion, which includes only natural selection, inheritance, and variation.

In the next outer ring, they place the *modern synthesis*, which includes things like *mendelian inheritance, speciation and trends*, population genetics, and gene mutation. In the last, outermost ring, which they call the *extended evolutionary synthesis*, *Pigliucci and Müller* place evo-devo theory, niche construction, multilevel selection, and genomic phenomena, etc., as shown in Fig. 42.4.

One apparent problem with this proposed structure, is: how does evo-devo theory interact with natural selection, when they are layers away in the onion? It seems that they have no integration, on this picture. Alan Love also complains that this model assumes a *core* of evolutionary population genetics, which is always in operation, even though many have no wish to make such an assumption, and may explicitly reject it [42.53, p. 6], [42.9, p. 8]. The general problem with the onion picture is thus the lack of integration and relations among its parts, which is available under analyses that approach the structure as hierarchical and intertwined families of models, which include the new evo-devo, detailed, models.

A valuable alternative approach to the structure of current evolutionary biology as well as its history is presented by *Telmo Piavani*, who offers a reconstruction inspired by philosopher *Imre Lakatos* [42.59]. *Piavani's* updated Lakatosian approach dispenses with some weaknesses in the original Lakatosian approaches, including any strong falsificationism, and analyzes contemporary evolutionary theory into a flexible structure that delineates the theory's *explanatory core* and its *protective belt*. The traditional *explanatory core* includes



**Fig. 42.4** The new, updated structure of expanded evolutionary theory, with its core and surrounding elements of theories (after [42.57])

such things as: evolution as a fact, the tree of life, common descent, and natural and sexual selection. The explanatory core of evolutionary theory serves as the basic theoretical and assumptive background when building evolutionary explanations, and can be combined with other explanatory and theoretical tools from the *protective belt*, which has expanded and contracted over time, depending on our state of biological knowledge. Currently, Piavani argues that the evolutionary core includes two hierarchies, one a genealogical hierarchy, namely of nested levels of transmission of genetic materials, organisms, demes, species, and monophyletic taxa. The other hierarchy is ecological, made up of, for example, organisms, populations like avatars, local ecosystems, and regional ecosystems, the key issue being the transfers of energy.

Using these two hierarchies in the explanatory core, Piavani then describes a “pluralistic protective belt,” that is, a set of models and assumptions, in addition

to the “explanatory core,” that may be appealed to by an evolutionary biologist when constructing an explanation [42.60, pp. 12–13]. This set of models and assumptions includes evo-devo, or evolutionary-developmental biology, which is now being appealed to in some applications of evolutionary theory. Piavani emphasizes that the theoretical changes that evolutionary theory has undergone since the modern synthesis in the 1930s and 1940s has been a progressive one, and now the theory is very healthy. The current dual core of the theory, and its various models and features of the protective belt demonstrate the growth of evolutionary knowledge since the modern synthesis (although they do have the disadvantage, according to someone like Love, of privileging some models as the *core* of evolutionary theory). Piavani also emphasizes that much of the theory is capable of *falsification*, but *has not been falsified*, which concerns the testability and confirmation of the theory, the topic of Sect. 42.2.

## 42.2 Confirmation in Evolutionary Biology

### 42.2.1 Confirming and Testing Models

In general, one evaluates a model by comparing the outcome of the model with empirical observations, and by independently testing the assumed conditions, including observations not available at the time of model construction [42.1, pp. 116–117]. In population genetics, the most obvious way to support a claim of the form *this natural system is described by this model*, is to demonstrate the simple matching of some part of the model with some part of the natural system being described. (Throughout this discussion, let us assume that each *match* of model to a natural system in the real world involves a specification of the model, by filling in variable and parameter values (on the model side), plus a determination from the real world of a *data model* into a similar format (on the real-world side). We will skip these specifications in our discussion, henceforth, and for simplicity, simply say that a *model matches or describes the natural world*.) In a population genetics model, the solution of an equation might yield a single genotype frequency value. The genotype frequency is a state variable in the model. Given a certain set of input variables (e.g., the initial genotype frequency value, in this case), the output values of variables can be calculated using the rules or laws of the model. The output set of variables (i. e., the solution of the model equation given the input values of the variables) is the *outcome* of the model. Determining the fit between model and natural system involves testing how well the genotype frequency trajectory cal-

culated from the model (the outcome of the model) matches that measured in the relevant natural (or experimental) populations. It can be evaluated by determining the fit of one curve (the model trajectory or coexistence conditions) to another (taken from the natural system); ordinary statistical techniques of evaluating curve-fitting are used.

Also, numerous assumptions are made in the construction of any model, and testing a model sometimes involves confirming some or all of these assumptions independently. These include assumptions about which factors influence the changes in the system, what the ranges for the parameters are, and what the mathematical form of the laws is. On the basis of these assumptions, the models take on certain features. Many of these assumptions thus have potential empirical content. That is, although they are assumptions made about certain mathematical entities and processes during the construction of the theoretical model, when empirical claims are then made about this model, the assumptions may have empirical significance.

For instance, the assumption might be made during the construction of a model that the population is panmictic, that is, all genotypes interbreed at random with each other.

#### **Definition 42.11 Panmictic, Panmixia**

All genotypes interbreed at random with each other.

The model outcome, in this case, is still a genotype frequency trajectory, for which ordinary curve-fitting tests

can be performed on the natural population to which the model is applied. But the model can have additional empirical significance, given the empirical claim that a natural system is a system of the kind described in the model. The assumption of panmixia, as a description of the population structure of the system under question, must be considered part of the system description that is being evaluated empirically. Evidence to the effect that certain genotypes in the population breed exclusively with each other (i. e., evidence that the population is not, in fact, panmictic) would undermine empirical claims about the model as a whole, other things being equal, and even if genotype frequencies (that is, the model outcome) provide good instances of fit. In other words, the assumption that genotypes are randomly redistributed in each generation is intrinsic to many population genetics model types. Hence, although the assumption that the population is panmictic often is taken for granted in the actual definition of the model type – that is, in the law formula – it is interpreted empirically and plays an important role in determining the empirical adequacy of the claim.

By the same token, evidence that the assumptions of the model hold for the natural system being considered will increase the credibility of the claim that the model accurately describes the natural system. Technically, we can describe this situation as follows. From the point of view of empirical claims made about it, the model has three parts: state variables, empirically interpreted background assumptions, and those aspects and assumptions of the model that are not directly empirically interpreted. Because the possible outcomes of the model (along with the inputs) are actually values of the state variables, we can understand the input and outcome of the model as a sort of minimal empirical description of the natural system.

#### **Definition 42.12 Three Parts of the Models**

State variables: such as genotypes. Empirically interpreted background assumptions: such as selection parameters, mutation parameters, relations between model parts. And those aspects and assumptions of the model that are not directly empirically interpreted.

If a model is claimed to describe accurately a natural system, at the very least, this means that the variables in which the natural system is described change according to the laws presented in the theoretical model. Under many circumstances, such models, in which only the state variables are empirically interpreted, are understood to be *mere* calculating devices, because the match with the (*data model* extracted from the) natural system is so limited.

Various aspects of the model – for instance, the form of its laws, or the values of its parameters – rely on assumptions made during model construction that can be interpreted empirically. The assumption of panmixia, discussed earlier, is an empirically interpreted background assumption of many population genetics models. Because it is empirically interpreted, the presence or absence of panmixia in the natural population in question makes a difference to the empirical adequacy of the model.

Finally, there may be aspects of the model that are not interpreted empirically at all. For instance, some parameters appearing in the laws might be theoretically determined and might have no counterpart in the natural system against which the model is compared.

Given that there can be aspects of the model structure that are not directly tested by examining the fit of the state variable curve, direct testing of these other aspects would give additional reason to accept (or reject) the model as a whole. In other words, it is taken that direct testing provides a stronger test than indirect testing, hence there is a higher degree of confirmation if the test is supported by empirical evidence. Direct empirical evidence for certain empirically interpreted aspects of the model that are not included in the state variables (and thus are confirmed only indirectly by goodness-of-fit tests), therefore provides additional support for the application of the model.

The above sort of testing of assumptions involves making sure that the empirical conditions for application of various parts of the model description actually do hold. In other words, in order to accept an explanation constructed by applying the model, the conditions for application must be verified. The specific values inserted as the parameters or fixed values of the model are another important aspect of empirical claims that can be tested. In some models, mutation rates, etc. appear in the equation – part of the task of confirming the application of the model involves making sure that the values inserted for the parameters are appropriate for the natural system being described. Finally, there is more abstract form of independent support available, in which some general aspect of the model, for instance, the interrelation between the two variables, or the significance of a particular parameter or variable, can be supported through evidence outside the application of the model itself. For example, a migration parameter can be measured independently of the application of the model by counting the numbers of organisms that move between designated populations.

Variety of evidence, of which there are at least three kinds, is an important factor in the evaluation of em-



irical claims. Three kinds of variety of evidence are discussed here:

1. Variety of instances of fit
2. Variety of independently supported model assumptions
3. Variety of types of support, which include fit and independent support of aspects of the model. In addition, we must consider the virtue of robustness of models.

First, let us consider variety of fit, that is, variety of instances in which the model outcome matches the value obtained from the natural system. Traditionally, variety of evidence has meant variety of fit. There are two issues that must be distinguished when considering the variety of fit. One involves the range of application of a model or model type, while the other involves the degree of empirical support provided for a single application. Both of these issues involve the notion of independence, which needs to be clarified before continuing.

One point of any claim to variety of evidence is to show that there is evidence for the hypothesis in question from different or independent sources. The notion of independence here is not the sort of independence that is found in the frequency interpretation of probability theory, where the independence depends only on accepting a particular reference class. Rather, in the context of scientific theories, independence is relative to whatever theories have already been accepted.

Significantly, in evolutionary theory, independence is usually relative to some assumption about natural kinds. Empirical claims are often made to the effect that a model is applicable over a certain range of natural systems. Inherent in these claims is the assumption that all of the natural systems in the range participate in some key feature or features that make it possible and/or desirable to describe them all with the same model type. Thus, there is some assumption of a natural kind (characterized by the key feature or features) whenever range of applications arises. The scientist, in making a claim that the model type is applicable to such and such systems, is making an empirical assumption about the existence of the key feature or features in the range of natural systems under question. Hence, for evolutionary models, testing for variety of fit depends on accepting an assumption about what constitutes different or independent instances of fit; this in turn, amounts to accepting a particular view of natural kinds. Part of what variety of evidence does, in a bootstrapping effect, is to confirm this original assumption about natural kinds. More technically, variety of evidence confirms the sufficiency of the parameters and state space.

Take the case in which a model type is claimed to be applicable over a more extended range than that actually covered by available evidence. This extrapolation of the range of a model can be performed by simply accepting or assuming the applicability of the model to the entire range in question. A more convincing way to extend applicability is to offer evidence of fit between the model and the data in the new part of the range. Provision of a variety of fit can thus provide additional reason for accepting the empirical claim regarding the range of applicability of the model.

For instance, a theory confirmed by 10 instances of fit involving populations of size 1000 (where population size is a relevant parameter) is in a different situation with regard to confirmation than a theory confirmed by one instance of each of ten different population sizes ranging from 1–1 000 000. If the empirical claims made about these two models asserted the same broad range of applicability, the latter model is confirmed by a greater variety of instances of fit. That is, the empirical claim about the latter model is better confirmed, through successful application (fits) over a larger section of the relevant range than the first model. Variety of fit can therefore provide additional reason for accepting an empirical claim about the range of applicability of a model. Variety of fit can also provide additional reason for accepting a particular empirical claim, that is, one of the form, *this natural system is accurately described by that model*. That is, variety of evidence can serve to increase confidence in the accuracy of any particular description of a natural system by a model.

Variety of fit is only one kind of variety of evidence. An increase in the number and kind of assumptions tested independently, that is, greater variety of assumptions tested, would also provide additional reason for accepting an empirical claim about a model. The final sort of variety of evidence involves the mixture of instances of fit and instances of independently tested aspects of the model. In this case, the variety of types of evidence offered for an empirical claim about a model is an aspect of confirmation. This kind of variety is not only related to robustness, which appears when numerous models all converge on the same result, but also is especially significant when the same model-type with a particular causal core is used to converge on the same result using a variety of assumptions, a condition that I call *model robustness*, and which warrants the conclusion that the core cause in the model-type has been confirmed by this configuration of evidence [42.61].

According to the view of confirmation sketched above, claims about evolutionary models may be confirmed in four different ways:

1. Through fit of the outcome of the model to a natural system
2. Through independent testing of assumptions of the model, including parameters and parameter ranges
3. Through a range of instances of fit over a variety of the natural systems to be accounted for by the model, through a variety of assumptions tested, and including both instances of fit and some independent support for aspects of the evolutionary model
4. Through *model robustness*, wherein the core of a model type is confirmed through robust outcomes and empirical support for the assumptions of the models.

## 42.3 Models in Behavioral Evolution and Ecology

### 42.3.1 The Phenotypic Gambit

The basic assumption underlying much behavioral biology and evolutionary theory is that the organisms under study are well adapted. This assumption amounts to the claim that many features of the organism are themselves adaptations, shaped by natural selection for their present uses. There are also many features of the organism that are byproducts of present adaptations, characteristics that Stephen Jay Gould and Elisabeth Vrba dubbed *exaptations*, or traits that may be useful, but were not adapted by natural selection for their present uses. We will continue to focus on the adaptations, for the present. We can ask, about any feature of an organism, *does this trait have a function?* And if the answer seems to be *yes*, then we can search for the specific, detailed function that it serves. Assuming that there is a genetic underpinning for a selected phenotypic trait, and that it does not matter what that genetic underpinning is called *the phenotypic gambit*. In researching traits in the population, the researcher delineates distinct *strategies* or behavioral characteristics, and then assigns fitnesses or reproductive success to those strategies. The phenotypic gambit [42.62, Ch. 3, p. 64]:

“Is to examine the evolutionary basis of a character as if the very simplest genetic system controlled it: as if there were a haploid locus at which each distinct strategy was represented by a distinct allele, as if the payoff rule gave the number of offspring for each allele, and as if enough mutation occurred to allow each strategy the chance to invade.”

Technically, the Gambit is almost always false, as few species studied by animal behaviorists are haploid, having only one set of chromosomes. But the soundness of behavioral ecology, whose “main aim is to undercover the selective forces that shape the character,” depends on ignoring population genetics, and hoping that their situations and traits do not violate their assumptions, especially the phenotypic gambit [42.62, Chap. 3, p. 64].

### 42.3.2 Evolutionary Stable Strategies, Animal Signalling

Using the phenotypic gambit, evolutionary models at the phenotypic level, usually game theoretic models, are created by animal behaviorists and behavioral ecologists to explore the evolutionary dynamics of animal behavior. W.D. Hamilton offered the rubric of his rule: if the degree of relatedness,  $r$ , times the benefit of a behavior, minus the cost of that behavior, would be greater than zero, then the behavior would evolve in the population. The benefits and costs are always counted in terms of numbers of offspring, not any other currency.

In the context of the phenotypic gambit, strategies in models borrowed from economics, in which stable equilibria are reached using simple strategies assigned to the phenotypes, are dubbed Evolutionary Stable Strategies. Such strategies are explored through optimal foraging theory, in models that represent searches for food and consumption and handling of that food in either simple or complex environments. Optimal strategies make the best use of the organisms’ net rate of energy intake and time, which are therefore optimized in the economic models adapted for animal behavior. Sometimes these optimality models are compared with experimental evidence regarding their assumptions, to good effect. For example, there is experimental evidence showing a connection between cognitive searching, a memory type of task, and the type of spatial searching done in optimal foraging theory, that are found to be evolutionary stable strategies in the models [42.63].

While *Lewontin* first suggested using game theoretical models in evolutionary biology [42.64], it was *Smith* who really established the use of such models in the field [42.55]. His analysis of adaptations using optimality models is now widely used. *Parker* and *Smith* [42.65] distinguish between general models and specific models, which are parts of a continuum. General models are used heuristically, to guide insights into the biological problem. Specific models are meant to be applied quantitatively to particular species, and

include measurable parameters. In creating an optimality model, an adaptation is assumed, and a range of strategies relating to the adaptive problem is defined, specifying the plausible alternatives. In the case of sex ratios, for example, the strategies would go from all males to half males and half females to all females. An assumption is made that Darwinian fitness is being maximized, usually the lifetime number of surviving offspring, or sometimes Hamilton's inclusive fitness, for each strategy. Assumptions are then made about the fitness consequences or payoffs of each strategy or behavior, which may require constructing mathematical models. For example, R.A. Fisher assumed, in his sex-ratio theory that all parents have equal resources to spend on sons and daughters, so more sons would be less daughters, etc.

The relationship between the fitness function and the strategy can sometimes be determined empirically, although this is often problematic. David Lack famously argued about optimal clutch size in birds, saying that bigger the clutch size, the more chicks that could be hatched. But the more chicks there are, the lower the chance that a given chick would survive, because the parent birds can only supply a limited amount of food to the fledglings. The cost to the parents can be manipulated experimentally by subtracting eggs from the nest as they are laid, to determine what the payoffs are. Ultimately, if the models do not fit, then we have misidentified the strategy set, the optimization criterion, or the payoffs, or possibly, the behavior identified is no longer adaptive. The assumptions of the model need to be revised and retested.

In an evolutionary stable strategy, individual strategies are optimized, which may not maximize fitness in a population sense. A strategy that might maximize population fitness may not be a competitive optimum or ESS (evolutionary stable strategy) because it is invadable by another strategy. Many optimality models have globally stable equilibrium solutions toward which selection is expected to converge. Thus, ESS models, as well as optimization models, may have multiple equilibria [42.65, 66].

### 42.3.3 Physiological/Evolution Models, Cognitive Ethology

Philosopher and cognitive scientist *Colin Allen* wrote in 2014 [42.67, pp. 82–83]:

“Cognitive scientists build many different kinds of models. There are various ways of classifying these models. They could, for instance, be classified according to the kinds of building blocks that they use: nodes and links for network models,

symbols and rules for computational models, probability distributions over hypotheses and conditional probabilities between hypotheses and evidence for Bayesian models, time-based differential equations for dynamical models [42.68], etc. They can also be classified according to the intended application of the model – some models seek only to capture overt behavior, while others are targeted at the mechanisms or processes underlying the behavior. Mathematical models (which, as the case of quantum mechanics shows, may also involve concepts from probability theory) use mathematical structures to provide theoretical insight and to generate empirical predictions that go beyond statistical interpolations or extrapolations of existing results.”

We see this in optimal foraging models. Optimal foraging theory involves optimality modeling of food search behavior in animals. Marmosets and tamarinds, two distinct species of monkeys, show different sets of behavior in their willingness to persist staying at a food source, in terms of maximizing returns, or patience. This makes sense in terms of the their ecology, because one eats insects, the other sap. The insects can easily be renewed, whereas when the sap is gone, it is gone. Thus, the cost to moving to a new resource versus staying at an exhausted resource is clear in each case. This is a model that involves cognitive, ecological, and evolutionary elements, thus is an example of evolutionary cognitive ethology.

Yet Fred Dyer notes that many optimal foraging models do not have any representation of the agents' own representation of their own environment, but they already know where things are in their environment. He does comparative work between species of bees to understand interactions between rewards. Bumblebees, which are diploid in their genetic systems, are individual foragers, while honeybees, which are haplodiploid, and share more of their genes with their hive-mates, are collective foragers, and are more inclined to hang around established sources to distribute the food [42.63].

Similarly, *Hills* finds physiological bases for optimal foraging and evolution, through connecting the rewards of foraging to the dopaminergic–glutamate axis in the brain [42.69]. Tracing the evolution of rewards for tasks through the brain rewards systems has enabled cognitive evolutionary ethologists to tie a variety of behaviors together. The ethologists talk about searching our memory, and offer models of that search as sharing mechanisms with a search in a spatial environment [42.70]. There is experimental evidence supporting the notion that there is a connection between cognitive searches and spatial searches [42.71].

### 42.3.4 Optimality Models Including Agent Models

Continuing the theme of the increased precision of the modeling of behavior, flocking and swarming behavior in birds and fish have been explained by the following

of very simple rules, and not bigger plans, by Craig Reynolds' boids simulations [42.72, 73]. Explain both swarming and predator-prey behavior by simple rules as well. The movement in modeling is toward more psychological modeling looking at agent behavior, rather than just strategies.

### References

- 42.1 E.A. Lloyd: The nature of Darwin's support for the theory of natural selection, *Philos. Sci.* **50**, 112–129 (1983)
- 42.2 J.R. Griesemer: Presentations and the status of theories. In: *PSA 1984*, Vol. 1, ed. by P.D. Asquith (Philosophy of Science Association, East Lansing 1984) pp. 102–114
- 42.3 B.C. Van Fraassen: *The Scientific Image* (Oxford Univ. Press, Oxford 1980)
- 42.4 E.A. Lloyd: Units and levels of selection. In: *Stanford Encyclopedia of Philosophy*, ed. by E.N. Zalta (Stanford Univ., Stanford 2005), <http://plato.stanford.edu/archives/win2012/entries/selection-units/>, Winter 2012 edn.
- 42.5 J. Beatty: What's wrong with the received view of evolutionary theory? In: *PSA 1980*, Vol. 2, ed. by P.D. Asquith, R.N. Giere (Philosophy of Science Association, East Lansing 1980) pp. 397–426
- 42.6 P. Thompson: The structure of evolutionary theory: A semantic perspective, *Stud. Hist. Philos. Sci.* **14**, 215–229 (1983)
- 42.7 J. Beatty: Chance and natural selection, *Philos. Sci.* **51**, 183–211 (1984)
- 42.8 E.A. Lloyd: A semantic approach to the structure of population genetics, *Philos. Sci.* **51**, 242–264 (1984)
- 42.9 E.A. Lloyd: *The Structure and Confirmation of Evolutionary Theory* (Princeton Univ. Press, Princeton 1994)
- 42.10 P. Suppes: Models of data. In: *Logic, Methodology, and Philosophy of Science*, ed. by E. Nagel, P. Suppes, A. Tarski (Stanford Univ. Press, Stanford 1962) pp. 252–261
- 42.11 R. Giere: *Explaining Science: A Cognitive Approach* (Univ. Chicago Press, Chicago 1988)
- 42.12 C. Alsina, R.B. Nelsen: *Math made Visual: Creating Images for Understanding Mathematics* (Mathematics Association of America, New York 2006)
- 42.13 H. Meyer-Ortmanns, S. Thurner (Eds.): *Principles of Evolution* (Springer, Berlin, Heidelberg 2011)
- 42.14 R. Dawkins: *The Extended Phenotype* (Oxford Univ. Press, Oxford 1982)
- 42.15 E.A. Lloyd, R.C. Lewontin, M. Feldman: The generational cycle of state spaces and adequate genetical representation, *Philos. Sci.* **75**, 140–156 (2008)
- 42.16 R.C. Lewontin: *The Genetic Basis of Evolutionary Change* (Columbia Univ. Press, New York 1974)
- 42.17 T. Dobzhansky: *Genetics of the Evolutionary Process* (Harper Rox, New York 1970)
- 42.18 R. Lewontin: Population genetics, *Annu. Rev. Genet.* **1**(1), 37–70 (1967)
- 42.19 R. Levins: *Evolution in Changing Environments* (Princeton Univ. Press, Princeton 1968)
- 42.20 M. Kimura, T. Ohta: Protein polymorphism as a phase of molecular evolution, *Nature* **229**, 467–469 (1971)
- 42.21 J.H. Gillespie: *Population Genetics: A Concise Guide*, 2nd edn. (Johns Hopkins Univ. Press, Baltimore 2004)
- 42.22 E. Mayr: Evolutionary challenges to the mathematical interpretation of evolution. In: *Mathematical Challenges to the Neo-Darwinian Interpretation of Evolution*, ed. by P.S. Moorehead, M.M. Kaplan (Wistar Institute, Philadelphia 1967) pp. 47–54
- 42.23 M. Wade: A critical review of the models of group selection, *Q. Rev. Biol.* **53**, 101–114 (1978)
- 42.24 M.J. Wade: *Adaptation in Metapopulations* (Univ. Chicago Press, Chicago 2016)
- 42.25 E.A. Lloyd: Confirmation of evolutionary and ecological models, *Biol. Philos.* **2**(3), 277–293 (1987)
- 42.26 R.A. Skipper Jr.: The heuristic role of Sewall Wright's 1932 adaptive landscape diagram, *Philos. Sci.* **71**(5), 1176–1188 (2004)
- 42.27 P. Forber: On biological possibility and confirmation, unpublished manuscript (2008)
- 42.28 P. Forber: Confirmation and explaining how possible, *Stud. Hist. Philos. Sci. Part C: Stud. Hist. Philos. Biol. Biomed. Sci.* **41**(1), 32–40 (2010)
- 42.29 L.A. Dugatkin, H.K. Reeve: Behavioral ecology and levels of selection: Dissolving the group selection controversy, *Adv. Study Behav.* **23**, 101–133 (1994)
- 42.30 K. Sterenly: Explanatory pluralism in evolutionary biology, *Biol. Philos.* **11**(2), 193–214 (1996)
- 42.31 E. Sober, D.S. Wilson: *Unto Others: The Evolution and Psychology of Unselfish Behavior*, (Harvard Univ. Press, Cambridge 1998), p. 57, pp. 98–99
- 42.32 K. Sterenly, P.E. Griffiths: *Sex and Death: An Introduction to Philosophy of Biology* (Univ. Chicago Press, Chicago 1999) pp. 168–169, p. 172
- 42.33 B. Kerr, P. Godfrey-Smith: Individualist and multi-level perspectives on selection in structured populations, *Biol. Philos.* **17**(4), 477–517 (2002)
- 42.34 C.K. Waters: Why genic and multilevel selection theories are here to stay, *Philos. Sci.* **72**, 311–333 (2005)
- 42.35 R.D. Alexander, G. Borgia: Group selection, altruism, and the levels of organization of life, *Annu. Rev. Ecol. Syst.* **9**, 449–474 (1978)
- 42.36 M.K. Uyenoyama, M.W. Feldman: Evolution of altruism under group selection in large and small populations in fluctuating environments, *Theor.*

- Popul. Biol. **17**, 380–414 (1980)
- 42.37 D.S. Wilson: *The Natural Selection of Populations and Communities* (Benjamin Cummings, Menlo Park 1980)
- 42.38 R.K. Colwell: Evolution of female-biased sex ratios: The essential role of group selection, *Nature* **290**, 401–404 (1981)
- 42.39 J.F. Crow, K. Aoki: Group selection for a polygenetic behavioral trait: A differential proliferation model, *Proc. Natl. Acad. Sci. USA* **79**, 2628–2631 (1982)
- 42.40 R.E. Michod: The theory of kin selection, *Annu. Rev. Ecol. Syst.* **13**, 23–55 (1982)
- 42.41 M.J. Wade: Soft selection, hard selection, kin selection, and group selection, *Am. Nat.* **125**, 61–73 (1985)
- 42.42 M. Smith: Evolutionary progress and levels of selection. In: *The Latest on the Best: Essays on Evolution and Optimality*, ed. by J. Dupre (MIT Press, Cambridge 1987)
- 42.43 D.C. Queller: Quantitative genetics, inclusive fitness, and group selection, *Am. Nat.* **139**, 540–558 (1992)
- 42.44 P. Godfrey-Smith, R.C. Lewontin: The dimensions of selection, *Philos. Sci.* **60**, 373–395 (1993)
- 42.45 M.W. Feldman, L.L. Cavalli-Sforza: Cultural and biological evolutionary processes, selection for a trait under complex transmission, *Theor. Popul. Biol.* **9**(2), 238–259 (1976)
- 42.46 M.W. Feldman, L.L. Cavalli-Sforza: *Cultural Transmission and Evolution: A Quantitative Approach (No. 16)* (Princeton Univ. Press, Princeton 1981)
- 42.47 E. Jablonka, M.J. Lamb: *Evolution in four dimensions: Genetic, Epigenetic, Behavioral, and Symbolic Variation in the History of Life*, *Life and Mind: Philosophical Issues in Biology and Psychology* (MIT Press, Cambridge 2005)
- 42.48 F. Suppe: What's wrong with the received view on the structure of scientific theories?, *Philos. Sci.* **39**, 1–19 (1972)
- 42.49 F. Suppe: *The Structure of Scientific Theories*, 2nd edn. (Univ. Illinois Press, Urbana 1977)
- 42.50 F. Suppe: Understanding scientific theories: An assessment of developments, *Philos. Sci.* **67**, S102–S115 (2000)
- 42.51 S.M. Downes: *The Importance of Models in Theorizing: A Deflationary Semantic View, PSA*, Vol. 1 (Univ. Chicago Press, Chicago 1992) pp. 142–153
- 42.52 P. Godfrey Smith: The strategy of model-based science, *Biol. Philos.* **21**, 725–740 (2006)
- 42.53 A. Love: Theory is as theory does: Scientific practice and theory structure in biology, *Biol. Theory* **7**(4), 325–337 (2012)
- 42.54 S. Scheiner: Toward a conceptual framework for biology, *Q. Rev. Biol.* **85**, 293–318 (2010)
- 42.55 J.M. Smith: *The Evolution of Sex* (Cambridge Univ. Press, Cambridge 1978), No. 574.1 S5
- 42.56 A. Love: Rethinking the structure of evolutionary theory for an extended synthesis. In: *Evolution – The Extended Synthesis*, ed. by M. Pigliucci, G.B. Müller (MIT Press, Cambridge 2010) pp. 403–441
- 42.57 M. Pigliucci, G.B. Müller: Elements of an extended evolutionary synthesis. In: *Evolution – The Extended Synthesis*, ed. by M. Pigliucci, G.B. Müller (MIT Press, Cambridge 2010) pp. 3–17
- 42.58 J. Griesemer: Presentations and the status of theories, *PSA: Proc. Bienn. Meet. Philos. Sci. Assoc.* (1984) pp. 102–114
- 42.59 I. Lakatos: *The Methodology of Scientific Research Programmes. Philosophical Papers*, Vol. 1 (Cambridge Univ. Press, Cambridge 2011)
- 42.60 T. Piavani: An evolving research programme: The structure of evolutionary theory from a Lakatosian perspective. In: *The Theory of Evolution and Its Impact*, ed. by A. Fasolo (Springer, Milan 2012) pp. 211–228
- 42.61 E.A. Lloyd: Model robustness as a confirmatory virtue: The case of climate science, *Stud. Hist. Philos. Sci.* **49**, 58–68 (2015)
- 42.62 A. Grafen: Natural selection, kin selection and group selection. In: *Behavioural ecology: An Evolutionary Approach*, Vol. 2, ed. by J.R. Krebs, N.B. Davies (Blackwell Scientific Publications, Oxford 1984) pp. 62–84
- 42.63 J.M. Townsend-Mehler, F.C. Dyer, K. Maida: Deciding when to explore and when to persist: A comparison of honeybees and bumblebees in their response to downshifts in reward, *Behav. Ecol. Sociobiol.* **65**, 305–312 (2011)
- 42.64 R. Lewontin: Interdeme selection controlling a polymorphism in the house mouse, *Am. Nat.* **96**, 65–78 (1962)
- 42.65 G.A. Parker, J.M. Smith: Optimality theory in evolutionary biology, *Nature* **348**, 27–33 (1990)
- 42.66 E.A. Lloyd, M.W. Feldman: Commentary: Evolutionary psychology: A view from evolutionary biology, *Psychol. Inq.* **13**(2), 150–156 (2002)
- 42.67 C. Allen: Models, mechanisms, and animal minds, *South. J. Philos.* **52**, 75–97 (2014)
- 42.68 C. Buckner: Two approaches to the distinction between cognition and 'mere association', *Int. J. Comp. Psychol.* **24**(4), 314–348 (2011)
- 42.69 T.T. Hills: Animal foraging and the evolution of goal-directed cognition, *Cogni. Sci.* **30**, 3–41 (2006)
- 42.70 T. Hills, P.M. Todd, R.L. Goldstone: Search in external and internal spaces: Evidence for generalized cognitive search processes, *Psychol. Sci.* **19**, 676–682 (2008)
- 42.71 P.M. Todd, T.T. Hills, T.W. Robbins: Building a foundation for cognitive search. In: *Cognitive Search: Evolution, Algorithms, and the Brain*, Vol. 9, ed. by P.M. Todd, T.T. Hills, T.W. Robbins (MIT Press, Cambridge 2012) pp. 1–7, Strüngmann Forum Reports
- 42.72 R.S. Olson, A. Hintze, F.C. Dyer, D.B. Knoester, C. Adami: Predator confusion is sufficient to evolve swarming behavior, *J. R. Soc. Interface* **10**, 20130305 (2013), <http://dx.doi.org/10.1098/rsif.2013.0305>
- 42.73 C. Reynolds: Flocks, herds, and schools: A distributed behavioral model, *Comput. Graph.* **21**(4), 25–34 (1987)

## 43. Models and Mechanisms in Cognitive Science

Massimo Marraffa, Alfredo Paternoster

In this chapter, we present and discuss models in the context of cognitive sciences, that is, the sciences of the mind. We will focus on computational models, which are the most popular models used in the disciplines of the mind.

The chapter has three sections. In the first section, we explain what is a computational model, give a pair of examples of it, illustrate some crucial concepts related to this kind of models (simulation, computational explanation, functional explanation, and mechanicism) and introduce a class of partially alternative models: dynamical models. In the second section, we discuss a pair of difficulties faced by computational explanation and modeling in cognitive sciences: the problem raised by the constraint of modularity, and the problem of the allegedly required integration between dynamical and computational models. Finally, in the third section, we provide a short recap.

43.1	<b>What is a Model in Cognitive Science?</b>	929
43.1.1	Computational Models .....	929
43.1.2	An Example of Computational Model...	932
43.1.3	Function and Functional Explanation..	934
43.1.4	Computational Models and Mechanistic Explanations .....	936
43.1.5	Dynamical Systems .....	939
43.2	<b>Open Problems in Computational Modeling</b> .....	940
43.2.1	Computationalism and Central Cognition .....	940
43.2.2	The Dynamicist Challenge: Is Integration Possible? .....	944
43.3	<b>Conclusions</b> .....	948
	<b>References</b> .....	949

### 43.1 What is a Model in Cognitive Science?

Cognitive science is the project of interdisciplinary study of natural and artificial intelligence that begins its maturation in the late 1950s and reaches a stable intellectual and institutional set-up in the early 1980s [43.1]. One point is worth emphasizing. Cognitive science (as a general framework) is the study of the mind as an information processing system, that is, a system behaving on the basis of detected and properly elaborated (in a variety of ways) information; yet research in cognitive science is typically about a *specific* type of information processor – for example, cognitive neuroscience investigates the *biological* processor, whereas artificial intelligence explores the *artificial* one. Therefore, cognitive science is better to be conceived of not as a discipline, but rather as a *doctrine* that has oriented inquiries in a number of disciplines (see [43.2, p. 521] and [43.3, p. 18]), some descriptive and empirical (e.g., cognitive psychology, linguistics and, more recently, neuroscience), some speculative and founda-

tional (e.g., philosophy), and some both speculative and applied (e.g., artificial intelligence) [43.4].

Although cognitive scientists are in agreement in considering the mind as an information processing system, their different disciplinary backgrounds lead them to make use of a large variety of approaches, methods, research styles and, for what we are concerned here, *models*. It suffices to mention symbolic systems, artificial neural networks, dynamical systems, robot-style artifacts, and so on.

#### 43.1.1 Computational Models

Yet, if we categorize models according to a sufficiently general (and arguably more useful) criterion, models in cognitive science are essentially of two kinds: mathematical and computational. (It seems reasonable to say that computational models are a subclass of mathematical models. However, it is a so relevant subclass that

it does make sense to assess them separately [43.5, p. 20]). In this chapter, we will mainly be concerned with computational models. We focus on these models for three reasons. First, we take computational models to be the best representative of the research in the discipline. This is confirmed by the fact that over 80% of articles in theoretical cognitive science focus on computational modeling [43.6]. Indeed, it could be argued (with some caution that will be justified later, see mainly Sect. 43.2) that the development of computational models is a *definitional* factor of the discipline [43.7]. Second, the class of computational models is large enough to cover many, arguably most, of the cases. Third, many among the most discussed foundational problems in cognitive science arise exactly from the role ascribed to computation. In other words, the future of the discipline depends crucially on the possibility of extending the use of computational models to a variety of further cognitive capacities and processes, and defending the computational approach from criticisms addressed by the supporters of the so-called *radical embodied cognition* [43.8] (see Sect. 43.2.1, *The Massive Modularity Hypothesis*).

The notion of computation – or algorithm – formalized by Turing [43.9], Church [43.10] and others, consists in an effective procedure to solve a given problem, that is, a finite sequence of elementary and totally explicit (= well defined and not ambiguous) instructions. In order to give a simple and concrete idea of what computations are, it is usual to mention cooking recipes for dummies: instead of saying, for example, *brown the onions*, the recipe specifies every possible elementary constitutive step of this action (*take a pan, put inside a pair of spoons of oil, put the pan on slow fire, add the onions, . . . , etc.*). Clearly, however, the paradigm cases of computation are computer programs. The leading idea in computational modeling is indeed that there is an interesting sense in which mental processes (such as perception, language understanding, reasoning, etc.) can be described as computer programs. In other words, some relevant aspects of mental processes are captured by certain features of computations.

As the reference to computer programs suggests, the notion of computation is closely related to the notion of information processing system; and this is indeed the main reason for the success of computational modeling and explanation in cognitive science: the notion of computation provides the *natural* way to develop the core idea of cognitive science, namely, that agents or their minds are regarded as information processing systems. Shortly put, taking seriously (quite literally) the idea that mental processes are computations is tantamount to saying that minds are information processing systems.

Weisberg [43.5, p. 7] defines *computational models* as sets of procedures that can potentially stand in relation to a computational description of the behavior of a system. The procedures constitutive of the model “take a starting state as an input and specify how this state changes, ultimately yielding an output” [43.5, p. 30]. Two points are to be highlighted in this definition. First, models are sets of *procedures*. It is mandatory, therefore, that the model is constituted of algorithms. However, to say that computational models are simply sets of procedures is arguably too strong. Models usually include some theoretical hypotheses that are not necessarily specified as parts of the algorithms. Weisberg seems to acknowledge this point when he says that “the procedure itself is the core component of the model, the structure in virtue of which parts of a target can be explained” [43.5, p. 30]. Therefore, procedures are *components* of the model, rather than being the model itself. Second, the model is in relation to a *computational description*. This means that a computational model is appropriate just in case the modeled system has a computational nature. What is usually intended by this claim is that the execution of one or more computations is causally responsible of the behavior of the system (see Sect. 43.1.1, *Computational Models and Computational Explanations*).

### Simulation

An execution of the *model-computation* (i. e., of the procedures constituting the model) is a *simulation*, an epistemologically crucial notion. Indeed, the results of simulation allow scientists to check whether a certain hypothesis embedded in the model concerning a cognitive ability or process is confirmed or falsified.

Vivid instances of the simulative method are robotic artifacts. For example, Grasso et al. [43.11] have studied the behavior of lobsters and developed a computational model to check the prediction that lobsters are able to locate sources of food through a very simple, rough mechanism to track turbulent odor plumes to their source (chemotaxis): the more a chemoreceptor is stimulated, the more is directly triggered the speed of the contralateral locomotive organ. Scientists constructed a biomimetic robot (RoboLobster) that works this way by implementing two algorithms:

1. Move forward when the gradient is below a minimal threshold; turn toward the side of the sensor detecting higher chemical concentration.
2. The same as algorithm 1 with one additional rule: back up when both sensors detect the lack of chemical substance.

Then RoboLobster was immersed in a stream of water to which a turbulent plume was added to check

whether it was able, in a variety of conditions, to reach the appropriate location. It turned out, however, that the robot failed in most trials. In trials running algorithm 1, it consistently failed to hit the source regardless of whether it began its trial 60 or 100 cm downstream from the source or ran in the normal or reverse sensor configuration; on average the robot approached the source much more when it started 60 cm from the source compared to the other conditions. In trials running algorithm 2, the robot never hit the source with reversed sensors; better results were obtained at the 60 cm, but not at the 100 cm starting distance, with sensors in the forward connectivity configuration. This allowed scientists to conclude that the hypothesized mechanism was not appropriate, too rough.

This case effectively shows that simulation is a very important empirical method in cognitive science. It is indeed part and parcel of cognitive science the idea that a good empirical practice for studying a phenomenon is trying to *reproduce* it. If the behavior, or certain behavioral abilities, of an organism are reproduced by an artifact or by a computer program, we have a reason to believe that the computational model is an explanation of the behavior of that system. Of course, this claim should be taken with much prudence. The legitimacy of regarding simulation as a kind of explanation is a crucial epistemological issue, and the simulative methodology calls undoubtedly for some methods of validation or evaluation [43.12, 13]. Nevertheless, it is hard to deny that the simulative approach has already demonstrated to be useful.

### Computational Models and Computational Explanations

Before providing another, more complex example of computational model, let us still spend some words on the distinction between computational models and computational explanations (or computational theories). This is important in order to assess properly the role of computational models.

According to *Piccinini* [43.14, 15], we have a computational *explanation* when the best way of accounting for the behavior of a certain system is to say that it performs a certain computation, that is, it is in virtue of the fact that the system computes a certain function that the system behaves the way behaves (as *Piccinini* puts it, the behavior of the system is *causally* explained by the computations it performs). For instance, it can hardly be denied that the behavior of a pocket calculator is explained by the fact that it computes certain (mathematical) functions; or that the behavior of a robot mounting a wheel on a car's body causally depends on computing the appropriate trajectory from the location of the stack of wheels to the hub (note that in

this case, as well as in many others, the behavior itself is a computational process, rather than being the result of a computation). Of course, it is not always so apparent that the behavior of a system is determined by a computation. In the most interesting cases, human and animal cognition among them, it has to be argued that describing the behavior of a system as the result of a computational process has an explanatory pay-off (more on this below).

In computational *modeling*, instead, a computation *C* is used to describe the behavior of another system *S* under certain conditions. *C* allows to generate subsequent descriptions of *S*, but the explanation of *S*'s behavior is based on *S*'s properties, not on features of *C*. As *Piccinini* points out [43.14, p. 96]:

“the situation is fully analogous to other cases of modeling: just as a system may be modeled by a diagram or equation without being a diagram or equation in any interesting sense, a system may be modeled by a computing system without being a computing system in any interesting sense.”

In light of these definitions, neither computational explanation implies computational modeling nor computational modeling of (the behavior of) a certain system implies a computational explanation of it, even if, generally speaking, given a computational explanation it is possible to derive a computational model from it. This characterization of computational models, however, is scarcely interesting for our goals. In fact, when we say, as cognitive scientists, that the mind is a collection of information processing systems, we are committed to the most theoretically interesting sense of this claim, that is, that the appeal to computation *explains* intelligent behavior. We suggest, therefore, modifying the definition of computational modeling, in such a way that computational models in cognitive science are intended to be descriptions of genuine computational systems, in the above-specified sense. Accordingly, one reasonable definition could be the following: a system  $S^*$  is a computational model of a computational system *S* when there is at least partial isomorphism between the behavior of  $S^*$  and the behavior of *S* (or vice-versa), so that (the behavior of) one of the two systems can be said to represent (the behavior of) the other. The idea is that a computational system (= the model) is used to reproduce some relevant aspects of the behavior of another (alleged computational) system. In other words, computational modeling *presupposes* computational explanation insofar as, when one builds a computational model, he is assuming from the start that the modeled system or phenomenon admits a computational explanation (at a certain level of description).



Of course, this leaves aside the aforementioned difficult problem of establishing whether a certain system admits a computational explanation – whether the system is genuinely computational, to put it shortly. One should not be misled by this latter formulation: the issue is epistemological, not metaphysics. As said earlier, the point is whether there is an explanatory pay-off; there is no question of discovering an objective fact of the matter establishing that the behavior of an organism is determined by computations. Yet, this does not mean that everything (every phenomenon, even every *thing* on earth!) could be described as a computation, as John Searle famously claims, making in this way the notion of computation completely vacuous. We do not want to say that, for instance, a rock, or a gas, are computational systems just because their dynamics could be modeled by computations. In order to escape this consequence, different constraints have been proposed on what counts as a computational system. Just to mention two of them, there is the *semantic* constraint, according to which something is a computational system only if it manipulates *representations*; and there is the *mechanistic* constraint, according to which something is a computational system only if it is a functional mechanism of a special kind, that is, (roughly), it has a complex, multilayered organization apt to cause its behavior (cf. [43.15]; the notion of mechanism is presented in Sect. 43.1.4). Both the constraints play an important role in the discussion on computational explanation (see Sects. 43.1.4 and 43.2), yet from now on we shall take for granted that only *some* phenomena, among which at least some cognitive capacities and processes, admit a computational explanation, without addressing the problem of what are the appropriate constraints that a system must satisfy in order to be regarded as computational in a nonvacuous sense – i. e., without making all phenomena computational (for convincing replies to Searle, see, for example, [43.14–16], [43.17, Chap. 3]). In the second section, we shall discuss some difficulties faced by the claim that minds are computational systems.

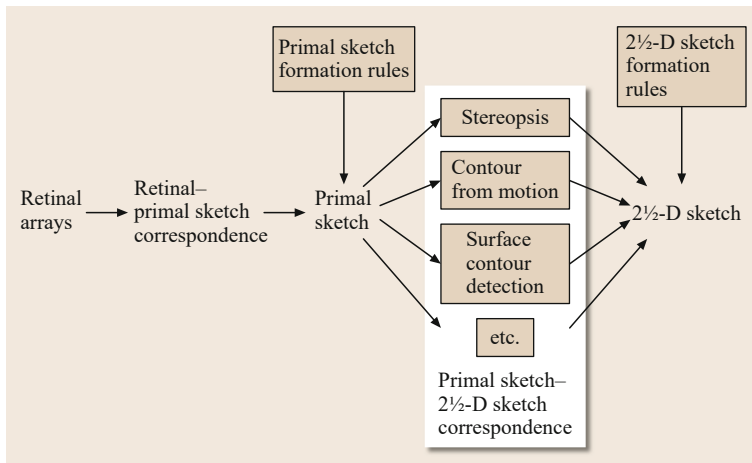
### 43.1.2 An Example of Computational Model

Computational models in cognitive science include at least classical symbolic models, that is, sequential algorithms working on data structures specified in some linguistic code, (artificial) neural networks, and artifacts-based models (robots). In order to give a concrete idea of computational modeling, here we present a classical symbolic model: Marr’s computational vision [43.18, 19]. Actually, what we are going to present is better to be called a computational *theory*, rather than a computational model; however, especially in light of

what we said in the previous section, the distinction is not much relevant. Indeed, Marr’s theory includes, as we shall see, a computational model. Our choice of Marr’s theory as an exemplar is motivated by three reasons. First, it has a formal, elegant definition and is extremely clear. Second, it is a paradigm case of (classical) cognitive science and third it provides an excellent starting point for the discussion in the following sections.

The target cognitive capacity of Marr’s theory is animal vision, more specifically a kind of vision typical of human beings and other superior animals. Vision is defined as a very complex computational system whose goal is to transform the pattern of light impinging retinas (the *input*) into a symbolic description of the observed scene (the *output*) – the collection of objects and properties present in the visual field of the observer.

Of course, this is an extremely complex task. For this reason, the task has to be decomposed in subtasks (each decomposable, in turn, in simpler tasks), so that we could say that each subtask is realized by a computational subsystem or mechanism (for the notion of mechanism, see Sect. 43.1.4). In particular, Marr identified three main subtasks: edge processing, surface processing, and object processing. Edge processing consists in the detection of strong variations of luminance. Luminance is the intensity of light in a retinal point. Therefore, a retinal image can be described as a function that associates to every point  $(x, y)$  of the retina its value of luminance. Surface processing consists in the determination, for each outgoing direction from the observation point, of the distance and orientation of the surface reflecting light in that particular direction. Object processing consists in the recognition of the (geometrical) shape of the objects present in the viewed scene. To put it simply, the first two processing stages or levels have the function of localizing an object, seen from a particular, egocentric, point of view (the observer’s perspective); the last stage has the goal of recognizing the object as an object possessing a certain shape. The shape is defined in an object-centered system of coordinates, that is, independently of the point of view. Each processing level produces an output, that is, a data structure containing the information relevant to the goal or function proper to that level. The output of the first stage is called *primal sketch*; the output of the second stage is called *2½-D sketch*; the output of the third stage (and of the whole system as well) contains one or more *three-dimensional (3-D) models*. As said earlier, each subtask admits further decomposition, not necessarily in serial tasks. Surface processing, for example, admits several *parallel* processing stages, each based on a different kind of information: shading, texture, movement etc. (Fig. 43.1).



**Fig. 43.1** The paths of information up to the  $2\frac{1}{2}$ -D sketch in Marr's architecture of the human visual system (after [43.20])

The first stage of processing occurs in the retina and the superior colliculus of human beings and other mammals. Its first product is the *raw primal sketch*, a description of the intensity changes in an image that is constructed using a primitive language of edge segments, bars, blobs, and terminations. From this, the *full primal sketch* is computed, namely a representation of the two-dimensional geometry of the field of view. The primal sketches of the visual fields of the two eyes are combined by the process of binocular fusion. The next stage in this process is the computation of the  $2\frac{1}{2}$ -D sketch, which represents the geometry of the surfaces visible to the observer – including their contours, depth, and orientation. Finally, the  $2\frac{1}{2}$ -D sketch is input to processes that construct the *3-D object-centered shape representation* that is vision's goal (Fig. 43.1).

What best qualifies Marr's model is another decomposition, vertical rather than horizontal. It is the idea that each task or subtask admits three different levels of description, all necessary in order to give an explanatorily adequate account of a given task or subtask: from the top, the computational level, the algorithmic level, and the implementation level. As we shall see, the articulation in a vertical collection of layers (the *stack*) is also a crucial characteristic of mechanistic explanations, which have indeed much in common with computational explanation (although they are not one and the same thing, see Sect. 43.1.4).

The computational level describes in the most general terms the *function* or cognitive task (or capacity) of the system to be modeled (and of its subsystems as well). This means providing a description of *what* the system does and what are the input and the output of the system. We already gave this specification for the entire system (the whole visual system) and for its immediate subsystems. Note that the individuation of the function is by no means easy, even at the highest level:

it cannot be taken for granted that the goal of vision – what the visual system does – is to individuate the (geometrical) shape of objects. Indeed many critics have pointed out that the function of the visual system, its basic goal, is allowing the agent to move effectively in the environment. Another important point concerns the vocabulary of the computational level. We could say that it is a macro-psychological, intentional vocabulary, in the sense that describes the mental capacities as relations between the agent and certain environmental properties (such as the surfaces present in the environment, or the shape of the object), and does not make use of any mathematical or (neuro)biological notions. At this level, the focus is on a certain capacity of an agent to do something in its environment.

The algorithmic level is what makes Marr's model specifically computational (despite his deserving the term *computational* to the upper level: Other, arguably better, labels for Marr's computational level have been proposed, such as *task level*, *project level*, *intentional systems level*.) Indeed, it is at this level that the programs that compute or realize functions postulated at the upper level are specified. For instance, the algorithm that processes edges (as we saw earlier, this is the first processing stage) is the Laplacian operator of the Gaussian function ( $\nabla^2 G$ ). Without entering in technical details, the idea is, intuitively, that a filter (the Gaussian one) is applied to the luminance function which describes the retinal image, and then the zero-points of the second derivative of the obtained image are calculated, since they correspond to the highest differences in luminance, namely to edges. For this reason, the algorithm is called *zero-crossing*. The role of the Gaussian filter is to *clean* the image, making more apparent the strongest bright/dark discontinuities.

Importantly, at the algorithmic level *representations* are also specified, that is, the input and the output of the

subsystem must be coded in a certain way in order to be manipulated by the algorithm. Other representations are also involved, since algorithms need further information to work successfully. Actually they embody certain assumptions on how the world is, which work as constraints restricting the space of the available solutions to the problem addressed by the algorithm. For instance, the visual system assumes that light comes always from above (although this could turn out to be false in a few cases).

Finally, the implementation (or hardware) level is constituted by the specification of the neural structures charged to execute the procedures specified at the algorithmic level. At the time Marr proposed his model (between 1976 and 1980), neuroscientific knowledge was not so developed, and the appeal to cerebral data in a model of a cognitive capacity was quite rare. Yet Marr proposed that the  $\nabla^2 G$  function is realized by off- and on-centred *X-cells* in the retina and lateral geniculate body, whereas some *simple cells* in striate cortex detect and represent zero-crossing segments [43.21, 22].

Thus, Marr was a pioneer of the current systematic attempt to integrate computational models with the specification of their neural implementation. Note that even though, strictly speaking, neural implementation is not part of the computational model, yet it is part of a mechanism that, taken as a whole, is computational (cf. Sect. 43.1.4). In other words, it is correct to say that the brain is a computational system, even if spelling out the computations performed by the brain is not something that can be done in a neurophysiological vocabulary.

Today, not only integrating computational models with brain data has become customary, but, often, brain data (neuropsychological and especially neuroimaging evidence) are even the starting point of the development of a computational model. We have here in mind what Zawidzki and Bechtel [43.23] call *interactive* view of explanation in cognitive neuroscience. On one hand, the functional knowledge obtained through psychological research allows us to identify the neural mechanisms, on the other hand the knowledge of structure is a heuristic guide to the development of more sophisticated psychological models [43.24]. In this context, *computational neuroscience* can be viewed as a *bridge* discipline between psychology and neuroscience which, on the one hand, puts bottom-up constraints on computational modeling, while on the other hand extends some principles of computational modeling to neuroscientific research, thus promoting the integration of neuroscientific theoretical constructs into computational psychology [43.25].

In the specific case of vision, a great impulse to the development of a neurophysiologically grounded com-

putational theory was the empirical discovery of two visual paths in the brain [43.26, 27], the *dorsal* one, which projects from the primary visual cortex (V1) to the parietal posterior cortex, and the *ventral* one, going from V1 to the infero-temporal cortex. This neurophysiological distinction grounds a corresponding functional distinction: the dorsal stream is associated with the visual control of action, having a *pragmatic* function, whereas the ventral stream realizes the identification of objects (*epistemic function*).

Based on this finding, Jacob and Jeannerod [43.28] put forward a dual computational model, one for each visual subsystem: the computational model of our capacity to identify objects, which is roughly based on a Marr's style architecture, and the computational model of our capacity to act effectively in the world, and specifically grasp objects, which is based on a sort of anticipatory scheme. The idea is that the dorsal stream codes (*represents*) only the pragmatically relevant features of the object, what J.J. Gibson called its *affordances* (i. e., the object's presenting itself as something that can be grasped, or can be filled, etc.), and processes this information so to produce, as output, a procedure that triggers the appropriate movement. The position of the object, which is crucial information in order to accomplish the task, is coded in a system of egocentric coordinates, namely relative to the axe of the body.

This should be enough to give an idea of what a computational model is. Now, in order to set the stage for the discussion of Sect. 43.1.4, concerning mechanistic explanation, it is worth to point out that, as Marr's example admirably shows, two important notions are closely related to computational models: the concept of *function* (as well as the related concept of functional explanation) and the concept of *module*. We address the former here in the following section and the latter will be introduced in the section that follows next.

### 43.1.3 Function and Functional Explanation

We saw that a certain cognitive ability is individuated in terms of its *function* (= the goal, what is for), and functions can be characterized as computations. Vision has (or is) the function of recognizing the shape of objects present in the visual field and in order to do that the visual system performs a certain I/O transformation: given as input retinal images in a certain description, it produces as output a symbolic description of the shapes present in the environment. In this sense, we could say that the notion of function involved in cognitive science is at the same time biological (more specifically, psychological) and mathematical (more specifically, computational). We have already explained what a com-

putation is; let us focus now on the biological notion of function and its importance for cognitive science.

In light of the analysis of the notion of cause predominant in the second half of the twentieth century, causally explaining a natural phenomenon amounts to subsuming it under laws (the so-called *nomological-deductive model*). In this *epistemic* perspective, causality is nothing over and above nomological regularity. However, philosophers of science see increasingly the nomological approach as inappropriate for special sciences like biology and psychology.

In most cases, the biologists' explanatory practices rest not on the concept of law, but rather on the notion of function. Mayr [43.29] has argued, however, that there are at least two ways of making biology – functional biology and evolutionary biology – which are very different from each other but co-exist and use different concepts and methods. In evolutionary biology, functions are conceived of in *etiological* terms, that is, they are characterized in terms of their history of natural selection (the *why* or *how come* question). Functional attributions are intended to account for the existence or maintenance of a trait in a given population, and functions are the effects of those traits that, by increasing fitness, have been favored by natural selection [43.30, 31]. In contrast, in anatomy and physiology the word *function* refers usually to the activities that an organism can perform, for instance, flying, digesting, finding viruses in one's own tissues, etc. Therefore, functional biology explains *how* organisms are able to do all that by means of what Cummins [43.32] termed *functional analysis*, where a certain capacity  $C$  is decomposed in a collection of (simpler) subcapacities  $S_1, S_2, \dots, S_n$ , in such a way that  $C$  emerges as a *programmed manifestation* of the exercise of  $S_1, \dots, S_n$ . In this context, the function of a trait is seen as the contribution that it makes to a given capacity of the system incorporating it. Such a contribution is termed *causal role function*. A causal role function is nonhistorical, that is, it ignores evolutionary history. The heart's function is pumping blood not by virtue of its history of natural selection, but because it is a part of a larger system, the circulatory system, in which it plays a crucial causal role. It is to be noticed, however, that in this case the causal-role function of a trait coincides with its etiological function – the heart pumps blood, and that is what it was selected to do. In other cases, the two kinds of functions diverge: it is very likely that birds' feathers evolved as a mechanism for regulating body temperature but their causal-role function is what they make for their owner in the present and in the future, namely to make flight possible [43.33, p. 223].

The distinction between causal role and etiological functions is also relevant for the philosophy of cognitive

science, since it may impart a different trajectory, internalist or externalist, to a *functionalist* theory of mind. Functionalism is indeed a very influential philosophical theory (or family of theories) concerning the nature of mental phenomena. At least in one of its versions, it is closely linked to computational explanation.

Functionalism is first and foremost a metaphysical theory, which characterizes psychological states according to the causal roles they play in a system (i. e., an agent's inner life), independently of how such roles are physically realized; or equivalently, the identity of a certain type of psychological state is established by the causal relations it entertains with stimulus input and behavioral outputs, as well as with other psychological (i. e., functional) and nonpsychological internal states of the system [43.34]. Putnam's [43.35, papers 16–21] formalization of functionalism via the theory of effective computability gave rise to an early version of computational functionalism, known as *machine functionalism*.

Machine functionalism was challenged on several fronts. For one thing, "it still conceived psychological explanation in the logical positivists' terms of subsuming observed data under wider and wider universal laws" [43.36, pp. 53–54]. However, Fodor [43.37], Dennett [43.38], and Cummins [43.39] noticed that psychology, like biology, does not traffic with nomological explanations, which are predictive tools based on laws; rather, psychological explanations are functional analyses in Cummins' sense [43.40, 41]. This explanatory practice has also been defined by decomposition, by identifying *homunculi*, by reverse engineering, by taking the design stance, by describing the articulation of parts, and by discovering mechanisms. As Craver [43.42, p. 107] suggests, these definitions can all be seen as contributions to a theory of *mechanistic explanation* (see Sect. 43.1.4 *Mechanisms and Mechanistic Explanation*).

In the *homuncular* version of functional analysis [43.38, 43, 44], behavioral data are manifestations of the complex cognitive capacities of an agent (vision, memory, face recognition, etc.); and those capacities are to be explained by assuming that an agent is a system of interconnected, hierarchically organized components. Every component (or module) is a *homunculus*, that is, an intelligent mechanism that is individuated by means of the function it performs; and swarms of homunculi cooperate with each other in such a way as to produce the overall activity of the system. The homunculi are in turn seen as teams of simpler homunculi, whose functions and interactions are similarly used to explain the capacities of the subsystems that they compose; and so on, *recursively*, until the sub-sub-... components turn out to be simple

neuroanatomical structures [43.45, p. 320]. Marr's computational theory of vision is an exemplar of functional analysis in cognitive science.

Another flaw in machine functionalism lies in its conception of the *functional realization* (the relation between an organism and the abstract program that it is supposed to instantiate) as a simple one-to-one correlation between the organism's repertoire of physical stimuli, inner states and behavioral responses, on the one hand, and the inputs, states and outputs specified by a machine table on the other. This criterion of realization, though, is too *cheap*: since virtually anything bears a one-to-one correlation of some sort to virtually anything else, *realization* in the sense of one-to-one correspondence is far too easily come by [43.46, Chap. 4]. To put this defect right, Lycan [43.43, 44] and Millikan [43.31] have imposed a *teleological* constraint on realization: a physical state of an organism will count as realizing a certain functional description only if the organism has genuine organic integrity and the state properly plays its functional role *for* the organism. The state must do what it does because this is its biological purpose. In this perspective, psychological functions are biological functions in the etiological sense, that is, effects that are promoted by natural selection.

The etiological notion of function implies environmental constraints. Here it can be useful to draw a distinction between *narrow* and *wide* function [43.47]. *Narrow* functional analysis only looks at the system as such. The function boils down to the causal role that a property plays in the *internal* economy of some system, which consequently is insulated from the *extra-cranial* environment where it is situated [43.48]. Famously, Fodor [43.49] has argued that narrow functional analysis is the only admissible form of functional analysis. According to the Fodorean *methodological solipsism*, psychological explanations should be restricted to quantifying over the formal, *intrinsic* properties of mental states, unlike the *naturalistic* psychology (exemplified by James' naturalism, learning theory, and Gibson's ecological optics), in which generalizations are defined by the relations between mental representations and their environmental, *extrinsic* causes – a research strategy, Fodor warns, that will hardly turn out to be fruitful.

In contrast, *wide* functional analysis refers to the extra-cranial environments in which an organism is situated, and it involves teleological functional considerations about the relationship between the organism and its environment. Pace Fodor, Harman thinks that “only a wide psychological functionalism can motivate appropriate distinctions between aspects of a system, irrelevant side effects and misfunctions” [43.50, p. 20]. This teleological notion of functional analysis leads us

to look beyond the mental apparatus as such to the way it deals with its environment.

Despite the fact that the first generation of research in cognitive science has largely adopted (at least implicitly) narrow functional analysis during last three decades, cognitive scientists have shown growing impatience against the idea that the scientific study of an individual's psychological states is required to abstract from all the environmental, extrinsic variables, to take into consideration only the *intra-cranial*, intrinsic properties of those states. Various factors contributed to this change of climate, including the rising success of Putnam's and Burge's semantic externalism, Searle's and Dreyfus' classical criticism of symbolic artificial intelligence, the consolidation of Gibson's influence in cognitive psychology. In this framework, research programs very different from each other – for example, the sensorimotor (or enactivist) approach to perceptual experience, situated robotics, and dynamic approach to cognition – have endorsed an *externalist* conception of explanation according to which biological cognition cannot be understood without taking into consideration its embodied and situated nature (see Sect. 43.2.2 *Active Externalism*).

#### 43.1.4 Computational Models and Mechanistic Explanations

##### Mechanisms and Mechanistic Explanation

Another serious shortcoming of machine functionalism is the extreme biological implausibility of its *two-level* view of human psychobiology. As Lycan [43.44, Chap. 4]; [43.45, p. 320] has incisively argued, neither living things nor even computers can be split into a purely structural level of neurobiological/physiochemical (in the case of computers: electronic) description and an abstract level of cognitive-algorithmic description. Rather, organisms and computers are hierarchically organized at many levels, where each level is *abstract* with respect to those beneath it but *structural* or concrete as it realizes those levels above it. The functional/structural or software/hardware distinction is entirely relative to one's chosen level of organization. In brief, Lycan concludes, “nature is functions all the way down” [43.44, p. 48].

This emphasis on the hierarchical organization of functional levels has been worked out by Craver [43.42], which represents a synthesis between Cummins' theory of causal-role functions and the thriving literature on mechanisms and mechanistic explanation.

The mechanistic explanation of a phenomenon consists in the specification of the mechanisms that produced it. But what is a mechanism, exactly? Accord-

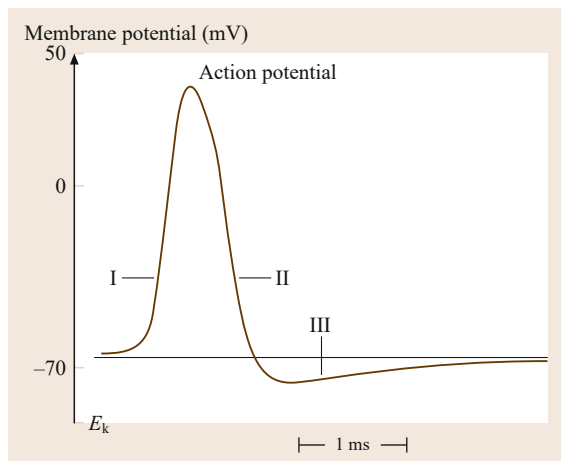
ing to Machamer et al. mechanisms are “entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions” [43.51, p. 3]. These systems have a spatiotemporal organization that explains the way they realize their activities. One example, taken from Craver [43.52, 53], is the mechanism of the action potential. Here the entities are the axonal membrane, the  $\text{Na}^+$  and  $\text{K}^+$  ions, and two types of voltage-dependent ionic channels that allow  $\text{Na}^+$  or  $\text{K}^+$  ions, respectively, to spread through the membrane. Membrane, ions, and ionic channels regularly act and interact to produce the action potential. These activities essentially depend on the *spatial* organization of the mechanism’s components (e.g., ionic channels span the membrane, thus producing ion fluxes across it); but still more crucial is *temporal* organization: the shape of the action potential is explained by the relative orders and durations of the activation and inactivation of ionic channels (Fig. 43.2).

Moreover, mechanisms are intrinsically organized in multilayers, that is, they are involved in a hierarchical organization in which lower level entities and activities are the components of higher level entities and activities. Therefore, a mechanism is a hierarchically organized (to produce a certain goal or behavior) system, and a mechanistic explanation is usually, if not always, multilayered. Following Craver [43.42, 54], we can say that an *ideally complete* mechanistic explanation describes a mechanism by integrating three perspectives. The *isolated* perspective (level 0) describes the mechanism at its proper, characteristic

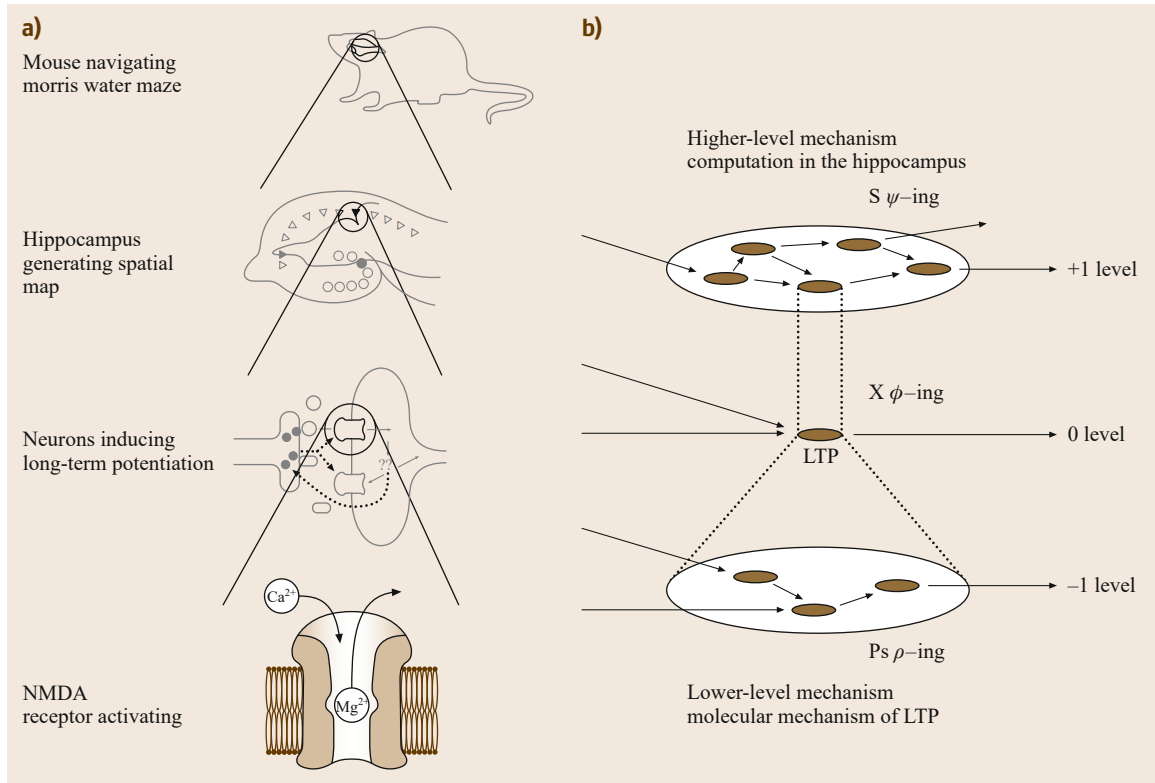
level. It is an ordinary causal explanation describing the input–output relations of the mechanism. The *contextual* perspective (level +1) locates the mechanism in the context of another mechanism, being the former a part of the latter; this means that the activities of the former contribute to the working of the latter. Finally, the *constitutive* perspective (level –1) breaks down the mechanism in its constitutive parts in such a way that we can understand how these parts enable the input–output relations of the mechanism defined at the level 0. Mechanistic explanation is thus aimed at integrating levels of mechanisms. However, ideally complete mechanistic integration is hard to attain, and ideally complete mechanistic explanations are correspondingly rare [43.52, p. 360]:

“Models that describe mechanisms can lie anywhere on a continuum between a mechanism sketch and an ideally complete description of the mechanism. [...] A mechanism sketch [...] characterizes some parts, activities, and features of the mechanism’s organization, but it has gaps. [...] At the other end of the continuum are ideally complete descriptions of a mechanism. Such models include all of the entities, properties, activities, and organizational features that are relevant to every aspect of the phenomenon to be explained. Few if any mechanistic models provide ideally complete description of a mechanism.”

To provide an example of this quest for the integration of levels of mechanisms, Craver [43.42] examines the development of the explanations of long-term potentiation (LTP) and spatial memory. He distinguishes at least four levels (Fig. 43.3a). At the top of the hierarchy (the behavioral-organismic level) are memory and learning, which are investigated by behavioral tests. Below that level is the hippocampus and the computational processes it is supposed to perform to generate spatial maps. At a still lower level are the hippocampal synapses inducing LTP. Also finally, at the lowest level, are the activities of the molecules of the hippocampal synapses underlying LTP (e.g., the *N*-methyl D-aspartate receptor activating and inactivating). These are mechanistic levels: the *N*-methyl D-aspartate receptor is a component of the LTP mechanism; LTP is a component of the mechanism generating spatial maps, and the formation of spatial maps is a part of the spatial navigation mechanism. Integrating these four mechanistic levels requires both a *looking up* integration, which will show that an item (LTP) is a part of an upper level mechanism (a computational-hippocampal mechanism); and a *looking down* integration, which will describe the lower level mechanisms underlying the



**Fig. 43.2** The action potential consisting of (I) a rapid rise in  $V_m$  to a maximum value of roughly +35 mV, followed by (II) a rapid decline in  $V_m$  to values below  $V_{rest}$ , and then (III) an extended after-potential during which the neuron is less excitable (known as the refractory period) (after [43.42])



**Fig. 43.3a,b** Levels of spatial memory (a). Integrating levels of mechanisms (b) (after [43.42])

higher level phenomenon (the molecular mechanisms of LTP) [43.42, 55, 56] (Fig. 43.3b).

Note that providing a mechanistic explanation of a certain phenomenon (capacity, system) is giving a certain type of *causal* explanation, since searching for mechanisms is searching the effective cause of a certain behavior. This is a much important point, since, on this perspective, mechanistic explanation confronts nomological explanation. Although it could be argued that there is some relationship between laws and causes (consider that it is not rare to hear that causal relations are instantiations of laws), from the point of view of the explanatory style mechanistic explanations do not involve laws. (For a recent discussion of the issue see [43.57].) We will come back later on the significance of this opposition.

### The Relation Between Mechanisms and Computational Models

Recently the notion of computational model has been embedded in the larger explanatory framework of mechanistic explanation. But what is, exactly, the relation between mechanisms and computational models? The relation is twofold. On one hand, mechanisms are physical systems that implement, or realize, computa-

tions. On the other hand, a computational model can be one among the multiple levels of organization that constitute a mechanistic explanation [43.14, 17]. For example, the above-mentioned computational process that hippocampus is supposed to perform to generate spatial maps. With regard to the first aspect, to the extent that mechanisms provide the nexus between abstract computations and concrete physical systems, they ground the notion of computation. The second aspect focuses instead on the issue of explanatory styles, subsuming computational explanation under the larger class of mechanistic, and thus causal, explanations.

To sum up, computations are realized by mechanisms, that is, if we give a computational explanation of a certain phenomenon or ability, there must be a mechanism – a system that admits a mechanistic description – which realizes or performs the computation. Also since computational systems may be articulated in several levels, as we have seen in the case of Marr’s model, it seems appropriate to say that a computational model is a complex part of a mechanistic explanation (and, correspondingly, a computational system is a certain kind of mechanism, usually a part of a complex mechanism).

Even if there are computational models that are defined with no reference at all to mechanisms (as it was

customary in classical cognitive science, since a full-blooded account of mechanistic explanation has been provided only about 15 years ago), we hold that computational models are always mechanistic too, since computational explanation is essentially mechanistic. It is mechanistic because the fundamental features of mechanistic explanations – entities realizing functions, components etc. – are saved in computational models (for a slightly different point of view, see [43.17]). This is one of the reasons to believe that computational models, *qua* mechanistic, can hardly be integrated with dynamical models (see Sect. 43.1.5).

### 43.1.5 Dynamical Systems

Neural networks and classical models are all computational models. Arguably, the same cannot be said for dynamical systems, at least if we consider them from the perspective of the explanatory style.

Dynamicism consists essentially in applying to behavior and cognition methods from the theory of nonlinear dynamical systems theory. A dynamical system is a physical system whose behavior is described by a set of differential equations. *Nonlinear* means that, given the initial conditions and the equations, the target behavior of the model cannot be analytically determined (except for some particular cases) and will be predicted only with a certain approximation. For instance, even once fixed the initial conditions, the system could evolve in two or more different states.

The dynamical approach focuses on the time evolution of a system, and seems particularly well equipped to deal with cases in which a system or a component of a system *A* constantly influences and is constantly influenced by another system or component *B* (which could in turn be sensible to another component *C*, and so on). Take, for instance, a tennis player who is going to reply the service: here the locations of the ball and of the other player change continuously and, at the same time, the player moves and acts, influencing in turn the other player, and so on so forth [43.58, p. 348]. In sum, “[e]verything is simultaneously affecting everything else” [43.59, p. 23]. These events, which are strictly *coupled*, appear to be difficult to model with the tools of classical computational models.

Let us examine an often-cited example of dynamical analysis [43.60, Chap. 2]. In one experiment, the subjects were asked to oscillate their index fingers back and forth with the same frequency in each finger. The oscillation could be in-phase (homologous muscle groups contracting simultaneously) and antiphase (homologous muscle groups contracting in an alternating fashion). At high frequencies of oscillation, however, the antiphase movement is unstable and at a critical frequency subjects spontaneously switch from the antiphase motion of the fingers to an in-phase symmetrical pattern.

A dynamical analysis of this pattern of results begins with plotting the phase relationship between the two fingers. This variable (= the relative phase) is constant for a wide range of oscillation frequencies but is subject to a strong shift at a critical value, namely, the moment of the antiphase/phase shift. Plotting the unfolding of the relative phase variable is plotting the values of a *collective variable*, that is, one whose value is set by a relation between the values of other variables (the ones describing individual finger motions). The values of these collective variables are fixed by the frequency of motion, which thus acts as a so-called *control parameter*. The dynamical analysis consists therefore in a set of equations displaying the space of possible temporal evolutions of relative phase as governed by the control parameter. This description fixes, in detail, the so-called *state space* of the system. In this framework, the collapse of the antiphase pattern of coordination into the phased one can be construed as the transition from a *landscape* where there are stable attractors for both patterns of coordination to one in which there is a stable attractor only for the in-phase motion. The dynamical analysis turned out to be very useful also in understanding the activity of simple robots [43.61] and the development of infant locomotion [43.62].

Although there are scholars who regard dynamical systems as a class of computational systems, essentially because differential equations are computable (and are a way to compute as well), we take dynamical models as an *alternative* to computational models, since the former are usually brought to bear in nonmechanistic explanations. We will discuss more extensively this issue in the next section.



## 43.2 Open Problems in Computational Modeling

In this section, we will take into consideration two different kinds of problem faced by modeling in cognitive science. The first problem concerns the scope of computational explanation and has specifically to do with its capacity of accounting of the so-called *central processes*. The second problem concerns the feasibility of integrating different styles of explanation.

### 43.2.1 Computationalism and Central Cognition

#### The Problem of the Central Cognition

Jerry Fodor's Computational and Representational Theory of Mind (CRTM) is one of the most important and, at the same time, controversial systematizations of computational functionalism. Among its virtues is a powerful response to Skinner's objection against the mentalistic explanation in psychology, the *homunculus fallacy*. This is a vital constraint on *any* serious mentalistic psychology: a plausible theory of cognition must avoid the infinite regress triggered by the attempt to explain a cognitive capacity by tacitly positing an internal agent with that very capacity. To discharge all the homunculi CRTM combines two ideas:

- *Formality condition*: The rules that govern the state transitions in a computational system are sensitive only to the *form* of representations, namely to their syntactic properties, whereas they are insensitive to the semantic ones.
- *Recursive decomposition*: Complex capacities are structured ensembles of much simpler capacities. Thus CRTM endorses the already-mentioned recursive decomposition of complex cognitive capacities into co-operating ensembles of simpler capacities.

Jointly, these two ideas ensure that a computational theory of cognition begs no questions. The theory *explains* intelligent agency rather than *presupposing* intelligent agency. For the formality condition guarantees that the elementary operations of a computational theory do not presuppose intelligence (do not involve *interpretations*); and the recursive decomposition guarantees that all cognitive functions are ultimately explicable by these elementary operations [43.63].

Is it really feasible to combine the formality condition with the recursive decomposition? In other terms, can the parts or modules into which the functionalists decompose the mind work as engines that satisfy formality condition? To answer this question, we need to distill from the heterogeneous collection of definitions of *module* and *modularity* those relevant to

our issue. For example, when many psychologists and philosophers speak of modules, they intend to refer to functionally individuated psychological systems. But then, *Fodor* justly notes, “everybody who thinks that mental states have any sort of structure that’s specifiable in functional terms qualifies as a modularity theorist in this diluted sense” [43.64, pp. 56–57]. It is therefore advisable to ascribe to the term a less diluted sense on pain of vacuity.

The first stronger sense of *module* is an epistemic one. A module is a *body of cognized information* – in Chomsky’s linguistics, the tacit beliefs of a speaker/hearer who masters the grammar of her own language, namely the principles of universal grammar, plus a choice of parameter values, plus a lexicon. Moreover, this psychological structure is *domain specific*, that is, it is dedicated to solving a restricted class of problems in a restricted content domain. By contrast, a *domain-general* or *general-purpose* psychological structure is one that can be used to do problem solving across many different content domains.

A knowledge base, however, cannot give rise to behavior through its propositional content alone. Mechanisms are necessary “to bring the organization of behavior into conformity with the propositional structures that are cognized” [43.65, p. 9]. If this mechanism is designed to compute only a restricted class of inputs (i. e., representations of the properties and objects found in a particular domain), we get a “vertical faculty,” namely a domain-specific computational mechanism. For example, the systems that compute the phonological analysis of speech are domain specific in that they operate only upon acoustic signals that are taken to be utterances: “[T]he very same signal that is heard as the onset of a consonant when the context specifies that the stimulus is speech is heard as a *whistle* or *glide* when it is isolated from the speech stream” [43.65, p. 49].

In some domains, a domain-specific computational mechanism and an *epistemic* module can form a single mechanism. For example, it may be supposed that a syntactic parser is a domain-specific computational mechanism that takes as input sensory (e.g., acoustic) representations of utterances and, in virtue of a database dedicated to linguistic information, delivers syntactic and semantic representations of physical sentence forms. It is important to note, however, that in other domains *general-* rather than specific-domain algorithms could employ domain-specific bodies of information. (See Sect. 43.2.1 *The Massive Modularity Hypothesis*.) A *Fodorean* module is precisely such a mechanism – a domain-specific innately specified processing system with an innately specified epistemic module as its pro-

proprietary database. Moreover, it has its own proprietary transducers; delivers *shallow* (nonconceptual) outputs; it is mandatory in its operation, swift in its processing, associated with particular neural structures, liable to specific and characteristic patterns of breakdown, and developed according to a paced and distinctively arranged sequence of growth.

The most important aspect of such a system, though, is that its information exchanges with other systems are architecturally constrained. First, the *central* cognitive processes (such as belief fixation and practical deliberation) can access only the output representations of the *macromodules* (= modules composed by *micromodules*). Consequently, the intermediate representations generated by the micromodules are *inaccessible* to central cognition. Second, Fodorean modules are isolated from the rest of cognitive system in that they are *informationally encapsulated*. This means that a module works by employing only its proprietary database, without any appeal to the more general knowledge available to the rest of cognition. Marr (and more recently Pylyshyn [43.66, 67]) thought that early vision mechanisms were informationally encapsulated; Fodor extends this idea to any peripheral input and output system – including audition, face-recognition, language-processing, and various motor-control systems.

Now, Fodor [43.64, 65, 68] rejects the idea that the entire human cognition is modular. For what he intends as *central cognition* are processes such as belief fixation and practical deliberation, as they are described in personal-level propositional attitude psychology. Also intentional psychology is distinctively flexible and nonmodular, as is shown by an analogy between the processes that lead to the formation of beliefs and the type of nondemonstrative inference that is characteristic of the confirmation of hypotheses in science. (Fodor's analogy is rejected by Carruthers [43.69] and Pinker [43.70]). Like scientific confirmation, central cognition is characterized by two properties: it is *isotropic*, in the sense that “in principle, any of one's cognitive commitments (including, of course, the available experiential data) is relevant to the (dis)confirmation of any new belief” [43.68, p. 115] and it is *Quinean*, in the sense that central systems compute over properties like simplicity, centrality, conservatism, which are fixed by the global structure of belief system. But if the central systems are Quinean/isotropic, then they are not encapsulated – processes of belief formation need to have free access to the whole belief system.

Fodor's claim that the architecture of the mind is mostly *non-modular* has very bleak implications for CRTM and computational psychology. The formality condition says that the rules that govern the state transi-

tions in a cognitive system are sensitive only to the *form* of representations, namely to their syntactic properties, whereas they are insensitive to the semantic ones. Here an analogy with a logical calculus holds: the form of the proposition  $P$  and  $Q$  is entirely a matter of the identity and arrangement of its parts; and this is everything one must know to infer that the proposition is true if and only if  $P$  and  $Q$  are both true – namely it is not necessary to know anything about either the meaning of  $P$  or  $Q$ , or the extra-linguistic world. Therefore, syntactic properties are *local* properties of representations, that is, “they are constituted entirely by what parts a representations has and how these parts are arranged. You don't, as it were, have to look ‘outside’ a sentence to see what its syntactic structure is, any more than you have to look outside a word to see how it is spelled” [43.64, p. 20]. In other terms, from the formality condition, which requires computations to be sensitive only to the form of representations, follows the *locality principle*, according to which computations work only upon the local properties of representations.

In the modular processes, it is encapsulation that assures the respect of the locality principle: for the computations of an encapsulated module are supposed to access only the information in the proprietary database, ignoring all other information held in the mind. However, central processes do not satisfy the locality condition, being isotropic and Quinean; consequently, they are not computationally tractable processes. In other terms, they face what is sometimes known as *the frame problem*.

### The Massive Modularity Hypothesis

In order to deal with Fodor's problem of the computational intractability of central cognition some cognitive scientists have pursued the *massive modularity hypothesis* (MMH), which is the core of evolutionary psychology, a research program in cognitive science that aims to combine computationalism, nativism, and adaptivism. Instead of a picture of domain-specific encapsulated modules feeding into a domain-general propositional attitude system, MMH portrays the mind as a swarm of overlapping *Darwinian* modules of varying degrees of specialization and domain specificity. Such systems are innate computational modules that take epistemic modules as their databases, and are the product of natural selection. The latter is their characteristic mark: they are adaptations, that is, cognitive mechanisms shaped by the forces of evolution to solve major problems posed in ancestral environments (foraging, avoiding predation, finding shelter, co-ordinating with others, choosing mates, etc.).

According to MMH, then, Fodor's problem of the huge informational load on the central cognition is the

result of misdescribing human cognitive capacities. The problem does indeed show that an intelligent general-purpose agent could not be thinking by computing. However, we are not such agents; our mind is a complex structure of Darwinian modules. This view avoids Fodor's problem of central cognition by rejecting the picture of a central functional arena where the contents of an agent's propositional attitudes are stored, processed, and poised for the control of behavior.

However, legitimate doubts have been raised about the empirical evidence offered in favor of MMH. *Samuels* [43.71], for example, pointed out three kinds of difficulties of the empirical arguments for MMH:

1. The supposed evidence for the existence of Darwinian modules can often be explained in terms of other less specific mechanisms. For example, it has been objected that the content effects in Wason's selection task are not cogent evidence for the existence of a module – the “social-cheater” mechanism – that would be specific to the domain of social reasoning, since they can be more plausibly explained in terms of general mechanisms of verbal understanding [43.72].
2. Even when there are data that suggest the existence of a specialized cognitive structure in a given domain, it is arduous to establish whether it is a computational module or an epistemic one. For epistemic modules can be easily integrated in a theory of central processing in which different corpora of knowledge are computed by a few nonmodular mechanisms. This point is very well illustrated by the debate on folk biology. As *Samuels* [43.71, p. 46] has rightly noticed, the problem here is that the main evidence for folk biology does not allow us to adjudicate between the domain-specific hypothesis [43.73] and the general-domain one [43.74]. All that it allows us to claim is that folk biology needs a dedicated and (perhaps) innate cognitive structure; but this, of course, is not enough to demonstrate the existence of a Darwinian module for folk biology.
3. Even when some robust evidence for computational modules is available, it is very hard to determine whether their domain of application is really central cognition. For example, the evidence for a computational module specialized in processing geometrical information [43.75] does not support MMH since it can be plausibly viewed as part of vision or visuomotor control [43.66]. The same holds for what is considered a strong candidate for central modularity, viz. Leslie's Theory of Mind Mechanism (ToMM). That ToMM is evidence for central modularity is highly controversial, mainly in the light of *Leslie's* most recent work, where ToMM

is viewed as a relatively low-level mechanism of selective attention [43.76], whose functioning rests on nonmodular executive systems like the selection processor.

### Narrowing the Scope of Central Processing

The difficulties afflicting the empirical arguments for MMH invite us to be very cautious about taking it as a *complete* account of human cognitive architecture. The most we can argue is that there are a number of domain-specific and/or encapsulated central systems, but there are also nonmodular (domain-general and unencapsulated) central systems as well.

In light of this, *Bermúdez* [43.77] argued that the interesting point about (this weaker form of) MMH is the suggestion that the role of propositional attitudes in human cognition has been overstated. MMH leads us to rethink the traditional nexus between intelligent behavior and propositional attitudes, realizing that much social understanding and social coordination are subserved by mechanisms that do not capitalize on the machinery of intentional psychology. The latter does not rule all social interactions; it is used far less frequently than is commonly assumed in philosophy of mind. In this perspective, MMH goes in the right direction since it narrows the scope of central processing, thus outlining [43.77, p. 242]:

“the possibility of more or less direct links between perception and action that are sophisticated enough to be characterized as forms of intentional behavior, and yet that do not engage the propositional attitude system.”

Emphasis on belief/desire as the source of action is strictly related to the endorsement of what can be called *the sandwich model of the mind* [43.78], whereby perception and action are systematically mediated by the central cognition layer. Central cognition is the sandwich filling: according to the sandwich model, perception yields beliefs, and these, in turn, trigger actions. The separation between the three layers, in particular between perception and cognition, is taken to be very neat, especially if the model is coupled (as often is) with the Fodorean view of modularity, according to which only perceptual systems are modular.

Therefore, if one rejects the sandwich model, beliefs and desires may no longer be considered as the unique causes of action. There are other mechanisms, not located at the level of folk psychology, which can explain action. These mechanisms, however, are not only the Darwinian modules. For instance, a mechanism of emotional sensitivity such as *social referencing* is a form of *low-level* mindreading that subserves social understanding and social coordination without in-

volving the attribution of propositional attitudes. Here Bermúdez is on the same wavelength as mental simulation theorists and social neuroscientists in drawing our attention to forms of *low-level* mindreading that have been largely neglected by philosophers. However, he goes a step beyond them and explores cases of social interactions that point in a different direction, that is, situations that involve mechanisms that can no longer be described as mindreading mechanisms. He offers two examples.

In game theory, there are social interactions that are modeled without assuming that the agents involved are engaged in explaining or predicting each other's behavior. In social situations that have the structure of the iterated prisoner's dilemma, the so-called *tit-for-tat* heuristic simply says: "start out cooperating and then mirror your partner's move for each successive move" [43.79]. Applying this heuristic simply requires understanding the moves available to each player (cooperation or defection), and remembering what happened in the last round. So here we have a case of social interaction that is conducted on the basis of a heuristic strategy that looks backward to the results of previous interactions rather than to their psychological etiology. We do not need to infer other players' reasons; we only have to coordinate our behavior with theirs.

There is another important class of social interactions that involve our predicting and/or explaining the actions of other participants, but in which the relevant predictions and explanations seem to proceed without us having to attribute propositional attitudes. These social interactions rest on what social psychologists call *scripts* (*frames* in artificial intelligence), that is, complex information structures that allow predictions to be made on the basis of the specification of the purpose of some social practice (e.g., eating a meal at a restaurant), the various individual roles, and the appropriate sequence of moves

According to Bermúdez, then, much social interaction is enabled by a suite of relatively simple mechanisms that exploit purely behavioral regularities. It is important to notice that these mechanisms subserve *central* social cognition (in Fodor's sense). Nevertheless, they implement relatively simple processes of template matching and pattern recognition, that is, processes that are paradigmatic cases of perceptual processing. For example, when a player *A* applies the tit-for-tat rule, *A* must determine what the other player *B* did in the preceding round. This can be implemented in virtue of a template matching in which *A* verifies that *B*'s behavioral pattern matches *A*'s prototype of cooperation and defection. Moreover detecting the social roles implicated in a script-based interaction is a case of template matching: one verifies whether the perceived

behavior matches one of the templates associated with the script (or the prototype represented in the *frame*).

Bermúdez notes that the idea that much of what we intuitively identify as central processing is actually implemented by mechanisms of template matching and pattern recognition has been repeatedly put forward by the advocates of the connectionist computationalism, especially by Paul M. Churchland. But unlike the latter, Bermúdez does not carry the reappraisal of the role of propositional attitudes in social cognition to the point of their elimination. For he argues that social cognition does not involve high-level mindreading when the social world is *transparent* or *ready-to-hand*. However, when we find ourselves in social situations that are *opaque*, that is, situations in which all the standard mechanisms of social understanding and interpersonal negotiation break down, it seems that we cannot help but appeal to the type of metarepresentational thinking characteristic of propositional attitude psychology.

### The Global Workspace Approach to Central Processing

To the extent that Bermúdez leaves room for forms of social interactions involving that type of metarepresentational thinking characteristic of propositional attitude psychology, he only made Fodor's problem of central cognition less pressing. The architecture for what is left of our central processing still consists in a freely accessible cognitive realm, or central arena, in which attitudes of all types can become active and enter in processes of reasoning and thinking. According to Carruthers [43.80], however, the evidence from cognitive science allows us to hold that there is not such an arena.

Carruthers's claim rests on the validity of a global workspace account of the conscious accessibility of our perceptual experiences. There is now extensive evidence supporting such models [43.81–87]. Moreover, subsequent analyses of functional connectivity patterns in the human brain have demonstrated just the sort of neural architecture necessary to realize the main elements of a global broadcasting account [43.88, 89]. Specifically, these studies show the existence of two main neurocomputational spaces within the brain, each characterized by a distinct pattern of connectivity.

The first space is a processing network, composed of a set of parallel, distributed, and functionally specialized processors or modular subsystems subsumed by topologically distinct cortical domains with highly specific local or medium-range connections that encapsulate information relevant to its function.

The subsystems compete each other to access the global neuronal workspace, which is implemented by long-range cortico-cortical connections, mostly originating from the pyramidal cells of layers 2 and 3 that are

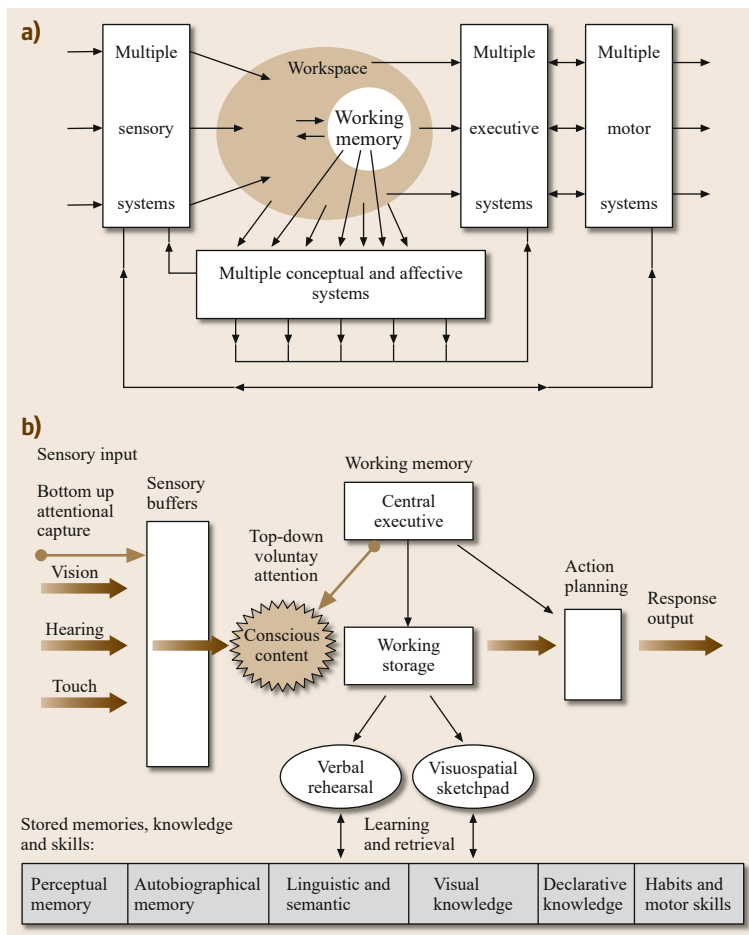
particularly dense in prefrontal, parieto-temporal and cingulate associative cortices, together with their thalamo-cortical loops. When one of these subsystems accesses the global neuronal workspace, its outputs (i. e., sensory information including perceptions of the world, the deliverances of somatosensory systems, imagery and inner speech) are broadcast to an array of specialized executive, conceptual, and affective consumer systems – for example, systems that consume the perceptual input to form judgments or make decisions (Fig. 43.4a).

Global broadcasting makes possible the development and subsequent benefits of a working memory system which exploits the mechanisms of global broadcast to subserve a wide variety of central cognitive purposes [43.90–92] (Fig. 43.4b). So there is, indeed, a central arena in the mind; but this is working memory. And what can be found within working memory are not propositional attitudes, but rather imagery, inner speech, and so forth; working memory’s operations are always sensory based.

Thus, a global broadcast architecture arranges in parallel specialized conceptual systems around the global broadcast of attended perceptual information, and makes entry into a general-purpose working memory system competitive. Such design features seem to enable us to circumvent many aspects of Fodor’s scepticism about being central cognition amenable to computational modeling [43.89, 93, 95, 96].

### 43.2.2 The Dynamicist Challenge: Is Integration Possible?

As mentioned in Sects. 43.1.4 and 43.1.5, dynamicism can better be regarded as a research paradigm alternative to mechanicism, therefore to computationalism too. In this connection, a standard reference is *van Gelder and Port* [43.59], which was the first major presentation of the dynamical approach to cognition. According to the authors, “to see that there is a dynamical approach is to see a new way of conceptually reorganizing cognitive



**Fig. 43.4** (a) Global broadcast and working memory (after [43.93]). (b) A functional framework for attention and conscious events (after [43.94])

science as it is currently practiced” [43.59, p. 4]. Such a reorganization takes a stand against not only classical computationalism but also the connectionist one – and this despite the fact that connectionists were the first to apply the dynamical systems theory to the study of cognition. However, van Gelder and Port argue, the limit of connectionism lies in the use of the dynamical systems tools within a paradigm that is still the computationalist and representationalist one, even though in a brain-like variant. The dynamicist wants to go beyond.

First, the dynamicist dissolves the boundary between the cognitive system and the system’s environment. Coupling between the equations describing a cognizing system and those describing the environment gives rise to complex *total system* behaviors. In this perspective [43.97, p. 373],

“the cognitive system is not just the encapsulated brain; rather, since the nervous system, body, and environment are all constantly changing and simultaneously influencing each other, the true cognitive system is a single unified system embracing all three.”

Second, the dynamicist expansion into the environment implies an explanatory model very different from the mechanistic one underlying the vertical expansion. In the 1950s, the appeal to the mechanistic explanatory strategy by early cognitivists was the logical conclusion of the battle waged against behaviorism and mathematical psychology, which conceived psychological explanation as discovery of laws or mathematical regularities in behavior [43.1, p. 96]. The dynamical approach, however, relaunches the covering law conception of explanation. The dynamical analysis identifies the critical variables characterizing the state of a system and attempts to construct laws (a set of differential equations) to account for the system’s trajectory through state space. The system can no longer be decomposed into subsystems (modules) that involve computations on representations. Consequently, the dynamical explanation is seen as incompatible with the explanatory style of the computationalist mechanistic (Fig. 43.5).

Dynamicism, then, puts forward the *radical embodied cognition thesis*: to understand the complex interplay of brain, body, and environment we do not need either the concepts of internal representation and computation or the mechanistic decomposition of a cognitive system into a multiplicity of inner neuronal or functional subsystems; all we need are the analytic tools and methods of dynamical systems theory [43.98, p. 148]. In this form, however, the dynamicist project is not a *third contender* (the other two contenders are Fodor’s CRTM – or, more generally, symbolic models –

and neural networks) in the controversy on the foundations of cognitive science but, rather, the denial of the possibility of such a science – to the extent, of course, that we are right in claiming that (some form of) computational functionalism is at the core of the very idea of a cognitive science.

However, a reformist perspective challenged the dynamicist obituary for cognitive science. It uses the objections to the individualism of classical cognitive science as guidelines to reconstruct the conceptual bases of cognitive science.

### Active Externalism

Such a reformism has been pursued by Andy Clark in the externalist framework introduced in Sect. 43.1.3. Clark thinks that the computational and representational framework can be reconstructed making due allowances for the embodied and world-embedded character of natural cognition but without collapsing into the radical embodied cognition thesis. Accordingly, he pursues the transformation of that framework into just one component in a three-tiered explanatory strategy [43.98, p. 126]:

- (i) A dynamicist account of the gross behavior of the agent–environment system
- (ii) A mechanistic analysis, describing how the components of the agent–environment system interact to produce the collective properties described in (i)
- (iii) A representational and computational account of the components identified in (ii).

Clark calls this tripartite explanatory strategy *minimal representationalism*, and puts it into a wider theoretical framework: *active externalism* [43.100, 101]. Unlike *semantic externalism*, where the mental *contents* of an agent are showed to partly depend on aspects of the environment which are clearly external to the agent, Clark’s externalism sees the environment as playing an active role in constituting and driving the agent’s cognitive processes. In the wake of Gibson’s

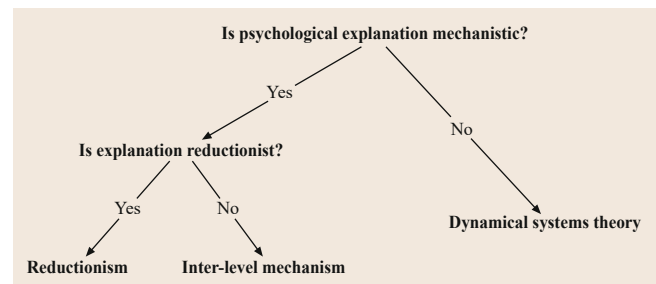


Fig. 43.5 The debate on explanatory style in cognitive science represented as a decision tree (after [43.99])

ecological optics, this environment is viewed as a collection of affordances that are the source of a particular variety of inner states, namely the *action-oriented* representations which, unlike the symbols in the language of thought, are *personal* (in that they are related to the agent's needs and the skills that she has), *local* (in that they relate to the circumstances currently surrounding the agent) and *computationally cheap* (compared with Marr's rich inner models of the visual scene).

It is important to notice, however, that action-oriented representations are only a representational *genus*. Clark rightly notices that the concept of inner representation was introduced in cognitive science to account for cases in which a cognitive system must coordinate its behaviors with environmental features that are not always reliably present to the system. In such cases, the cognitive system is able to decouple from the external environment and act in an offline fashion by creating some kind of inner item that stands in for the absent phenomena. These inner stand-ins are what cognitive scientists have termed *inner representations* [43.102]. Such cases of environmentally decoupled cognition are really a tough nut to crack for the antirepresentationalists, who are concerned exclusively with cases of *adaptive hookup*, that is, cases in which the inner states of a system (e.g., a sunflower, or a light-seeking robot) are supposed to coordinate its behaviors with specific environmental contingencies [43.98, p. 147]. Such cases of adaptive hookup, however, cannot ground a *general* antirepresentationalist argument: they are not sufficiently "representation hungry" [43.103].

In light of these considerations, Clark replaces the classical notion of mental representation with a *continuum of representational genera*. At one end of the spectrum there are the inner states that border the simple causal correlation and environmental control. At the other end of the spectrum, we find the type of inner stand-in that allows us to deal with the representation-hungry problems. Then between these two poles are the action-oriented representations.

According to Clark, therefore, depending on the coupling or decoupling between agent and environment, one must appeal to the dynamical nonrepresentational explanation or the representational one respectively. It can be objected, however, that this implies a division of labor between the two styles of explanation, and not their complementarity; as a result, they cannot be the tiers (i) and (iii) of a single explanatory strategy, as Clark would want.

Moreover, it is not clear how Clark's model of explanation can motivate the integration between the tiers (i) and (ii): if the interactions between the components of the global system can be described in mechanis-

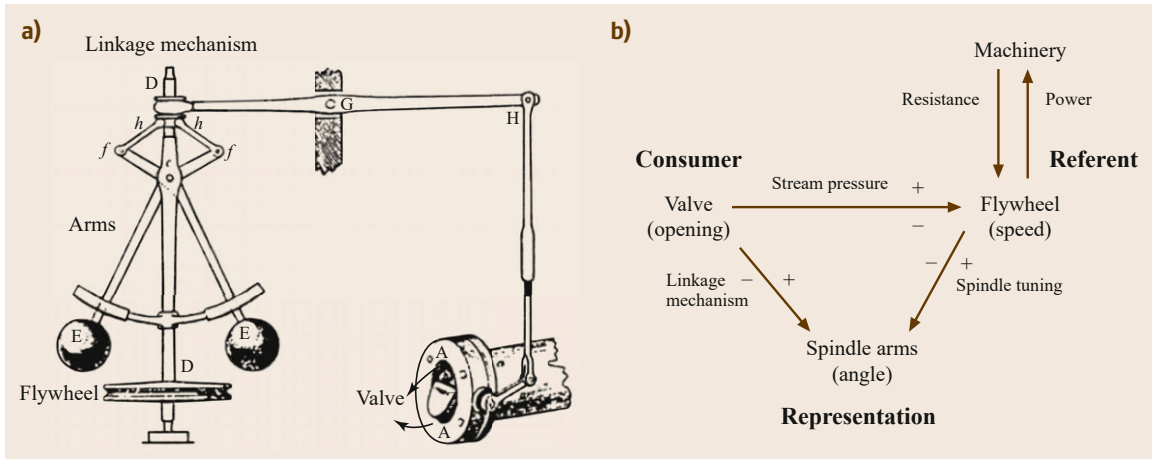
tic terms, is there still need to conceive the system in dynamicist terms? In the next section, we will see how *William Bechtel* and his collaborators [43.104–110] tried to deal with the issue.

### Integrating Dynamical Modelling and Mechanistic Analysis

*Bechtel* [43.105] has rejected antimechanistic construals of the dynamical approach to cognition in that they can be traced to a misunderstanding of the nature of mechanistic explanations. The source of this misunderstanding is *van Gelder's* [43.97] analysis of the Watt governor (Fig. 43.6). The latter is contrasted with a hypothetical computational governor that is characterized by the following classical features: it uses computations on representations, is sequential and cyclic, and has a *homuncular* (= mechanistic) character. According to van Gelder, the Watt Governor fails to exhibit these features because of the continuous and simultaneous relations of causal influence among the various factors involved. It is the continuous reciprocal causation that requires dynamical analysis.

However, mechanistic explanations are not required to have the sequential and cyclic character that van Gelder ascribes to the computational governor. *Bechtel* and *Richardson* [43.109, Chap. 7] note that in the early stage of the process of developing mechanistic models scientists often assume that the processes that they are considering are performed serially. However, when it is not possible for scientists to develop a linear model that is adequate to the phenomenon, they start to introduce feedback loops and other nonlinearities in attempting to develop adequate models. The outcome is what the authors define as *functionally integrated systems*, namely, systems that are not sequential and cyclic in van Gelder's sense.

Again, as in the case of representation, a continuum emerges. At one end of the spectrum, we have *fully decomposable* (or *highly modular*) systems, which are composed of subsystems that are completely independent except for the mutual exchange of outputs (this is the case of Fodor's encapsulated modules). If the interactions amongst the subsystems are weak but not negligible, the system is *nearly decomposable* [43.111]. As the complexities of interaction amongst parts increases, the explanatory burden shifts from the parts (or more precisely, the interactions *within* subsystems) to their organization (i.e., the interactions *between* subsystems). Thus, we reach the other end of the spectrum, where we find *holistic* systems whose components are functionally equivalent and hence interchangeable. In between the nearly decomposable systems and the holistic ones, there are the integrated systems. In these systems, unlike the holistic systems, it is possible to iso-



**Fig. 43.6** (a) Watt's centrifugal governor for a steam engine. (b) A schematic representation showing that the angle of the spindle arms carries information about the speed of the flywheel for the valve, which uses the angle to determine the opening, thereby regulating the speed of the flywheel (after [43.107])

late different parts that make distinctive contributions but also give rise to a complex set of interactions that are nonlinear, and hence much stronger than those of a nearly decomposable system.

Now, the Watt governor is an integrated system. Although it does not use the sequential and cyclic elements of the computational governor, it nonetheless can be explained in mechanistic terms. In explaining how it works, we identify separate modules, each of which contributes something different to its operation; the components are tightly coupled with each other, but no more so than in the case of fermentation. Analogously with psychobiological cognition – Bechtel [43.106] and Clark [43.98] suggest that much of it is likely to take up the intermediate space between nearly decomposability and holism.

Some proponents of a radically holistic view that rejects the very possibility to decompose the mind–brain are neuropsychologists who use dynamical systems tools to revive the Gestaltist principle that higher level activities depend upon the dynamical organization of the entire cortex [43.104, §3]. For example, Van Orden et al. [43.112] criticized the double-dissociation and neuroimaging studies that explain cognitive activities in terms of *single causes*, and promoted the approaches that rest on the notion of continuous reciprocal causation.

A piece of evidence that is supposed to confirm this holistic view is the presence in the brain of a vast number of feedforward, feedback, and collateral connections. However, as Bechtel [43.113, 114] has convincingly argued, the hypothesis of the neurobiological reality of holism is scarcely plausible. Important contrary evidence comes from the studies by David van Essen and

his collaborators, who have almost completely mapped the areas of the Macaque monkey's visual system over the last two decades [43.115–119]. The researchers have identified 32 different areas in the macaque visual cortex and more than 300 connections between these areas; and the tool-kit of dynamical analysis can be very useful to model this vast number of feedforward, feedback, and collateral connections. However, although these regions are highly interconnected, we can still determine what each area contributes to visual information processing; i. e., it is not a holistic system, but an integrated one. Indeed, Bechtel takes this work to be an exemplar of mechanistic analysis of how the brain performs a cognitive function. And in an integrated system mechanistic analysis “provides the foundation for dynamical analysis” [43.106, p. 483] since the latter has explanatory force only insofar as it describes “the operations of the underlying mechanism” [43.110, p. 443], only to the extent that it reveals “aspects of the causal structure of a mechanism” [43.120, p. 602].

Bechtel and Abrahamsen [43.108] refer to accounts integrating mechanistic decomposition of systems into parts and operations with the quantitative tools provided by dynamical systems theory as *dynamic mechanistic explanations*.

And yet Bechtel's dynamic mechanistic explanations do not appear to be really successful in harmonizing mechanistic–computational explanations with the dynamical ones. To see why, let us go back to the notion of integrated system.

Integrated systems have parts (subsystems) that are individuated according to a mechanistic principle; at the same time, however, since the inter-relations among parts are nonlinear (e.g., they cannot be reduced to sim-



ple input/output connections), their global organization requires a dynamical description, i. e., the whole system turns out to be a dynamical system. Note that in this picture the burden of explanation is carried by the mechanistic component, since the mechanistic decomposition of the system in parts is a non-negotiable condition. In other words, dynamical explanations make sense only against a mechanistic background – their role consists only in, so to speak, *filling the (explanatory) gaps*.

This idea is highly plausible since, in the case of cognition, computational models seem to possess a higher explanatory force [43.120], and computational models are mechanistic. But here again, as in Clark's case, how the integration actually works remains to a large extent obscure. First, computational explanations require (at least preferably) *modular* subsystems, whereas, according to Bechtel's model, the richness of interactions makes it difficult to regard subsystems as modules (and, if all parts were modules, of course we could dispense with dynamical explanations completely). Second, it is by no means obvious how to link the output of modules to the relevant dynamical variables of the whole system. The notion of integration is required to put together, in some way, computational descriptions and dynamical descriptions; by contrast, in the current view, the two kinds of explanations are merely alternative.

### 43.3 Conclusions

In this chapter, we gave a comprehensive account of *computational models* and showed their relevance in cognitive science. We discussed computational models from a variety of perspectives. First of all, from an explanatory perspective, making explicit the relation among computational models, computational explanation, and mechanistic explanation; second, from a metaphysical perspective, illustrating the relation between computational explanation and functionalism in the philosophy of mind. Also, we introduced dynamical models, focusing more on their differences, rather than similarities, from computational models. Then, in the second section of the chapter, we took into consideration and extensively discussed a pair of important difficulties faced by computational models (considered as parts of computational explanations): the problem of central cognition and the challenge of dynamical models, considered as a thoroughly alternative explanatory style. Dealing with these difficulties is indispensable if we really want to understand the mind: the future of cognitive science depends crucially on our capacity to accept this challenge. We argued that there is

In short, the appeal to dynamical models is typically invoked for integrated (sub)systems in Bechtel's sense, which are very *weakly* modular, since each of their parts is influenced by the activity in some other parts of the system. This degree of modularity, though, is hardly sufficient for those cognitive scientists who argue that a mechanistic–computational explanation requires constraints on the concept of part far more demanding than what is required for the notion of integrated system. What they need is a form of modularity that is not vulnerable to the problem of central cognition (Sect. 43.2.1). Carruthers [43.96], for instance, distinguishes a *narrow-scope* form of encapsulation from a *wide scope* variety. Influenced by the *simple heuristics* research program of Gigerenzer and Todd [43.121], he argues that the computational tractability does not require Fodor's *narrow-scope* form of encapsulation whereby – as we have seen – the encapsulated system cannot draw on information held outside of it in the course of its processing [43.96, p. 58]. Rather, the computational tractability only requires that a system is encapsulated in the sense that it can draw on the information that is present in other systems during the course of its processing but, on any given occasion, can draw only on a *subset* of the *exogenous* information – a property that Carruthers calls *frugality* or *wide-scope encapsulation*.

a promising account of central cognition, based on a development of Baars' global broadcast model; as to the problem of dynamicism, we presented a sort of *ecumenical* proposal, based on the idea of integrating mechanistic explanations with dynamical models in different degree, according to the kind of cognitive task, we have to account for. Yet, difficulties remain in realizing this integration.

In short, we can say that, despite some strong criticisms that have been addressed to the concept of computation and the related notion of representation, computational models are still at the core of the disciplines of the mind. Computational models and, more generally, mechanistic explanations are still the dominant methods in cognitive science. Indeed, on one hand, the complexity of animal and specifically human behavior requires an appropriately complex model, such as computational models, on the other hand, more traditional nomological explanations appear not to be explanatorily much apt to hit the target: psychological explanation is closer to the biological one rather than the physical one. Moreover, the alterna-

tives to computational models appear to be not very solid.

This claim should not be intended as a way of hiding certain difficulties of computationalism, as we saw in the previous section. To put it generally, it is far from being definitely established that computational explanation can be extended to all mental processes and capacities. For this reason, we believe that the appropriate attitude in the epistemology of mind is explanatory *pluralism*, at least at the moment. This has to be intended in a twofold sense. On one hand, mental phenomena require a variety of explanatory levels, whose inter-relations are of two kinds: decomposition and contextualization (in other words, we think that mechanistic explanation vindicates pluralism). On the

other hand, the arguably quasi-holistic character of some cognitive tasks suggests that the mechanistic style of explanation has to be integrated in these cases with a dynamical explanatory style.

*Cartwright* [43.122] suggested that the most appropriate metaphor of explanation in cognitive science is the *patchwork*: a disparate arsenal of modeling *weapons*. Not only each mental capacity calls for its distinctive collection of explanatory layers, but it is also necessary to relax the constraints on the overall explanatory architecture. We are essentially in agreement with such a point of view, though, at the same time, we try to maintain confidence in computationalism as a general explanatory framework in the sciences of the mind.

## References

- 43.1 W. Bechtel, A. Abrahamsen, G. Graham: The life of cognitive science. In: *A Companion to Cognitive Science*, ed. by W. Bechtel, G. Graham (Blackwell, Oxford 1998) pp. 1–104
- 43.2 N. Block: Mental pictures and cognitive science, *Philos. Rev.* **90**, 499–541 (1983)
- 43.3 D. Marconi: *Filosofia e Scienza Cognitiva* (Laterza, Roma–Bari 2001)
- 43.4 R. Bogdan: L'histoire de la science cognitive. In: *Dictionnaire Critique de la Communication*, ed. by L. Sfez (PUF, Paris 1993)
- 43.5 M. Weisberg: *Simulation and Similarity: Using Models to Understand the World* (Oxford Univ. Press, Oxford 2013)
- 43.6 J.R. Busemeyer, A. Diederich: *Cognitive Modeling* (Sage Publications, Thousand Oaks 2010)
- 43.7 M. Marraffa, A. Paternoster: Functions, levels, and mechanisms: The explanation in cognitive science and its problems, *Theory Psychol.* **23**, 22–45 (2013)
- 43.8 A. Chemero: *Radical Embodied Cognitive Science* (MIT Press, Cambridge 2009)
- 43.9 A. Turing: On computable numbers, with an application to the Entscheidungsproblem, *Proc. Lond. Math. Soc. (Ser. 2)* **42**, 230–265 (1936/37)
- 43.10 A. Church: An unsolvable problem of elementary number theory, *Am. J. Math.* **58**, 345–363 (1936)
- 43.11 F.W. Grasso, T.R. Consi, J. Atema: Biomimetic robot lobster performs chemo-orientation in turbulence using a pair of spatially separated sensors: Progress and challenges, *Robot. Auton. Syst.* **30**, 115–131 (2000)
- 43.12 E. Datteri, G. Tamburrini: Biorobotic experiments for the discovery of biological mechanisms, *Philos. Sci.* **74**, 409–430 (2007)
- 43.13 E. Datteri: *Filosofia Delle Scienze Cognitive. Spiegazione, Previsione, Simulazione* (Carocci, Rome 2012)
- 43.14 G. Piccinini: Computational modeling vs. computational explanation: Is everything a Turing machine, and does it matter to the philosophy of mind?, *Australas. J. Philos.* **85**, 93–115 (2007)
- 43.15 G. Piccinini: Computation in physical systems. In: *The Stanford Encyclopedia of Philosophy*, ed. by E.N. Zalta <http://plato.stanford.edu/archives/fall2012/entries/computation-physicalsystems> (Fall 2012 Edition)
- 43.16 G. Piccinini: Computationalism in the philosophy of mind, *Philos. Compass* **4**, 515–532 (2009)
- 43.17 M. Milkowski: *Explaining the Computational Mind* (MIT Press, Cambridge 2013)
- 43.18 D. Marr: *Vision* (Freeman, San Francisco 1982)
- 43.19 L.M. Vaina (Ed.): *From the Retina to the Neocortex: Selected Papers of David Marr* (Birkhauser, Boston 1990)
- 43.20 R.S. Jackendoff: *Consciousness and Computational Mind* (MIT Press, Cambridge 1987)
- 43.21 D. Hubel, T. Wiesel: Receptive fields, binocular interaction and functional architecture in the cat's visual cortex, *J. Physiol.* **160**, 106–154 (1962)
- 43.22 D. Hubel, T. Wiesel: Receptive fields and functional architecture of monkey striate cortex, *J. Physiol.* **195**, 215–243 (1968)
- 43.23 T. Zawidzki, W. Bechtel: Gall's legacy revisited. Decomposition and localization in cognitive neuroscience. In: *Mind as A Scientific Object: Between Brain and Culture*, ed. by C.E. Erneling, D.M. Johnson (Oxford Univ. Press, Oxford 2004) pp. 293–318
- 43.24 W. Bechtel, J. Mundale: Multiple realizability revisited: Linking cognitive and neural states, *Philos. Sci.* **66**, 175–207 (1999)
- 43.25 A. Clark, C. Eliasmith: Philosophical issues in brain theory and connectionism. In: *The Handbook of Brain Theory and Neural Networks*, ed. by M.A. Arbib (MIT Press, Cambridge 2002) pp. 886–888
- 43.26 L. Ungerleider, M. Mishkin: Two cortical visual systems. In: *Analysis of Visual Behavior*, ed. by D.J. Ingle, M.A. Goodale, R.J. Mansfield (MIT Press, Cambridge 1982) pp. 549–586

- 43.27 D.A. Milner, M.A. Goodale: *The Visual Brain in Action* (Oxford Univ. Press, Oxford 1995)
- 43.28 P. Jacob, M. Jeannerod: *Ways of Seeing* (Oxford Univ. Press, Oxford 2003)
- 43.29 E. Mayr: *Toward a New Philosophy of Biology* (Belknap, Cambridge 1988) pp. 148–159
- 43.30 L. Wright: *Teleological Explanations: An Etiological Analysis of Goals and Functions* (Univ. California Press, Berkeley 1976)
- 43.31 R.G. Millikan: *Language, Thought, and Other Biological Categories* (MIT Press, Cambridge 1984)
- 43.32 R. Cummins: Functional analysis, *J. Philos.* **72**, 741–765 (1975)
- 43.33 K. Sterelny, P. Griffiths: *Sex and Death: An Introduction to the Philosophy of Biology* (Univ. Chicago Press, Chicago 1999)
- 43.34 T. Polger: Functionalism as a philosophical theory of the cognitive sciences, *WIREs Cogn. Sci.* **3**, 337–348 (2012)
- 43.35 H. Putnam: *Mind, Language, and Reality. Philosophical Papers Volume 2* (Cambridge Univ. Press, Cambridge 1975)
- 43.36 W. Lycan: The mind–body problem. In: *Mind and Cognition*, ed. by S.P. Stich, T.A. Warfield (Blackwell, Oxford 2003) pp. 47–64
- 43.37 J.A. Fodor: The appeal to tacit knowledge in psychological explanation, *J. Philos.* **65**, 627–640 (1968)
- 43.38 D.C. Dennett: *Brainstorms* (MIT Press, Cambridge 1978)
- 43.39 R. Cummins: *The Nature of Psychological Explanation* (MIT Press, Cambridge 1983)
- 43.40 R. Cummins: “How does it work?” versus “what are the laws?” Two conceptions of psychological explanation. In: *Explanation and Cognition*, ed. by F. Keil, R. Wilson (MIT Press, Cambridge 2000) pp. 117–144
- 43.41 M. Roth, R. Cummins: Two tales of functional explanation, *Philos. Psychol.* **27**, 773–788 (2014)
- 43.42 C.F. Craver: *Explaining the Brain* (Oxford Univ. Press, Oxford 2007)
- 43.43 W. Lycan: Form, function, and feel, *J. Philos.* **78**, 24–50 (1981)
- 43.44 W. Lycan: *Consciousness* (MIT Press, Cambridge 1987)
- 43.45 W. Lycan: Functionalism (1). In: *Companion to the Philosophy of Mind*, ed. by S. Guttenplan (Blackwell, Oxford 1994) pp. 317–323
- 43.46 N. Block: *Consciousness, Function, and Representation* (MIT Press, Cambridge 2007)
- 43.47 S. Bem, H. Looren de Jong: *Theoretical Issues in Psychology. An Introduction* (Sage, London 2006)
- 43.48 E. Sober: Panglossian functionalism and the philosophy of mind, *Synthese* **64**, 165–193 (1985)
- 43.49 J.A. Fodor: Methodological solipsism considered as a research strategy in cognitive psychology, *Behav. Brain Sci.* **3**, 63–109 (1980)
- 43.50 G. Harman: Wide functionalism. In: *Cognition and Representation*, ed. by S. Schiffer, S. Steele (Westview., Boulder 1988) pp. 11–20
- 43.51 P.K. Machamer, L. Darden, C. Craver: Thinking about mechanisms, *Philos. Sci.* **67**, 1–25 (2000)
- 43.52 C.F. Craver: When mechanistic models explain, *Synthese* **153**, 355–376 (2006)
- 43.53 C.F. Craver, W. Bechtel: Mechanism. In: *Philosophy of Science: An Encyclopedia*, ed. by S. Sarkar, J. Pfeifer (Routledge, London 2006) pp. 469–478
- 43.54 C.F. Craver: Role functions, mechanisms, and hierarchy, *Philos. Sci.* **68**, 53–74 (2001)
- 43.55 W. Bechtel: Looking down, around, and up: Mechanistic explanation in psychology, *Philos. Psychol.* **22**, 543–564 (2009)
- 43.56 C.F. Craver: Beyond reduction: Mechanisms, multifield integration, and the unity of science, *Stud. Hist. Philos. Biol. Biomed. Sci.* **36**, 373–396 (2005)
- 43.57 C.F. Craver, M.I. Kaiser: Mechanisms and laws: Clarifying the debate, *Hist. Philos. Theory Life Sci.* **3**, 125–145 (2013)
- 43.58 A. Clark: An embodied cognitive science?, *Trends Cogn. Sci.* **3**, 345–351 (1999)
- 43.59 T.J. van Gelder, R. Port: It’s about time: An overview of the dynamical approach to cognition. In: *Mind as Motion*, ed. by R. Port, T. van Gelder (MIT Press, Cambridge 1995) pp. 1–43
- 43.60 S. Kelso: *Dynamic Patterns* (MIT Press, Cambridge 1995)
- 43.61 R. Beer: A dynamical systems perspective on agent–environment interaction, *Artif. Intell.* **72**, 173–215 (1995)
- 43.62 E. Thelen, L. Smith: *A Dynamic Systems Approach to the Development of Cognition and Action* (MIT Press, Cambridge 1994)
- 43.63 K. Sterelny: *The Representational Theory of Mind* (Blackwell, Oxford 1990)
- 43.64 J.A. Fodor: *The Mind Doesn’t Work that Way* (MIT Press, Cambridge 2000)
- 43.65 J.A. Fodor: *The Modularity of Mind* (MIT Press, Cambridge 1983)
- 43.66 Z. Pylyshyn: Is vision continuous with cognition? The case for cognitive impenetrability of visual perception, *Behav. Brain Sci.* **22**, 341–365 (1999), discussion pp. 366–423
- 43.67 Z. Pylyshyn: *Seeing and Visualizing* (MIT Press, Cambridge 2003)
- 43.68 J.A. Fodor: *LOT 2: The Language of Thought Revisited* (Oxford Univ. Press, Oxford 2008)
- 43.69 P. Carruthers: Moderately massive modularity. In: *Minds and Persons*, ed. by A. O’Hear (Cambridge Univ. Press, Cambridge 2003) pp. 67–90
- 43.70 S. Pinker: So how does the mind work?, *Mind Language* **20**, 1–24 (2005)
- 43.71 R. Samuels: Is the mind massively modular? In: *Contemporary Debates in Cognitive Science*, ed. by R.J. Stainton (Blackwell, Oxford 2006) pp. 37–56
- 43.72 D. Sperber, V. Girotto: Does the selection task detect cheater detection? In: *From Mating to Mentality: Evaluating Evolutionary Psychology*, ed. by J. Fitness, K. Sterelny (Psychology Press, Hove 2003) pp. 197–226
- 43.73 S. Atran: Folk biology and the anthropology of science: Cognitive universals and cultural particulars, *Behav. Brain Sci.* **21**, 547–609 (1998)
- 43.74 S. Carey: On the origins of causal understanding. In: *Causal Cognition*, ed. by D. Sperber,

- D. Premack, A. Premack (Clarendon Press, Oxford 1995) pp. 268–308
- 43.75 K. Cheng: A purely geometric module in the rat's spatial representation, *Cognition* **23**, 149–178 (1986)
- 43.76 A.M. Leslie, O. Friedman, T.P. German: Core mechanisms in 'theory of mind, *Trends Cogn. Sci.* **8**, 528–533 (2004)
- 43.77 J.L. Bermúdez: *Philosophy of Psychology* (Routledge, London: 2005)
- 43.78 S. Hurley: *Consciousness in Action* (Harvard Univ. Press, Cambridge 1998)
- 43.79 R. Axelrod: *The Evolution of Cooperation* (Basic Books, New York 1984)
- 43.80 P. Carruthers: *The Centered Mind: What the Science of Working Memory Shows us About the Nature of Human Thought* (Oxford Univ. Press, Oxford 2015)
- 43.81 B. Baars: *A Cognitive Theory of Consciousness* (Cambridge Univ. Press, Cambridge 1988)
- 43.82 B. Baars: *In the Theater of Consciousness: The Workspace of the Mind* (Oxford Univ. Press, Oxford: 1997)
- 43.83 B. Baars: The conscious access hypothesis: Origins and recent evidence, *Trends Cogn. Sci.* **6**, 47–52 (2002)
- 43.84 B. Baars: How brain reveals mind: Neuroimaging supports the central role of conscious experience, *J. Conscious. Stud.* **10**, 100–114 (2003)
- 43.85 S. Dehaene: *Consciousness and the Brain* (Viking, New York 2014)
- 43.86 S. Dehaene, L. Naccache: Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework, *Cognition* **79**, 1–37 (2001)
- 43.87 S. Dehaene, J.P. Changeux, L. Naccache, J. Sackyr, C. Sergent: Conscious, preconscious, and subliminal processing: A testable taxonomy, *Trends Cogn. Sci.* **10**, 204–211 (2006)
- 43.88 E. Bullmore, O. Sporns: Complex brain networks: Graph theoretical analysis of structural and functional systems, *Nat. Rev. Neurosci.* **10**, 186–198 (2009)
- 43.89 M.P. Shanahan: *Embodiment and the Inner Life* (Oxford Univ. Press, Oxford 2010)
- 43.90 M. D'Esposito: From cognitive to neural models of working memory, *Philos. Trans. R. Soc. B* **362**, 761–772 (2007)
- 43.91 E. Knudsen: Fundamental components of attention, *Annu. Rev. Neurosci.* **30**, 57–78 (2007)
- 43.92 J. Jonides, R. Lewis, D. Nee, C. Lustig, M. Berman, K. Moore: The mind and brain of short-term memory, *Annu. Rev. Psychol.* **59**, 193–224 (2008)
- 43.93 P. Carruthers: *The Opacity of Mind* (Oxford Univ. Press, Oxford 2011)
- 43.94 B. Baars, N.M. Gage: *Cognition, Brain, and Consciousness* (Elsevier, Oxford 2010)
- 43.95 E.C. Deise: Frame problems, Fodor's Challenge, and Practical Reason, Ph.D. Thesis (Univ. Maryland, College Park 2008)
- 43.96 P. Carruthers: *The Architecture of the Mind* (Clarendon Press, Oxford 2006)
- 43.97 T.J. van Gelder: What might cognition be, if not computation?, *J. Philos.* **92**, 345–381 (1995)
- 43.98 A. Clark: *Being There* (MIT Press, Cambridge 1997)
- 43.99 A. Chemero, M. Silberstein: After the philosophy of mind: Replacing scholasticism with science, *Philos. Sci.* **75**, 1–27 (2008)
- 43.100 A. Clark: *Natural-Born Cyborgs* (Oxford Univ. Press, Oxford 2003)
- 43.101 A. Clark: *Supersizing the Mind* (Oxford Univ. Press, Oxford 2008)
- 43.102 J. Haugeland: Representational genera. In: *Having Thought*, ed. by J. Haugeland (Harvard Univ. Press, Cambridge 1998) pp. 171–206
- 43.103 A. Clark, J. Toribio: Doing without representing?, *Synthese* **101**, 401–431 (1994)
- 43.104 W. Bechtel: Dynamics and decomposition: Are they compatible?, *Proc. Australas. Cogn. Sci. Soc.* (1997), Retrieved from: <http://mechanism.ucsd.edu/research/dynamics.htm>
- 43.105 W. Bechtel: Representations and cognitive explanations: Assessing the dynamicist challenge in cognitive science, *Cogn. Sci.* **22**, 295–318 (1998)
- 43.106 W. Bechtel: The compatibility of complex systems and reduction: A case analysis of memory research, *Minds Mach.* **11**, 483–502 (2001)
- 43.107 W. Bechtel: *Mental Mechanisms* (Routledge, London 2008)
- 43.108 W. Bechtel, A. Abrahamsen: Dynamic mechanistic explanation: Computational modeling of circadian rhythms as an exemplar for cognitive science, *Stud. Hist. Philos. Sci. A* **1**, 321–333 (2010)
- 43.109 W. Bechtel, R.C. Richardson: *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*, 2nd edn. (MIT Press, Cambridge 2010)
- 43.110 D.M. Kaplan, W. Bechtel: Dynamical models: An alternative or complement to mechanistic explanations, *Top. Cogn. Sci.* **3**, 438–444 (2011)
- 43.111 H. Simon: *The Sciences of the Artificial*, 3rd edn. (MIT Press, Cambridge 1996)
- 43.112 C.G. Van Orden, B.F. Pennington, G.O. Stone: What do double dissociations prove?, *Cogn. Sci.* **25**, 111–172 (2001)
- 43.113 W. Bechtel: Decomposing the mind-brain: A long-term pursuit, *Brain Mind* **3**, 229–242 (2002)
- 43.114 W. Bechtel: Referring to localized cognitive operations in parts of dynamically active brains. In: *Perception, Realism and the Problem of Reference*, ed. by A. Raftopoulos, P. Machamer (Cambridge Univ. Press, Cambridge 2012)
- 43.115 D.J. Felleman, D.C. van Essen: Distributed hierarchical processing in the primate cerebral cortex, *Cereb. Cortex* **1**, 1–47 (1991)
- 43.116 D. van Essen, J.L. Gallant: Neural mechanisms of form and motion processing in the primate visual system, *Neuron* **13**, 1–10 (1994)
- 43.117 J.W. Lewis, D.C. van Essen: Corticocortical connections of visual, sensorimotor, and multimodal processing areas in the parietal lobe of the macaque monkey, *J. Comp. Neurol.* **428**, 112–137 (2000)

- 43.118 D. van Essen: Organization of visual areas in macaque and human cerebral cortex. In: *The Visual Neurosciences*, Vol. 1, ed. by L.M. Chalupa, J.S. Werner (MIT Press, Cambridge 2004) pp. 507–521
- 43.119 G.A. Orban, D. van Essen, W. Vanduffel: Comparative mapping of higher visual areas in monkeys and humans, *Trends Cogn. Sci.* **8**, 315–324 (2004)
- 43.120 D.M. Kaplan, C.F. Craver: The explanatory force of dynamical and mathematical models in neuroscience: A mechanistic perspective, *Philos. Sci.* **78**, 601–627 (2011)
- 43.121 G. Gigerenzer, P. Todd: *the ABC Research Group: Simple Heuristics that Make Us Smart* (Oxford Univ. Press, Oxford 1999)
- 43.122 N. Cartwright: *The Dappled World* (Cambridge Univ. Press, Cambridge 1999)

## 44. Model-Based Reasoning in the Social Sciences

Federica Russo

Social scientists use different types of model to reason about social objects and to study social phenomena. In this chapter, I provide an overview of various forms of model-based reasoning in social research, especially quantitative and qualitative. In the course of the chapter, I highlight differences with other variants of model-based reasoning, notably the one inherited from logical positivism, and I discuss the use of experiments and simulation in social contexts. The chapter also investigates intersections between model-based reasoning and other notions, such as explanation and causality, truth and validity.

<b>44.1 Modeling Practices in the Social Sciences</b> .....	954
44.1.1 Social-Scientific Objects .....	954
44.1.2 Quantitative Modeling .....	955
44.1.3 Qualitative Modeling.....	957
44.1.4 Experimental and Quasi-Experimental Modeling .....	957
<b>44.2 Concepts of Model</b> .....	958
44.2.1 Models as Representations .....	958
44.2.2 Models as Objects .....	960
<b>44.3 Models and Reality</b> .....	962
44.3.1 Mediators .....	962
44.3.2 Isolations.....	962
44.3.3 Maps.....	963
<b>44.4 Models and Neighboring Concepts</b> .....	963
44.4.1 Simulations .....	963
44.4.2 Causation and Explanation.....	964
44.4.3 Truth and Validity.....	965
<b>44.5 Conclusion</b> .....	967
<b>References</b> .....	968

The notion of model occupies an important part of the debate in the philosophy of science. One reason why this notion deserves so much attention is that models lie at the interface between the epistemic agent (in this context, the scientist) and the system under investigation (be it physical, biological, or social). Thus, models allow us to study, understand, and interpret the surrounding reality. The literature is incredibly vast, spanning very many scientific disciplines as well as philosophical traditions. As a consequence, providing an exhaustive summary of the contributions, or finding the main conceptual research lines, is far from being an easy task. A further difficulty is that the term *model* is currently used by scientists coming from different backgrounds, as well as by logicians and philosophers having different perspectives on the topic. This created distortions that might be dubbed *episodes of conceptual imperialism*: a discipline and its basic concepts are used as benchmarks to evaluate or discuss another discipline and its basic concepts. This has often happened

in the social sciences, for instance in the debate about their status, that is, whether they rightly belong to the realm of the sciences, alongside the natural sciences, and physics in particular. Thus, the social sciences have been often judged according to the standards of *other* disciplines, typically physics. What is at stake here is a question about their methods – and, consequently, about their *objectivity* – typically so different than those of the natural sciences. For this reason, it is worth opening the chapter with an overview of the methods and models used in the social sciences, before discussing in detail the concept of model.

The chapter is organized as follows. In Sect. 44.1, I examine the objects of study of the social sciences and different types of models: quantitative vs qualitative, experimental vs quasi-experimental. To be sure, this is not the only possible categorization of social science models, and I shall refer, when appropriate, to other types of models, such as theoretical models or simulations. In Sect. 44.2, I distinguish two concep-

tions of models: as representations and as objects. For each of these, I examine two variants. For models as representations (Sect. 44.2.1), I consider the notion of set-theoretic structures and of a family of probability distributions. For models as objects (Sect. 44.2.2), I analyze those positions that interpret models as fictional entities and as epistemic objects.

The idea that a model is a representation of a given system is, in some way, the received view which any other position subsequently developed has to confront with. In particular, the first conception, namely representation in the set-theoretic sense, is the most discussed one in philosophy of science. The second conception (family of probability distributions) corresponds instead to what a statistician would think of as a model. As we shall see, there is a specificity in the use and meaning of the concept of model in the social sciences that is not entirely captured by the received view and that is instead captured by the idea that models are families of probability distributions *and* epistemic objects.

In Sect. 44.3, I discuss a selection of philosophical positions concerning the relationship between models and reality. I focus on three positions: models as mediators, isolations, and maps. These positions should not be seen in opposition with each other, and, in fact, they complement each other, as they highlight different aspects of modeling practices and of model-based reasoning in the sciences.

Finally, in Sect. 44.4, I examine the concept of model in relation to types of model that had not been discussed earlier in Sect. 44.1 – viz. simulations – and to some central issues in model-based reasoning: causality and explanation, validity and truth. The philosophical questions tackled in this section arise in social research and in any other scientific domain alike, but here I will privilege discussion on issues concerning modeling in the social sciences.

Before starting, it is worth lingering on the position of this essay in the *geography* of the philosophy

of science. To begin with, the choice of starting from a thorough description of the scientific practice is not accidental. In fact, I adhere to the view according to which philosophical discussions must be rooted in a practice, or a problem, or an issue that emerges in contemporary science or in the history of science [44.1, 2]. This, however, does not rule out a priori philosophical investigations as an important part of the process (for a nuanced view on the relations between science and metaphysics, see, for example, [44.3]). I will then examine those philosophical positions that help us clarify controversial aspects or important conceptual issues, as they arise in the scientific practice. The comparison with the received view (model in the set-theoretic sense) is important for two reasons. On one hand, the methodological literature in social research developed its own version of the representation view (i. e., as a family of probability distributions); on the other hand, such comparison will help us foster a dialogue between different subdisciplines in philosophy of science.

It is also necessary to locate this essay in another area, namely within the (philosophy of the) social sciences themselves, and especially within the tradition of hermeneutics, of historicism, and of critical theory that developed since the mid-nineteenth century. There is no doubt about the value of the methodological contribution of authors such as Wilhelm Dilthey, Theodor Adorno, Jürgen + Habermas, or Max Weber. No doubt the scientific practices presented in Sect. 44.1 have, in different ways, a conceptual debt to these traditions. However, these will not be examined here.

Finally, I felt it important to give voice to scientific practices that are relatively less discussed in the literature, such as qualitative models (Sect. 44.1.3). However, the reader will notice that the arguments discussed in later sections (notably, Sects. 44.2 and 44.3) focus more on quantitative than qualitative modeling, which makes the discussion somewhat incomplete. Yet, the hope is to draw the attention to a form of model-based reasoning in need of further philosophical inquiry.

## 44.1 Modeling Practices in the Social Sciences

### 44.1.1 Social-Scientific Objects

The social sciences study individuals and societies from different perspectives; to do so, they use methods and approaches that are highly heterogeneous. Demography, for instance, studies populations according to the parameters of fertility and mortality, morbidity, and migration. It does so with the help of statistics and of quantitative analysis of data, which allow us to get

a snapshot of a population, to see how it changes over time, and to predict how it will be at some future time. Economics focuses on the behavior of individuals and of groups (a family, a company, a market, or a state) with respect to the management of resources. Sociology is interested in the social behavior of individuals and groups, identifying specific contexts and environments as the sociology of work, of science, or of health. Anthropology studies human beings within a given society

in their various dimensions, such as cultural, emotional, or spiritual. Epidemiology deserves a special note, as it is at the border that separates the social sciences from the biomedical sciences. Epidemiology studies the distribution and variation of mortality and morbidity of the population, according to biological and socioeconomic characteristics of individuals. Needless to say, the scientific objects of these disciplines do not always have sharp and precise contours. Understanding and conceptualizing the objects of study of the social sciences is in itself an interesting and important topic – see for instance [44.4].

The label *social science* has been persistently used since the mid-nineteenth century, which is when the social sciences *come into existence* as autonomous fields of scientific investigation. Yet, the social sciences are, in fact, highly heterogeneous both in their objects and in their methods. On one hand, some social scientists, for instance sociologist Émile Durkheim, worked hard to establish the autonomy of their discipline with respect to other areas of investigation (notably, psychology). On the other hand, economics has an object of study that clearly falls within the realm of the social, and yet it is often considered apart from other social sciences. Examples abound. This is the case, for instance, in disciplinary panels for the allocation of national or European funds. To give another example, this series of volumes on model-based reasoning or the Italian online journal *APhEx* devoted specific essays to *models in economics* ([44.5] and Chap. 19) and to *models in the social sciences* (Chap. 52, [44.6], and this chapter). Consequently, it is difficult to find *one* concept of model that fits the social sciences *as a whole*; nevertheless, it is possible to isolate some common conceptual points.

In the following, I will present models and approaches in the social sciences according to a classification that is quite widespread and that distinguishes two lines of research that are seemingly very different: *quantitative models*, based on statistical analyses of data (Sect. 44.1.2), and *qualitative models*, typically based on the direct study of small groups of individuals (Sect. 44.1.3). As we will see, however, this does not mark a difference between models that are intrinsically better, more robust, or more valid. The quality of models, or rather their *validity*, is to be assessed on different grounds and concerns them all – see later Sect. 44.4.3. In the social sciences, quantitative models are very often *observational* models, in the sense that, once data is collected, it is analyzed using the tools of probability theory and statistics, but the process of data generation is not repeated, as can happen in experimental modeling practices instead. Experimental models may generate data repeatedly, control certain experimental conditions in labs with more precision,

and use sophisticated instrumental apparatus for measurement. To be sure, experimental models are used in the social sciences too, and I illustrate some aspects of this modeling practice in Sect. 44.1.4. The reader interested in learning about scientific practices inside and outside the laboratory will find [44.7] an inspiring read. Another possible classification distinguishes between empirical and theoretical models. Empirical models use empirical data, collected in different ways, while theoretical models are developed theoretically, that is, in the abstraction of any empirical investigation and are sometimes formalized. Theoretical models are often used in economics and try to reconstruct, a priori, various processes involved in economic behavior. One example is Schelling’s model of segregation or Friedman’s hypothesis on permanent income. I will not discuss these models in detail – for an extensive discussion, see [44.5] and Chap. 19.

### 44.1.2 Quantitative Modeling

Quantitative analysis has a long tradition, tracing back to pioneering methodologists such as Adolphe Quetelet, demographer and astronomer, and Émile Durkheim, sociologist, both active in the second half of the nineteenth century. During the first half of the twentieth century, and until the 1970s, the social sciences saw continuous improvements and refinements of techniques for data analysis, developing increasingly sophisticated statistical models and tests. Several reknown scholars have been protagonist of these developments, for instance, Sewall Wright was active in the field of population genetics in the first half of the last century, or Otis Dudley Duncan and Raymond Boudon both sociologists have been active since the 1960s and 1970s of the last century. In the last 30 years or so, economists, statisticians, and computer scientists such as James Heckman, Kevin Hoover, Judea Pearl, Clark Glymour, or Donald Rubin (and their collaborators) prompted further progress in the use of probability, statistics, and automated reasoning for the study of the social.

Probability theory and statistics provide us with useful means to analyze random phenomena like flipping a coin, waiting time in a queue, or other more complex social phenomena such as migration or changes in the morbidity of a population.

Data analyzed in a statistical model typically come from censuses, surveys, or other similar methods. But what is data? Data consist of observations or measurements of characteristics of populations or individuals under study. This seemingly simple characterization, hides a conceptual complexity which I will only gloss over. The generation, use, and re-use of data are all activities that deserve a deep epistemological, method-



ological, and metaphysical reflection (for a thorough discussion, see, for instance, [44.8, 9]). It is worth emphasizing that data are not *just* data, but are already heavily *theory-laden*, to borrow the famous expression of N.R. Hanson. Some data are generated in a rather simple and noncontroversial way. For instance, nowadays population registers in Western countries allow us to determine the age of an individual very easily and precisely. Other data, however, are much harder to generate. For example, there is not a unique way of measuring *socioeconomic status* or the level of education of individuals. Similarly, other data are indirectly generated by measuring *other* characteristics. For instance, school motivation can be measured by recording class attendance.

Once generated, data are then organized and grouped according to variables. Here is a possible taxonomy of variables (see also [44.10, Chap.3]):

1. Gender and scale: continuous/discrete; quantitative/qualitative
2. Role: observed, latent, instrumental, proxy
3. Level: individual, aggregate
4. Scope: socioeconomic, demographic, biological, etc.

Using a type of variable rather than another depends on reasons that can be methodological, empirical, or other. Variable *age*, for instance, is clearly a continuous variable, but for convenience scientists analyze populations by *age structures* (taxon 1). Age, in turn, can provide information regarding socioeconomic characteristics of individuals – for example, a 10-year-old person is likely to be in school age and a 30-year-old to be in work age; age may also provide information on biological aspects – for example, we record hearing loss typically after 60 years of age, and yet with the increased use of headsets, the phenomenon begins appearing much earlier, thus providing information on *behavioral* aspects (taxon 4). Thus, *age* can be an observed (and directly measured) variable, or a proxy, that is, a variable that *stands for something else* and that is not directly measured (taxon 2). Finally, most characteristics are measured for each individual in the sample or population, for example (individual) income; however, some phenomena are better modeled using aggregate variables, for example, the average income for a given population (taxon 3).

Once data are organized, we have to organize the variables. This is the task of *quantitative* models. Typically, a quantitative model consists of (a system of) equations and of a graphical representation of these equations. The social sciences have, in the course of time, developed more sophisticated quantitative models, tailored to specific scientific problems. For instance,

multilevel models are designed specifically to study the relationships between *individual* and *aggregate* variables. As mentioned above, an individual variable measures a certain characteristic for each individual in the sample, for example, *individual* income. An aggregate variable *sums up* individual measurements into aggregate measurements, such as mean income at the regional or national scale. Or, structural models are specifically used to model the *structure* of the relationships between variables, which in turn serves to explain the socioeconomic mechanism(s) underlying a given phenomenon.

Quantitative analysis can be used to provide a description of a social phenomenon. In this case, the model studies how a variable *changes* depending on how other variables change. For example, we can trace how the variable that records births in Alsace varies depending on the variable that records the presence of storks in the same region, thus establishing a correlation (or co-variation) between the two variables. At this point, the model only attests to a statistical dependence between the variables. This dependence is symmetrical: the birth rate in Alsace changes depending on the change in the number of storks in the region. But we can reverse the equation: the number of storks in Alsace changes depending on the change in the birth rate.

As is well known, correlation alone does not allow us to determine whether there is a causal relationship between two variables, and in which direction it flows, nor it allows us to explain a phenomenon. This is because of the *third variable problem*. Given a correlation between two variables, it is possible to find a third variable such that, when included in the model, makes the correlation disappear. Let me illustrate with a toy example. Yellow fingers are correlated with lung cancer. Include now the variable *cigarette smoking* in the model. It is easy to show that the correlation disappears because cigarette smoking is the cause of *both* yellow fingers and lung cancer. The philosophical literature discussed the issue under the headings of *screening off* and of the *common cause principle*, while the methodological literature more often talks about *confounding variables* and *statistical control*.

How to infer causation from correlation is a vexata quaestio in philosophy of causality. Positions proposed in the literature vary, but there seems to be agreement at least on the fact that a causal model has more constraints than an associational model. In other words, causal models have some *augmented* technical features compared to an associational model, notably with respect to assumptions, the types of test, and the use of background knowledge (for a discussion, see, for example, [44.11]). I will briefly mention just two types of tests. Exogeneity tests are used to check if cause

and effect are properly separated, that is, if the (probabilistic) structure of the model is correct from a causal point of view. Exogeneity is typically explained by saying that exogenous variables are caused *outside* the model, while the endogenous ones are caused *within* the model (by the exogenous variables) – for a discussion, see [44.12]. Invariance tests, instead, are used to check that the causal structure is sufficiently stable across different partitions of the population of reference, or under interventions or manipulations [44.13].

Introductions to quantitative modeling and to the concepts mentioned above, also accessible to non-experts in probability and statistics, are available in [44.10, 14].

### 44.1.3 Qualitative Modeling

Qualitative models are mainly used in ethnography and anthropology, in some branches of sociology, and also in educational science. A first difference with the quantitative models is of *scale*: the larger the sample size of a quantitative model, the better its quality. The same, however, does not necessarily hold for qualitative models. A second difference concerns the *techniques* for data analysis. While a quantitative model typically uses the tools of probability theory and statistics, in a qualitative model, researchers select small groups of individuals and study them in detail, for example, by integrating into their community and observing them *from the inside*. In this case, data not only *quantitatively* measure certain characteristics, but also *qualitatively* describe social practices, behaviors, language use, etc.

It is worth noting that this does not necessarily mean traveling to remote or distant places to study particular ethnic groups. Ethnographic research also concerns the societies close to us, whether geographically or culturally. Ethnographers and sociologists are interested, for instance, in how young people socialize in the digital era in Western countries, or how doctors interact as a team in an operating room, or in the way in which citizens can be part of decision-making processes relating to the environment or the like – on ethnography conduct *at home*, see [44.15]. What triggers anthropological interest is the object of anthropology itself – on this, see, for example, *Montuschi* [44.4], who discusses the distinction between ethnography and anthropology, or *Eriksen and Nielsen* [44.16], who introduce the concept of *home blindness*, viz. the difficulty of seeing and studying our own culture, because we belong to it.

It is a widespread misconception that qualitative methods are less rigorous than quantitative ones, but this view is wrong. Scientific rigour is not an intrinsic feature of a method. Rigour is instead a property of the *process* of project design, model building and test-

ing, and of the interpretation of results, and it depends on how scientists *in practice* carry out such a process. *Cardano* [44.17] offers an interesting presentation of qualitative methods and explains, step by step, what are the aspects to be taken into account in the preparation of an ethnographic study: what individuals to study, when and how long for, what are the assumptions, how to perform tests on the hypotheses, how to use theoretical assumptions in the interpretation of results, etc. All these are elements that belong to the long and complex process of modeling, even when formal methods or quantitative models are not at stake.

Cardano's approach is interesting because, if we have previously pointed to some of the differences between the quantitative and qualitative models, we can now identify a similarity. The way Cardano describes ethnographic research is perfectly in line with the precepts of modern scientific method and of a hypothetico-deductive methodology: the formulation of hypotheses, data collection and data analysis, hypothesis testing, and validation of the model. Therefore, the difference in the techniques for data analysis (quantitative or qualitative) does not draw a line between scientific and nonscientific, objective and nonobjective. For a discussion on the issue of objectivity, see also [44.18].

### 44.1.4 Experimental and Quasi-Experimental Modeling

Models presented earlier belong to the category of *observational models*, which are often opposed to *experimental models*. I will use, in this context, the term *experimental model* to refer to those modeling strategies that make use of experiments. Often, experimental models are conceptually associated with the natural sciences (physics, biology, etc.) and observational models with the social sciences and humanities. This is not entirely correct. Psychology, for instance, increasingly uses experimental methods to study the mechanisms that regulate certain phenomena, such as attention or memory. Economics too uses experiments, as it seeks to develop and validate economic theories based more on empirical data and less on theoretical hypotheses such as the *homo oeconomicus*, who always maximizes expected utility and has perfect knowledge. The relations between theory, experiment, and reality raise many questions, some are epistemological, others are methodological and, of course, some others are ethical and moral. The literature is vast. The following contributions touch on various issues related to the use of experiments in social sciences and economics: [44.19–24]. These issues are all the more pressing in areas where experimentation has severe limitations for ethical or practical reasons, as in the social sciences.

To illustrate the use of experiments to model social phenomena, consider the case of the *invisible gorilla*, a famous experiment in psychology [44.25]. The researcher asks a person to observe two teams playing and to count the number of times that the players pass the ball. At the end of the game, the researcher asks the person whether he or she has also seen a gorilla walking through the playing field. In about half of the cases, the gorilla goes unnoticed. This is because attention works in a very selective way. Researchers talk, in cases like this, of *inattentional blindness*. The experiment has been repeated several times, with significant variations, in order to confirm the stability of the results and of the assumptions. This allowed researchers to capture different aspects and dimensions of the phenomenon of attention. The results of studies go far beyond the understanding of the phenomenon itself. For example, results of the *invisible gorilla* have been used for campaigns to raise awareness about the problem of *invisible cyclists* among city drivers. Experiments like the *invisible gorilla* may have other functions within the modeling process. For instance, they can be used to test a theory, or to formulate more precise assumptions, or to examine different aspects and dimensions of the same phenomenon. These issues are, in different ways, aspects of external validity (see also [44.26, 27], and Sect. 44.4.3).

Of course, experimentation in the social sciences, as well as in the biomedical sciences, is subject to significant restrictions. We cannot force people to smoke to study the effects of nicotine, as we cannot force people to work 20 hours per day to understand the effects of stress. An ethnographer can, however, go to the City of London and study the behavior of young and ambitious people who try to make a career in a prestigious financial company. Experimentation in the social sciences (as well as any other method) has limits. For this reason,

observational models, both quantitative and qualitative, are an invaluable resource for studying those phenomena on which we cannot directly intervene.

Modeling social phenomena – whether concerned with economic, medical, or psychological dimensions – also make use of quasi-experiments or natural experiments. There are empirical studies designed to assess the impact of an intervention (e.g., a socio-economic or public health policy, or a natural event) in a given population. The basic idea is the same as in randomized trials (RCTs), but with an important difference. In quasi- or natural experiments, the allocation of individuals to the treatment is quasi-random. Potential outcome models also use techniques called *propensity scores* to match individuals for the cases and the controls. Let us reason about an example. Suppose we want to study the effects of attending private or public school on income. Researchers will select individuals that are as similar as possible for most characteristics (age, social class, family situation, etc.) and differ only in the type of school attended. The two groups will then be compared in order to find differences in income that are due to the type of school attended.

Sometimes, however, this quasi-random allocation is made by nature, or by the course of events. A famous example is the outbreak of cholera in London in 1854, stopped by epidemiologist John Snow. Snow was able to stop the epidemic because he figured out that the exposure to the bacterium, resulting in contraction of the disease, was associated with the use of public water pumps. These were, in fact, served by two aqueducts, which filtered water in a different way, the one holding the bacterium and the other releasing it into the water. Two groups were naturally created, the exposed and the nonexposed, and they could be studied and compared *as if* a real experiment had been performed.

## 44.2 Concepts of Model

In the previous section, I offered an overview of the models used in social sciences. Yet, two crucial questions were left in the background: What is a model? And what is it for? I will answer these questions starting from the most classic position offered in philosophy of science: models are representations. Thus, I will be able to locate the debate on models in the social sciences within the broader framework of the debate in general philosophy of science. Moreover, I will be able to isolate some peculiarities of model-based reasoning in social research.

### 44.2.1 Models as Representations

According to an established tradition, a model is a representation of a phenomenon or of a certain portion of reality. Models can be represented in at least two ways. Let us examine them in order.

#### Set-Theoretic Structures

A model is a representation of a phenomenon, or of a certain portion of reality, in the sense that it captures the main features of the phenomenon, and expresses

them in a formal manner. Set-theoretic or mathematical models fall under this category. Here, a model consists of a set of statements having a set-theoretic structure. Statements are verifiable, either directly, because they contain terms that refer to observable entities, or indirectly, because they contain terms referring to theoretical entities, for which we have correspondence rules that bring us to observational statements. The motion of a pendulum, the motion of particles like electrons or protons, or the Higgs mechanism, are examples of this sense of model. This characterization, however, does not so much answer the direct question: *What is a model?*, but rather the question about the nature of scientific theories, in particular physical theory.

This sense of *model* is clearly a legacy of logical positivism, which based the methodology of science on the idea of meaningfulness and verifiability. Let me elaborate on this point. Neopositivists were interested, among other things, in the nature of knowledge, particularly scientific knowledge. This can be read as a legacy of the first Wittgenstein. In the *Tractatus*, Wittgenstein claims that “To understand a proposition means to know what is the case if it is true” (Proposition 4.024). To understand a proposition, we have to establish a relationship between language (expressed in well-formed formulas) and the world. This position found fertile ground among the neopositivists, who applied it to scientific knowledge. On this approach, scientific knowledge can be expressed in well-formed formulas and verified against empirical experience. This was the origin of the formulation of the verification criterion for scientific statements and of the criterion of demarcation between science and nonscience as developed by Karl Popper. Both criteria greatly influenced the philosophy of science in the years to come.

Thus science, according to the neopositivists, *produces* theories; but what is a theory? The short version of the answer may be formulated as follows: theories are sets of statements that must meet very specific requirements. What requirements? Those developed by the neopositivists (and inspired, by and large, by Wittgenstein). The verification criterion occupies a special place: a theory is scientific if it is verified to a sufficiently high degree. Later on, Popper proposed replacing this criterion with a criterion of falsification. A theory is scientific if it is falsifiable, that is to say, if, from the theory, we can deduce observational statements that we can empirically falsify. Both of these criteria are based on a specific understanding of *theory*. Theories are not just sets of statements. For a set of sentences to be considered a theory, it must have a certain structure, in particular a set-theoretic structure. Without going into the technicalities, this means that from a certain set of axioms (for which no proof

is required) and following certain rules of inference (in particular, deduction), we can prove other statements as theorems. Basically, the set-theoretic structure is what gives *certainty* to the theory and, consequently, to scientific knowledge.

Let us get back to the structure of the theory. We defined a model as a set-theoretic structure. More generally, a model is an abstract structure, such as a mathematical structure, or a set of statements formalized in first-order logic or other logic. This makes the starting axioms of the model true, in the sense that we will now see.

To see if what happens corresponds to what the theory would predict, we need corresponding rules. For instance, if the theory states that Fuji apples are red, we need to establish a correspondence between the theoretical terms *Fuji apple* and *red*, and for all objects covered by such terms. We will then be in a position to tell if it is true that Fuji apples are red and not, for example, green or yellow. The difficulty lies not so much in verifying properties of observable entities, but of the unobservable ones. These difficulties are also at the basis of the realism–antirealism debate. Think of physical theories developed in the first half of the last century and the difficulty of establishing, for example, the correctness of statements about electrons or about other theoretical entities not directly observable. Consider, also, all the theoretical and experimental apparatus to confirm the existence of the Higgs boson, recently *found* by researchers at CERN.

In philosophy of science, the position that models are representations or structures (in the sense explained above) has also been developed by *Suppes* [44.28], *van Fraassen* [44.29], *French* and *Ladyman* [44.30], or *Boniolo* [44.31]. For a recent discussion of representation through mathematical structures, see also *Pincock* [44.32] or *Molinini* [44.33].

The idea that a model represents a phenomenon, or a portion of reality, intuitively captures some aspects of modeling processes described in Sect. 44.1.2. However, the sense in which models in the social sciences represent a certain reality is not given by the set-theoretic structures. So let us examine a possible alternative.

### Families of Probability Distributions

In the social sciences, there is a sense in which models – especially quantitative models – *represent*. If you ask a statistician what a model is, the answer will most likely be that a model consists in a family of probability distributions. These probability distributions, in turn, represent some aspects of the reality under examination, in a way that I will now explain.

Let us consider an example. Suppose we run a survey to see how Italians are doing. To do this, we can

measure their well being with a number of indicators in additions to the gross domestic product (GDP). This means that we will try to measure well being not only from an economic perspective, but also physical or psychological. A study like this was discussed back in February 2012 in front of the Italian Parliament, where the president of Italian National Institute of Statistics (ISTAT) presented different ways to measure well being. *Neodemos*, an online journal of demography, published a popular science article about the ISTAT study, offering also the point of view of the social sciences on the issue [44.34].

Let us try to reconstruct the key moments of the modeling process of a study like the one just mentioned. First, we collect the data and organize the observations (i. e., the answers of the respondents) into variables. To answer the original question, we have to understand the structure of the relationship between these variables. Such models are called *probabilistic* because the probability distributions are related to the variables in the database (see also Sect. 44.1.2). A probability distribution is a function that assigns a probability value to each of the possible values of a variable. To say that a model is a family of such distributions means *putting together* the probability distributions for each of the variables in the database and then studying their behavior.

The use of probability theory and statistics to study phenomena (social or natural) presupposes a *stochastic representation* of reality, rather than a deterministic one. This is mirrored in the inclusion of *error terms* that can stand for measurement errors, latent variables, or even for the fact that the phenomena are genuinely indeterministic [44.35]. In other words, it can be argued without contradiction that a *phenomenon* is deterministic, and that our *representation* of the phenomenon is stochastic. A corollary of this position is that the representation of a phenomenon is *partial* – the presence of terms of errors or latent variables means that we cannot take into account all possible aspects of the phenomenon.

It is worth explaining this sense of model, also with respect to the distinction made in the previous section between *theory* and *model*. In set-theoretic models, there is a close relationship between the two. In some ways, the model is the formal part of a theory, notably of a physical theory and, typically, for each physical theory we admit more models, or interpretations – think, for instance of the various models, or interpretations, of quantum mechanics.

However, in the social sciences, which interest us here, we are in a different situation. With the exception of economics and some branches of sociology, the social sciences do not have theories. Or, at least, they do not have strong ones (for a discussion, see, for exam-

ple, *Wunsch* [44.36]). It is also worth noting that the majority of theories developed in classical economics fall into the category of *theoretical models* (mentioned at the end of Sect. 44.1.1) but that will not be discussed here. Instead, the empirical models presented in Sect. 44.1 are used precisely to develop theories of the social (broadly understood) through the analysis of empirical data. Think of studies on migratory movements that are carried out systematically in different countries and at different times. One goal is to try to formulate a *general* theory of migration, that can be applied to different populations, times, and cultures. Part of the difficulty in developing such general theories in social contexts is due to the object of study: human behavior changes, and it does so very quickly in time and space, across cultures, and also as a result of the implementation of socioeconomic policies. Of course, this distinction is not, in scientific practice, so sharp. Empirical models are sometimes used to test and refine theoretical models. Yet, it is still controversial whether theoretical models (like rational choice theory) should be amended on the basis of empirical studies in behavioral economics or psychology.

To conclude, the idea that a model is a representation of a given reality belongs to the classic philosophical debate about the nature and function of models as well as to the methodological literature in the social sciences. However, as the previous discussion hopefully showed, this idea is cashed out, in the social sciences, in a way that has some important differences with respect to the neopositivist position.

#### 44.2.2 Models as Objects

In the literature, another position has also been proposed to answer the question about the nature of the models. Models, according to this position, are *objects*. More precisely, according to one variant of this position, they are fictional entities, while according to another variant they are epistemic objects. We will see how this second account, in particular, offers interesting insights for the social sciences.

##### Fictional Entities

The sciences produce different types of model. Some are *physical* objects, such as a relief map of an archeological site, or a globe. Many others, however, are *abstract* objects, such as Bohr's atomic model, or the model of the inverted pendulum.

Both characterizations of *model* presented earlier (in the set-theoretic sense and as a family of probability distributions) hinge upon the idea that models are representations of reality. In turn, these representations are

structures (set theoretic or probabilistic). According to philosopher Frigg [44.37], however, this conception of models leaves unanswered the question about their *nature*. What kind of *object* is a model? Frigg proposes to conceive of models as *imagined physical systems*. Models are *hypothetical* entities that have no actual space-time existence but they are not mere set-theoretic structures either. In Frigg's words, "they would be physical things, if they were real" [44.37, p. 253].

There are, according to Frigg, two reasons to embrace his thesis. The first is that it better mirrors the use that scientists make of *model* – here, *model in physics*. Frigg comments on popular physics textbook, written by Young and Freedman [44.38]. The two scientists explain that the physics model describing the motion of a baseball abstracts from a number of aspects present in the real system, such as the air friction or the mass of the ball. This makes complex systems manageable. In such a description, says Frigg, we find no reference to the mathematical structure of the system, but rather to a simplified hypothetical situation. Mäki's account that will be examined later in Sect. 44.3.2 also emphasizes aspects related to abstraction and isolation.

The second reason is more fundamental and has to do with the relationship between the *structure* and the *real system*. The problem, for Frigg, is that there is no relationship of *morphism* (isomorphism, homomorphism, ...) between the structure and the real system. These types of relationships hold between two *structures*, but not between a structure and a worldly system. This should prompt us to re-think the relationships between mathematical representations of a system, models, and worldly systems. Frigg does not abandon completely the notion of representation. Mathematical representations *are* part of the modeling process. But his argument is that the model is *not* a representation. Models are simplified and idealized systems, distinct entities that share many of the characteristics of fictional entities in fiction, just like Sherlock Holmes or any other character or object in a novel.

This position emphasizes the role that abstraction plays in the process of modeling (for a discussion, see also [44.39]). Making assumptions about the a-dimensional nature of atoms, or about the absence of friction in the motion of the pendulum means eliminating some empirical elements and reasoning about a distilled version of reality, which typically is too much complex to be modeled as such. Some of these aspects are also discussed by other authors, and their positions will be discussed in Sect. 44.3. Frigg's position, however, stays silent on an important aspect: the role that these objects (the models) play in several activities carried out by the epistemic agent (typically, the scientist).

This is instead explicitly discussed by Tarja Knuuttila, as we shall now see.

### Epistemic Objects

Tarja Knuuttila, together with other scholars, proposed considering models as *epistemic objects* [44.40–42].

Models are *objects* because they are concrete, tangible products that we can manipulate in different ways. We can not only manipulate a physical model as a globe, but we can also manipulate a theoretical model, for example, by changing or setting the value of a variable. For Knuuttila, it is important to highlight what aspects of modeling allow us to produce knowledge. Consequently, in her account, it is not vital to distinguish different types of manipulations on the models. It is instead important to isolate those elements common to the various practices of modeling.

Models are *epistemic* objects because they mediate between the epistemic agent and the system examined, and because they provide an understanding of the phenomenon. This idea, as we shall see, is central also in another debate, that is about the relationships between models and reality. Knuuttila's account is, in fact, closely related to the one developed by Morgan and Morrison (Sect. 44.3.1).

This account of model is largely *instrumentalist*, but not so much in the classic sense of the term, that is, leading to antirealist positions about models. Rather, they are instrumentalist in analogy to the role of technology. Models are *tools* that we build, manipulate, and use to gain the knowledge of a given phenomenon. In this sense, they share many of the properties of technological artifacts. Of course, we lose a clear demarcation between the scientific objects and the tools to acquire knowledge about them. The ontology of model also becomes less neat, and the boundaries between the natural and the artificial are now blurred. But, perhaps, this is a price worth paying, if the expected gain is a better understanding of the scientific practice. The instrumental role of the models will be further discussed in Sect. 44.3.1.

At first sight, this position may seem rather unconventional, especially if one is used to engage with the mainstream literature, according to which models provide us with knowledge because they *represent* (in one way or another) a system. However, if we expand our horizon, it will not be difficult to see that this position fits very nicely in another stream of philosophy of science, one that is interested in more practical – and less abstract – aspects of the scientific practice. Paradigmatic contributions in this area have been those of *Hacking* [44.43], *Daston* [44.44], and *de Regt* et al. [44.45], to name just a few.

## 44.3 Models and Reality

The account of models as epistemic objects allows us to introduce the next theme of this chapter: the relationship between models and reality. In fact, besides the question of the nature of the model, it makes sense to inquire about its function within the process of acquiring knowledge of reality. The first two accounts examined below have been developed in the philosophy of the social sciences (specifically, philosophy of economics). The third account is more general in scope, and it offers some interesting ideas for our topic.

### 44.3.1 Mediators

The first approach I consider is the one developed by *Morgan* and *Morrison* [44.46]. The purpose of the discussion of Morgan and Morrison is to clarify the dynamics of the construction of models, their function, and their use. In particular, Morgan and Morrison try to articulate the idea that models have *autonomy* and that their function in scientific practice is to be *mediating instruments*.

To begin with, models are said to be *autonomous*. But with respect to what? Models, according to Morgan and Morrison, have partial autonomy compared to the *theories* on the one hand, and the *reality* on the other hand. But, note, partial autonomy also means partial *dependence*. This partial autonomy (and dependence) is already clear at the model-building stage. According to some schools of thought (especially in economics), models are derived entirely from theory (for a discussion, see, for example, [44.47]), and according to others models are instead entirely bootstrapped from data alone (as in data mining). But Morgan and Morrison argue that *both* theory and data are involved (as well as other elements that do not concern us here).

Morgan and Morrison also want to defend the autonomy of the *function* of the model. Consider, by analogy, the use of a hammer. A hammer is separated (autonomous) from both the wall and the nail, and its function is to *connect* the nail to the wall. In this sense the models mediate – and here kicks in the second idea, that of mediating instruments – between the two sides: reality on the one hand and theory on the other hand. The analogy of the hammer is, however, insufficient to understand the use and function of models. Models are useful, in fact, also because of their ability to represent something, which allows us to use them as epistemic tools (see Knuuttila discussed earlier in *Epistemic Objects*). Yet, while the hammer only allows us to connect the wall and the nail, a model also allows us to *learn* about the two sides that it connects. An interesting aspect of this view is that we do not learn from the model

just by looking at it, but by building it and manipulating it, and that is why they are *tools* (or, as Knuuttila would say, epistemic objects).

It is worth emphasizing that conceiving of models as (some sort of) instruments does not commit us to an instrumentalist position about models. Typically, instrumentalist accounts of models are accompanied by antirealism. Models do not give epistemic access to an objective reality, which is independent of the epistemic agent. But Morgan and Morrison in no way deny that there is a reality out there to be discovered and studied. They emphasize, instead, the instrumental function of the models as they mediate this *access* to reality. Models allow us to gain the knowledge of reality, and this also in virtue of their representative function. Similar positions have also been developed by *Hesse* [44.48] and *Cartwright* [44.49].

### 44.3.2 Isolations

In his work, *Mäki* [44.50, 51] is interested in the modeling process in economics, with particular reference to the broader problem of realism and antirealism. Mäki makes the point that the entities studied in economics are not really independent, in the same sense as a realist in physics can think of the electron as an independent entity. Some economic entities certainly are dependent on the epistemic agent. That is to say, some objects in economics, such as the preferences of economic agents, are not directly accessible through our senses, but are instead *mind dependent*. Yet, many of the objects described and studied in economics are part of our common sense of understanding of the social world, to which belong other economic entities such as prices, wages, and taxes. Some entities, such as wages, have also the physical counterparts (assuming we grant physical reality to our bank account balance!), while others remain theoretical constructs, such as, for example, preferences or values.

Mäki is interested in the modeling process of those economic phenomena that affect these *common sense* entities. Mäki points out that since the early history of economic theory – think of John Stuart Mill, Karl Marx, Carl Menger, or Alfred Marshall – we proceeded by *abstraction* and *isolation*. On one hand, economic theory starts from premises that are incomplete, and in a sense, even false. For instance, economic rationality, which does not take into account all factors involved in the choices taken by economic agents. This incompleteness is also accompanied by a form of idealization, or *isolation*, of those factors considered instead relevant. Isolating means that, in explaining a phenomenon,

some elements are deliberately removed in order to simplify it. This process of isolation makes a complex phenomenon more tractable from a theoretical and also a practical point of view.

The purpose of these idealized – and strictly speaking *false* – assumptions is to implement some *theoretical isolations* in a controlled manner. This allows, according to Mäki, to make complex phenomena understandable and manageable. A similar account of modeling can also be found in the work of Nowak [44.52].

### 44.3.3 Maps

Giere [44.53] develops an account of the relationships between models and reality. Giere supports two related positions. One is that the results of science are *perspectival*. This means that scientific results are the product of the perspective adopted in studying a phenomenon. A relevant analogy is with color vision, which is not an objective fact, but depends both on the inputs received and on the instrumental apparatus used (including our perceptual system). More explicitly, this means that, whatever science establishes – or, whatever the *scien-* *tists* establish – vitally depends on a several aspects, from the data used to the analytical methods used (instrumental and experimental apparatus, and various types of models). How is this related to the discussion about models?

First, for Giere, models are models *of data*, not of theories. Giere emphasizes the empirical aspect of models: once data are collected, they must be modeled. Surely we need theories to develop models, but *data* are the starting point. This aspect is important because, contrary to what is sometimes said, data do not speak for themselves, not even if you torture them. The adage “If you torture the data enough, Nature will confess” is attributed to British economist Ronald Coase, Nobel Prize for Economics in 1991, who claimed the importance of studying real, rather than hypothetical markets.

## 44.4 Models and Neighboring Concepts

Modeling is at the very heart of scientific reasoning and rightly occupies an important place in the philosophy of science. But we cannot discuss models and modeling abstracting from other important issues. In this section, I will examine some selected topics that are closely related to model-based reasoning.

### 44.4.1 Simulations

In Sect. 44.1, I offered an overview of the various models used in the social sciences. That overview, however,

To use the classification of Sect. 44.1, Giere focuses on empirical models (either quantitative and qualitative), not on theoretical models. Second, models are like *maps*. Maps, to be sure, are not true or false, but useful or useless for a specific purpose. An important consequence of this position concerns the concept of *truth*, which ceases to have a *metaphysical load*, and is instead used only in a *minimal* sense. I shall get back to the issue of truth in Sect. 44.4.3.

A notable aspect of this perspectival account is that it pays careful attention to the scientific practice, or rather the practices of the scientists, that form a community. In other words, Giere discusses the meaning and use of models in very specific terms, anchored to the scientific practice and the activities of those who practice science. Thus, the epistemic activities involved in science, in which modeling is certainly central, are then *distributed* – distributed across all the individuals that belong to the scientific community (on distributed cognition, see also [44.8]).

Although Giere’s account is not specifically tailored to the social sciences, it is helpful in order to address the vexata quaestio of *objectivity*. The development and the increasing use of quantitative methods in the social sciences can be read as an attempt to give objectivity to disciplines which, historically, have been blamed for being too subjective and not rigorous enough (on this, see, for example, *Montuschi* [44.18]). However, if we embrace a perspectival approach such as Giere’s, then the role of the epistemic agent (the scientist) in model-based reasoning should not be reduced as much as possible, but should instead be studied and understood as much as possible. Some of these issues are also touched upon in the positions discussed earlier (Knuuttila, Morgan and Morrison, Mäki) and more generally in the literature on scientific understanding, whose impetus was given by the article by *de Regt* and *Dieks* [44.54] and continued in the volume edited by *de Regt* et al. [44.45].

has a striking omission: simulations. Models have been classified according to their quantitative or qualitative character, and according to whether they are experimental or observational. Yet, some models do not fit in this categorization: these are the simulations. Simulations are increasingly used in the sciences, including the social sciences, and offer insights into philosophical investigations.

The goal of a simulation is to emulate a certain system, also called a *target* system. For instance, one can simulate the flight of an airplane; the airplane is repro-



duced on a smaller scale, placed in a wind tunnel and its behavior studied in simulated weather conditions, such as a storm. In these cases, we try to reproduce the target system on a smaller scale. This type of simulations raises quite a number of questions, not least because they are based on the assumption that a reduction of scale leaves unchanged the essential characteristics of the system.

There are other types of simulations, which perhaps raise even more philosophical questions. These are *computer simulations*. These simulations mimic or attempt to reproduce the operation of a real system in silico. We can simulate biological systems, such as the cell's apoptosis, or physical systems, such as collisions between particles. In the social sciences, simulations are used to study processes like urbanization of a city, of socialization between groups, etc. Simulations are often used to model complex systems, whose behavior is not easily predictable (think of models in meteorology). The aim is to be able to reproduce the state of a system as it evolves from certain initial conditions that are set in the computer program. Simulations have an interesting, hybrid status between experiment and theory.

In the social sciences, we are often confronted with a further difficulty: available background knowledge may be quite poor. In these contexts, in fact, simulations are used precisely to *acquire* new knowledge of a phenomenon. Part of the debate is about whether, and to what extent, computer simulations accrue the explanatory and predictive power of models. This is far from obvious, as the output of the computer program *depends* on the instructions entered by the programmer. This does not mean that the simulations are not useful. On the contrary, their extensive use both in physics and biology, as well as in the social sciences, suggests that there is a conceptual and methodological potential to be explored, and to which philosophy is paying increasing attention ([44.55–59], and for the social sciences more specifically see, for example, [44.60–62]).

In sum, simulations are an important topic for a chapter on model-based reasoning for two reasons. On one hand, simulations *are* part of the methodological baggage of the social sciences; on the other hand, because of their hybrid status between theory and experiment, they are a fertile area for philosophical investigation.

#### 44.4.2 Causation and Explanation

In Sect. 44.1.2, I quickly outlined the distinction between associational and causal models. It is worth developing the issue further, as it is not just about technical and methodological aspects, but also (and especially) of philosophical and conceptual importance.

At the very basis of this distinction lies the perennial question about causal inference: how/when/under what conditions we can infer causation from correlations and probabilities? In turn, this question brings us to the following one: What is causality? Does causality have a special meaning in social contexts?

Instead of addressing the question directly, I shall get to it via a distinction that has been made in the recent debate in the philosophy of causality: production versus difference making (or, in the terminology introduced by Hall [44.63], *production and dependence*). Here, I shall apply concepts, more narrowly, to the *evidence* that supports a given causal claim, rather than to *types of causality*. Thus, for instance, establishing that smoking is a risk factor for lung cancer (actually, for most types of cancer), means establishing, in the first place, a *difference-making* relation between two variables: smoking and (deaths due to) lung cancer. A difference-making relation would state that variations in the quantity of smoked cigarettes (e.g., less than 10, between 10 and 20, more than 20) are associated with the number of deaths due to lung cancer (in a given population and in a given timeframe). It is in this sense that smoking *makes a difference to* lung cancer.

Often, however, to determine whether these difference-making relations are causal, we also need evidence of production, namely of *how* smoking causes cancer. Evidence of production includes information coming from biomedicine, for instance about the mechanisms of carcinogenesis triggered by smoking. But it also includes information about social, psychological, or behavioral mechanisms. These help understand the production of cancer in a nonreductive way (on this point, see [44.64]). An overview of the question causality and evidence is offered in [44.14], who frame the problem of evidence not only within the debate on evidence-based medicine, but also within the broader question of the methods for causal inference.

The distinction between evidence of production and of difference making, and their complementarity, is relevant to model-based reasoning in the social sciences. In fact, while the quantitative models presented in Sect. 44.1.2 generate evidence of difference making, it is less clear how they also generate evidence of production. The issue is investigated in [44.11], who examine the case of econometric models in particular.

Another important aspect concerns the explanatory power of a quantitative model. In analytical sociology and in the structural modeling tradition, some scholars emphasized the role played by mechanisms in explaining social phenomena. *Ruzzene* [44.65] clearly explains how the social sciences developed and used the concept of mechanism, building a bridge between the social science and the philosophical literature. Another inter-

esting line of research worth mentioning concerns the interpretation of structural models. To be more precise, what is at stake is the interpretation of a probabilistic structure as a mechanism (for a phenomenon) – on this point, see *Mouchart et al.* [44.12, 66], who discuss different positions found in the literature.

### 44.4.3 Truth and Validity

In the course of the chapter, I mentioned several possible purposes and uses of models, some of which have just been discussed: models may be used to explain, to establish causal relations, or to simulate a phenomenon. In this section, I address the following question: What makes a model *good* or *useful* or *real*?

To begin with, it is worth emphasizing that models, strictly speaking, are neither true nor false. This idea underlies the account of Giere, discussed in Sect. 44.3.3, as he emphasizes the *usefulness* of the models. The view is also supported by Mäki, discussed in Sect. 44.3.2, as he highlights the process of idealization and simplification of certain hypotheses and assumptions in model-based reasoning. *Basso and Marchionni* [44.5] discuss the issue of the falsity of the assumptions of model-based reasoning in economics. All these arguments are based on the analysis of scientific practice. In the following, I would like to offer different line of argument instead.

Truth and falsity apply, strictly speaking, to statements. For instance, the statement *The program Word is used to write texts with the computer* is true. Most competent speakers will also consider true counterfactual statements such as the following, although their truth conditions are notoriously more difficult to establish: *Had I heard the alarm, I would not have missed the train.* The results of models too can be formulated as well-formed statements and consequently – one might argue – they can be true or false. As we have seen in Sect. 44.2.1, this was the strategy of the neopositivists at the beginning of the last century in order to clarify the notions of model and theory.

How do we *build* a scientific statement? Suppose that we study the effects of smoking cessation on mortality rates due to lung cancer. Suppose that the analysis of the data allows us to establish that the mortality in the target population (that quit smoking, or smoked less) actually decreases. We will be inclined to express the results of such a study as follows *Quitting smoking reduces the chance of developing cancer by 60%* (the numbers are clearly imagined). Note that this is what we routinely find in popular science articles, TV programs, etc. Is it true that eating broccoli and cauliflower prevents cancer? Is it true that banner ads influence our purchasing decisions? These questions can be answered

with simple, well-formed statements summarizing the results of scientific studies. However, I would like to suggest that the crucial question is *not* whether these statements are true or false, but rather whether the entire model behind it is *valid* or not. The suggestion is to *freeze* the question about truth until we clarify how we build such statements, and then get back to it. The intention is not to create a conceptual opposition between truth and validity. On the contrary, the proposal is to sketch a possible path of research. So the first step is to shift the discussion toward a different notion: validity.

In the social sciences, *Cook and Campbell* [44.67] laid down the foundation for a systematic discussion of the concept of validity. These scholars distinguished two types of validity within quantitative modeling: internal and external. *Cook and Campbell* [44.67] actually distinguish four types of validity: internal, external, statistical, and construct. For the sake of simplicity, I present only internal and external validity, which are the most discussed ones in the literature. Internal validity refers to the possibility that the relationship between two variables, within a given model, is causal or, conversely, to the possibility that, given the lack of correlation between two variables, we conclude that there is no causal relation between them. External validity concerns the possibility of generalizing a causal relationship, established within a specific model, to different populations or settings. External validity, or extrapolation, is the object of another debate, extremely relevant to the philosophy of science and to the scientific practice. Some philosophers, notably *Guala* [44.68] and *Steel* [44.69], tried to explain the conditions and procedures that allow us to generalize, or extrapolate, the results of a study to other populations. Paradigmatic cases concern the biomedical sciences, where it is far from obvious that we can draw conclusions about the aetiology of a disease or the active ingredient in a drug for human patients from animal models.

For instance, demographer Daniel Courgeau developed a multilevel model to explain the phenomenon of migration in Norway using data from censuses in 1960, 1970, and 1980. Simply put, Courgeau explained that the farmers, usually nonmigratory, underwent pressures to migrate, because the percentage of farmers within their region considerably increased. An explanation based on a multilevel model delivers results that are valid for Norway during those years. However, this is not necessarily applicable to other countries, including Norway itself, but in a different historical period (or under different socioeco-demographic conditions). The account of Cook and Campbell sparked a lively debate in the methodology of the social sciences. For instance, some think that internal validity is more important, some the opposite. Under debate is also the

issue of whether external validity is necessary, or even attainable, in the social sciences. The concept of validity is sometimes given different meanings, as it can refer to the data, the model, or the results. So far I have addressed a question about the meaning of the concept of validity. Let us see how we decide whether a model is valid or not.

In Sect. 44.1, I touched upon the issue of building and testing models. I mentioned that data can be collected, for example, with surveys or interviews. I also mentioned that we analyze data in different ways, for instance using statistical models, which in turn may be very different depending on the type of data to be analyzed and on the phenomenon to be studied. I also briefly sketched the main features of the hypothetico-deductive method that embraces different stages of modeling building and testing, from data collection to the interpretation of results. To determine whether a model is valid or not means to determine whether the entire modeling process is cogent. At each and every step of the process, we can ask if things have been done correctly, or if they could have been done differently, or better, or if there were errors that influenced later stages of the modeling procedure.

Now, the question is: Why should validity be opposed to truth? In fact, whether there is opposition or not depends on what conception of truth one adopts.

According to a well-established philosophical tradition that traces back to Aristotle, the truth of a statement is established via a correspondence between a linguistic expression and some state of affairs. According to a Tarskian analysis, *The snow is white* is a true statement if, and only if, the snow is white. That is, I open the window, I verify that snow is actually white, and declare the statement true. This interpretation of truth is plausible until we deal with simple things like white snow, the desk I work on, and any other situation within the remit of unsophisticated empirical tests. This is an over simplification, however. Think of the empirical test of the color of my desk, made by myself and by my colleague, who is color blind on some frequencies. The statement *This desk is grey* will be true for me, and most likely false for my colleague. The problem is known to philosophers of science, who, however, addressed the issue in relation to nonobservable entities (including, most famously, electrons). This simple example shows that the problem of empirical control, and consequently of the truth of a (scientific) statement, arises already at the level of the observables. For a discussion of a verificationist approach and realism Dorato [44.70].

Consider again the (fictitious) study on the effects of smoking cessation on mortality due to lung cancer. The question is whether we can determine the truth of

the statement *Quitting smoking reduces the chance of developing cancer by 60%* by some kind of correspondence. In this case, as in many others, it is not obvious to find a *fact*, *state of affairs*, or *truthmaker*. Every study has a population of reference; does it follow that there is a truthmaker for each of these? Or is there *one* truthmaker that transcends them? The majority of studies make use of variables constructed from different indicators. Does it follow that there is a unique way to isolate facts or states of affairs like *quitting smoking*? In sum, the difficulty in finding an easy and clear correspondence between our scientific statements and some facts, state of affairs, or truthmakers stems from the fact that social reality is notoriously elusive and, to some extent *constructed* (but some may argue the same holds for natural phenomena) – an idea that has been developed by Latour [44.71] or Hacking [44.72], to name only the most prominent theorizers.

One option is to reduce the metaphysical burden of the concept of truth, and to make it less rigid and more usable. We can do this by trying out the conceptual tools that come from the philosophy of information. There, *truth* is not cashed out in terms of correspondence but with respect to the information network in which a certain expression is embedded. Truth, according to this view, does not collapse into the concept of validity. The truth of a statement (e.g., a scientific statement) is established (also) on the basis of the validity of the model within which it has been formulated. Let us see how truth can be conceptualized from an informational perspective, and then go back to the relationship between truth and validity.

The concept of information has a relatively recent history. In the last decades, mathematicians, engineers, and information technologists developed theories of information that revolutionized, in many ways, our information and communication technologies. Philosophers also deserve credit, as they initiated a new branch in philosophy that makes central the notion of information and also offers new methods for philosophical inquiry. The main theorizer and proponent of the philosophy of information is Floridi [44.73].

Floridi offers the following general definition of information (GDI) coupled with the *veridicality thesis*. Simply put, according to the veridicality thesis, truth *is* part of (semantic) information: *p* is an instantiation of information, understood as semantic content, if, and only if:

- (GDI1) *p* consists of *data*.
- (GDI2) Data in *p* are *well formed*.
- (GDI3) Well-formed data are *meaningful*.
- (GDI4) Meaningful well-formed data are *truthful*.

According to this definition, the four basic elements of semantic information are:

- (i) Data
- (ii) The *structure* of data
- (iii) Their *meaningfulness*
- (iv) Their truthfulness.

For Floridi, structuring data is not confined to giving a set-theoretic structure to well-formed formulas, but includes any *rule* for governing a given system, code, or language being analyzed. Thus, a probabilistic structure of the type discussed in Sects. 44.1.2 and families of probability distributions counts as structuring the data. *Meaningful* refers to any way of complying with the semantics of a given system, but not confined to language. Thus, nonsense correlations, such as the correlation between the rise of bread prices in England and sea level in Venice, would be *meaningless* under this account. The next step is to connect meaning to truth. In this context, *true* does not have a correspondentist meaning. Floridi defends a *correctness* theory of truth that provides an account of how well formed and meaningful data become truthful. Simply put, a truthful expression is one that is *correct* within the modeled system. There is not a truthmaker making a statement true, in a correspondentist sense. There is a network of information that the agent (the competent speaker, the scientist, etc.) processes in order to establish the

truth of a statement, and this highly depends on the data considered and on the way they are (or are not) well formed.

Let us get back to scientific statements. How shall we determine if they are true? The network of information includes considerations made to decide about the validity of the model. We establish the truth of a statement within the framework of a modeled system, analyzing the various stages of model building and model testing. Therefore, in establishing the truth of a statement, we are not referring to a fact or state of affairs or truthmaker that makes it true. We refer to a whole network of information that includes the accuracy of the modeling phase, the adequacy of the empirical results with respect to background knowledge, the role of scientific evidence in its various forms (correlations, mechanisms, etc.). Then the *plot* or informational network that the epistemic agent builds – and that constitutes their knowledge – gives us the truth of a statement.

There is no doubt that these sections do not exhaust the problem. I lack space, here, to develop this approach in detail. The interested reader can consult *Floridi* [44.73] and also the introductory textbook available on the website of the Society for the Philosophy of Information, outlining the main arguments about truth, knowledge, and validity from an informational perspective.

## 44.5 Conclusion

The objective of this chapter was to provide an overview of model-based reasoning in the social sciences, to locate it within the debate on models that extends outside social research, and to highlight some of the aspects worthy of further philosophical investigation.

I opened the chapter with a description of several modeling practices in the social sciences and categorized them according to the techniques for data analysis they use (quantitative or qualitative) and according to their observational or experimental character. What emerges from this overview is a multifaceted concept. There is not one account of the concept of *model* that captures all the aspects involved in these different modeling practices. Also, the received view – according to which models represent portions of reality using set-theoretic structures – does not seem to fit the case of the social sciences. Yet, models in social research (especially quantitative models) do represent portions of reality, but by means of families of probability distributions. These are the heart of statistical models and

raise important questions about the stochastic character of phenomena (and of causality) and about the possibility of explaining phenomena mechanistically.

Models are also *objects* that we can manipulate to gain knowledge of phenomena. This view offers an account of the concept of “model” and also offers ideas to investigate relationships between models and reality. The three accounts I presented (models as mediators, maps, and isolations) do not answer *exactly* the same question. Rather, these approaches tackle *some* aspects of the relationship between models and reality highlighting different stages of the modeling process. These accounts should not be seen in opposition or competition to each other, but should instead be seen as being part of a complex mosaic portraying model-based reasoning in science.

From the discussion of the previous section, it should be clear that the interest in the concept of the model is not confined to finding the most appropriate definition. Model-based reasoning is the heart of the scientific process and this is closely linked to other no-

tions, all central in the philosophy of science. In this chapter, I could only provide a preview of debates and issues that are far more complex and fascinating. Some, such as the relationship between the model, causality and explanation, belong to the perennial questions in philosophy of science. Others, for instance, concerning the explanatory or predictive role of simulations (especially in silico), are more recent but no less controversial or relevant. Finally, I suggested that the question of truth, at the basis of science as well as philosophy, can be addressed in the light of another concept: the validity of a model.

Many issues have not been covered in this chapter, sometimes for the lack of space and sometimes because they are outside the area of competence of the author. For instance, philosophy of science has traditionally given a lot of importance to the predictive power of models and theories (especially in physics). In the social sciences, the problem of prediction has rather different contours, however. I will just mention two points, hoping to stimulate the reader's interest. A demographic projection is not meant to test the predictive power of a theory, that is, to predict a *new* observation. A demographic projection aims instead to anticipate the structure of the society in 20 or 30 years, in order to design adequate socioeconomic or public health policies. Another aspect deserves attention. Typically, in the social sciences, the best predictive models are the ones with less explanatory power. That is to say,

simpler statistical models, with fewer variables, deliver more reliable projections than more complex, structural models with many explanatory variables. We are then confronted with an interesting asymmetry between prediction and explanation which, to my knowledge, has not been sufficiently investigated.

To conclude, a philosophy of modeling – covering both a discussion of the concept of the model and the various epistemological and methodological aspects of model-based reasonings – must seek synergies with the scientific practice, on the one hand, and with other branches of the philosophy of science, on the other hand. In fact, given the hyperspecialization of both science and philosophy, we need to engage in a dialogue, in order to formulate relevant questions and useful answers. At the same time, and again because of this hyperspecialization, philosophy of science must try to build an integrated view of the scientific practice (current practices, or found in the history of science) where various concepts find their place, like in a mosaic.

**Acknowledgments.** I wish to thank the editors, Mauro Dorato and Matteo Morganti, for the opportunity to contribute to this volume. I owe much of what I know and understand about model-based reasoning “in practice” to social scientists Michel Mouchart and Guillaume Wunsch. Finally, I wish to thank my friend and colleague Phyllis Illari – for her insightful comments and most enjoyable conversations.

## References

- 44.1 R. Ankeny, H. Chang, M. Boumans, M. Boon: Introduction: philosophy of science in practice, *Eur. J. Philos. Sci.* **1**(3), 303–307 (2011)
- 44.2 T. Arabatzisa, D. Howard: Introduction: Integrated history and philosophy of science in practice, *Stud. Hist. Philos. Sci. Part A* **50**, 1–3 (2015), doi:[10.1016/j.shpsa.2014.10.002](https://doi.org/10.1016/j.shpsa.2014.10.002)
- 44.3 M. Morganti: *Combining Science and Metaphysics. Contemporary Physics, Conceptual Revision and Common Sense* (Palgrave, New York 2013)
- 44.4 E. Montuschi: *The Objects of Social Science* (Continuum, London 2003)
- 44.5 A. Basso, C. Marchionni: I modelli in economia, *APhEx* **11** (2015)
- 44.6 F. Russo: Modelli nelle scienze sociali, *APhEx* **12** (2015)
- 44.7 M. Boumans: *Science Outside the Laboratory Science outside the laboratory. Measurement in field science and economics* (Oxford Univ. Press, Oxford 2015)
- 44.8 S. Leonelli: On the locality of data and claims about phenomena, *Philos. Sci.* **76**, 737–749 (2009)
- 44.9 S. Leonelli: Data interpretation in the digital age, *Perspect. Sci.* **22**, 397–417 (2014)
- 44.10 F. Russo: *Causality and Causal Modelling in the Social Sciences. Measuring Variations*, Methodos Series, Vol. 5 (Springer, New York 2009)
- 44.11 A. Moneta, F. Russo: Causal models and evidential pluralism in econometrics, *J. Econ. Methodol.* **21**(1), 54–76 (2014)
- 44.12 M. Mouchart, F. Russo: Causal explanation: Recursive decompositions and mechanisms. In: *Causality in the Sciences*, ed. by P.M. Illari, F. Russo, J. Williamson (Oxford Univ. Press, Oxford 2011) pp. 317–337
- 44.13 F. Russo: What invariance is and how to test for it, *Int. Stud. Philos. Sci.* **28**(2), 157–183 (2014)
- 44.14 P. Illari, F. Russo: *Causality: Philosophical Theory Meets Scientific Practice* (Oxford Univ. Press, Oxford 2014)
- 44.15 R. van Ginkel: The repatriation of anthropology: Some observations on endo-ethnography, *Anthropol. Medicine* **5**(3), 251–267 (1998)
- 44.16 T.H. Eriksen, F.S. Nielsen: *A History of Anthropology*, 2nd edn. (Pluto, London 2013)

- 44.17 M. Cardano: *Ethnography and Reflexivity. Notes on the Construction of Objectivity in Ethnographic Research*, Tech. Rep. 1, NetPaper del Dipartimento di Scienze Sociali, Univ. Turin (2009)
- 44.18 E. Montuschi: *Oggettività e Scienze Umane* (Carocci Editore, Rome 2006)
- 44.19 A. Oakly: *Experiments in Knowing: Gender and Method in the Social Sciences* (Polity, Cambridge 2000)
- 44.20 F. Guala: Models, simulations, and experiments. In: *Model-Based Reasoning*, ed. by L. Magnani, N. Nersessian (Springer, New York 2002) pp. 59–74
- 44.21 U. Mäki: Models are experiments, experiments are models, *J. Econ. Methodol.* **12**(2), 303–315 (2005)
- 44.22 M. Morgan: Experiments versus models: New phenomena, inference and surprise, *J. Econ. Methodol.* **12**(2), 317–329 (2005)
- 44.23 D. McArthur: Good ethics can sometimes mean better science: Research ethics and the Milgram experiments, *Sci. Eng. ethics* **15**(1), 69–79 (2009)
- 44.24 M. Morgan: Nature's experiments and natural experiments in the social sciences, *Philos. Soc. Sci.* **43**(3), 341–357 (2013)
- 44.25 D.J. Simons, C.F. Chabris: Gorillas in our midst: Sustained inattentive blindness for dynamic events, *Perception* **28**, 1059–1074 (1999)
- 44.26 D.G. Mook: In defence of external validity, *Am. Psychol.* **38**, 379–387 (1983)
- 44.27 J.W. Lucas: Theory-testing, generalization, and the problem of external validity, *Sociol. Theory* **21**(3), 236–253 (2003)
- 44.28 P. Suppes: A comparison of the meaning and uses of models in mathematics and the empirical sciences. In: *Studies in the methodology and foundations of science. Selected papers from 1951 to 1969*, (Reidel, Dordrecht 1960/1969)
- 44.29 B. van Fraassen: Structure and perspective: Philosophical perplexity and paradox. In: *Logic and Scientific Methods*, ed. by M.L. Dalla Chiara (Kluwer, Dordrecht 1997) pp. 511–530
- 44.30 S. French, J. Ladyman: Reinflating the semantic approach, *Int. Stud. Philos. Sci.* **13**(2), 103–121 (1999)
- 44.31 G. Boniolo: *On Scientific Representations On scientific representation. From Kant to a New Philosophy of Science* (Palgrave, New York 2007)
- 44.32 C. Pincock: *Mathematics and Scientific Representation* (Oxford Univ. Press, Oxford 2012)
- 44.33 D. Molinini: La spiegazione matematica, *APHEx* **7** (2013)
- 44.34 G. De Santis, Benessere: <http://www.neodemos.info/benessere/>
- 44.35 D. Fennell: The error term and its interpretation in structural models in econometrics. In: *Causality in the Sciences*, ed. by P.M. Illari, F. Russo, J. Williamson (Oxford Univ. Press, Oxford 2011) pp. 361–378
- 44.36 G. Wunsch: God has chosen to give the easy case to the physicists. In: *Evolution or Revolution in European Population. European Population Conference*, (Franco Angeli, Milan 1995) pp. 201–224
- 44.37 R. Frigg: Models and fiction, *Synthese* **172**, 251–268 (2010)
- 44.38 H.D. Young, R. Freedman: *University Physics. With Modern Physics*, 10th edn. (Addison-Wesley, San Francisco, Reading 2000)
- 44.39 R. Frigg, S. Hartmann: Models in science. In: *The Stanford Encyclopedia of Philosophy*, Winter 2016 edn., ed. by E.N. Zalta, <https://plato.stanford.edu/archives/win2016/entries/models-science/> (2012)
- 44.40 T. Knuuttila, A. Voutilainen: A parser as an epistemic artefact: A material view on models, *Philos. Sci.* **70**, 1484–1495 (2003)
- 44.41 T. Knuuttila: Models, representations, and mediation, *Philos. Sci.* **72**, 1260–1271 (2005)
- 44.42 T. Knuuttila, M. Merz: Understanding by modelling: An objectual approach. In: *Scientific Understanding. Philosophical Perspectives*, ed. by H. de Regt, S. Leonelli, K. Eigner (Univ. Pittsburgh Press, Pittsburgh 2009) pp. 146–168
- 44.43 I. Hacking: *Representing and Intervening: Introductory Topics in the Philosophy of the Natural Sciences* (Cambridge Univ. Press, Cambridge 1983)
- 44.44 L. Daston: *Biographies of Scientific Objects* (Univ. Chicago Press, Chicago 2007)
- 44.45 H. de Regt, S. Leonelli, K. Eigner (Eds.): *Scientific Understanding. Philosophical Perspectives* (Univ. Pittsburgh Press, Pittsburgh 2009)
- 44.46 M. Morgan, M. Morrison: Models as mediating instruments. In: *Models as Mediators. Perspectives on Natural and Social Science*, ed. by M. Morgan, M. Morrison (Cambridge Univ. Press, Cambridge 1999)
- 44.47 H.-K. Chao: *Representation and Structure in Economics. The Methodology of Econometric Models of the Consumption Function* (Routledge, London 2009)
- 44.48 M. Hesse: *Models and Analogies in Science* (Cambridge Univ. Press, Cambridge 1966)
- 44.49 N. Cartwright: *How the Laws of Physics Lie* (Clarendon, Oxford 1983)
- 44.50 U. Mäki: On the method of isolation in economics. In: *Idealization IV: Intelligibility in Science*, Poznan Studies in the Philosophy of the Sciences and the Humanities, Vol. 26, ed. by C. Dilworth (Rodopi, Amsterdam 1992) pp. 319–354
- 44.51 U. Mäki: Realism and antirealism about economics. In: *Philosophy of Economics*, ed. by U. Mäki (North-Holland, Amsterdam 2012) pp. 3–24
- 44.52 L. Nowak: *The Structure of Idealizations. Towards a Systematic Interpretation of the Marxian Idea of Science* (Springer, Dordrecht 1980)
- 44.53 R. Giere: *Scientific Perspectivism* (Univ. Chicago Press, Chicago 2006)
- 44.54 H. de Regt, D. Dieks: A contextual approach to scientific understanding, *Synthese* **144**, 137–170 (2005)
- 44.55 F. Rohrlich: Computer Simulation in the physical sciences, PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association 1990 (1991) pp. 510–518
- 44.56 W.J. Kauffmann, L.L. Smarr: *Supercomputing and the Transformation of Science* (Freeman, Oxford 1993)
- 44.57 S. Schweber, M. Wächter: Complex systems, modelling and simulations, *Stud. Hist. Philos. Mod.*

- Phys. **31**, 503–609 (2000)
- 44.58 M. Winsberg: *Science at the Age of Simulation* (Univ. Chicago Press, Chicago 2010)
- 44.59 F. Varenne, M. Silberstein (Eds.): *Modéliser et Simuler. Epistémologies et Pratiques de la Modélisation et de la Simulation* (Matériologiques, Paris 2013)
- 44.60 N. Gilbert, K.G. Troitzsch: *Simulation for the Social Scientists* (Open Univ. Press, Maidenhead 2005)
- 44.61 N. Gilbert, P. Terna: How to build and use agent-based models in social science, *Mind Soc.* **1**(1), 57–72 (2000)
- 44.62 F. Varenne: Les simulations computationnelles dans les sciences sociales, *Nouvelles Perspect. en Sci. Sociales* **5**(2), 17–49 (2010)
- 44.63 N. Hall: Two concepts of causation. In: *Causation and Counterfactuals*, ed. by L.A. Paul, E.J. Hall, J. Collins (MIT Press, Cambridge 2004) pp. 225–276
- 44.64 M.P. Kelly, R.S. Kelly, F. Russo: The integration of social, behavioural, and biological mechanisms in models of pathogenesis, *Perspect. Biol. Medicine* **57**(3), 308–328 (2014)
- 44.65 A. Ruzzene: Meccanismi sociali nelle scienze sociali, *APhEx* **5** (2012)
- 44.66 M. Mouchart, F. Russo, G. Wunsch: Inferring causal relations by modelling structures, *Statistica* **70**(4), 411–432 (2010)
- 44.67 T.D. Cook, D.T. Campbell: *Quasi-Experimentation. Design and Analysis Issues for Field Settings* (Rand MacNally, Paris 1979)
- 44.68 F. Guala: *The Methodology of Experimental Economics* (Cambridge Univ. Press, Cambridge 2005)
- 44.69 D. Steel: *Across the Boundaries. Extrapolation in Biology and Social Science* (Oxford Univ. Press, Oxford 2008)
- 44.70 M. Dorato: *Che Cosa c'Entra l'Anima con gli Atomi?* (Laterza, Rome 2007)
- 44.71 B. Latour: *La Science en Action. Introduction à la Sociologie des Sciences* (Découverte, Paris 1987)
- 44.72 I. Hacking: *The Social Construction of what?* (Harvard Univ. Press, Cambridge 1999)
- 44.73 L. Floridi: *The Philosophy of Information* (Oxford Univ. Press, Oxford 2011)

---

# Models in Part I

## Part I Models in Engineering, Architecture, and Economical and Human Sciences

Ed. by Cameron Shelley

**45 Models in Architectural Design**

Pieter Pauwels, Ghent, Belgium

**46 Representational and Experimental  
Modeling in Archaeology**

Alison Wylie, Seattle, USA

**47 Models and Ideology in Design**

Cameron Shelley, Waterloo, Canada

**48 Restructuring Incomplete Models in  
Innovators Marketplace on Data Jackets**

Yukio Ohsawa, Bunkyo-ku, Tokyo, Japan  
Teruaki Hayashi, Bunkyo-ku, Tokyo,  
Japan

Hiroyuki Kido, Bunkyo-ku, Tokyo, Japan

**49 Models in Pedagogy and Education**

Flavia Santoianni, Naples, Italy

**50 Model-Based Reasoning  
in Crime Prevention**

Charlotte Gerritsen, Amsterdam,  
Netherlands

Tibor Bosse, Amsterdam, The Netherlands

**51 Modeling in the Macroeconomics  
of Financial Markets**

Giovanna Magnani, Pavia, Italy

**52 Application of Models  
from Social Science to Social Policy**

Eleonora Montuschi, Venice, Italy

**53 Models and Moral Deliberation**

Cameron Shelley, Waterloo, Canada



The chapters contained in this part provide an overview of model-based reasoning in a variety of humanistic disciplines. Some of these disciplines relate to the material practices of humanity, including architecture, archaeology, design, and technological innovation. Others relate to humanity through its policy practices, such as pedagogy, crime, economics, the social sciences, and moral reasoning. Of course, the disciplines discussed here are diverse, and have specialized histories and concerns. At the same time, each one is devoted to what might be termed *inhabitation*, that is, ways that people have of living in and adapting the world around them.

For present purposes, inhabitation refers to how people interact with their material environment, especially for the purpose of adapting it to their wants and needs. It also refers to how people organize and regulate their activities, adapting their actions in view of their significance for others. The demands of inhabitation are a crucial part of human nature and history.

It is striking how significant model-based reasoning is to studies of inhabitation. In reading the chapters of this volume, readers will be impressed with the diversity of models employed and the variety of insights that can be gained through them. It becomes quite apparent that inhabitation itself is crucially dependent on different kinds of model-based reasoning. The same can be said for studies of inhabitation in divergent disciplines. So, it is fitting to have an overview of human sciences with model-based reasoning at its center.

**Chapter 45** provides an overview of modeling as it figures in the practice of architecture. This chapter well illustrates the, not always appreciated, fact that what counts as a model, and as model-based reasoning, depends upon the tools at hand. Traditionally, architects employed sketches of plans and elevations, as well as scale models. Recently, the tool kit has expanded to encompass an array of computer-based aids. Modern architecture, then, requires command of an expanding set of modeling tools and a facility for integrating them in the minds of the architects. The chapter shows how such integration occurs through a recurrent process of abductive, deductive, and inductive reasoning.

**Chapter 46** presents a taxonomy of model-based reasoning as used in archaeological interpretation. Models have long been employed to characterize and understand the archaeological record, either through abstraction from archaeological traces or comparison across cultures. A key issue of debate has been how well models can support substantial, scientific conclusions rather than merely whimsical speculations. The chapter illustrates challenges posed by models in archaeology

and how practitioners have employed models with satisfactory results.

**Chapter 47** illustrates how ideologies affect the use of models by the designers of material technology. Besides aiming to meet technological requirements, designers also use models to promote social ideals. Those ideals could include a religion, such as Catholicism, social arrangements such as universalism, lifestyles such as consumerism, or disciplinary traits such as reductionism. The chapter indicates how model-based reasoning can be an expression of the reasoner as well as a response to external requirements.

**Chapter 48** illustrates how models may constitute and facilitate technological innovation. Traditionally, innovation is viewed as the domain of producers of technology. However, consumers also participate in innovation through ways they find to adapt and find new value in products or services. The Innovation Marketplace shows how this process of invention relies on models and model-based reasoning, and has also been adopted as the standard method for data exchange by government and industries in Japan.

**Chapter 49** surveys the sciences of education from a variety of perspectives. In the history of the subject, models of pedagogy have been constructed on psychological, philosophical, institutional, and ideological premises. In all cases, accounts of pedagogy focus on the central problem of teaching and learning for which models are constantly being proposed, applied, and assessed. Continuing focus on this process shows how important models and modeling are to it.

**Chapter 50** shows how models may figure in the study of crime and law enforcement. Computational modeling has been employed chiefly to clarify theories of crime. As the ambient intelligence model suggests, computational models may also be employed to guide responses to crime by law enforcement through analysis based on the concept of crime displacement. Thus, model-based reasoning is crucial to both understanding crime and reacting to it.

**Chapter 51** reviews and discusses models of the behavior of financial markets, their response to risk and uncertainty, and the resulting tendency toward economic crises. As the crisis of 2007 illustrates, modern capitalism is prone to sudden periods of instability. Famously, this same crisis brought about a resurgence of attention to Keynesian ideas about finances and financial policy. This crisis and Keynesian models for its explanation are the point of departure for this chapter, which explores how uncertainty conditions the sometimes vertiginous behavior of financial markets.

Chapter 52 considers the role of causal models in social science research and their use in social policy construction. Although models are used throughout the social sciences, their ability to account rigorously for social phenomena and to motivate effective interventions is disputed. This chapter explores the nature of causal models and the qualities they have that support rigorous and applicable social science research.

Chapter 53 discusses various ways in which models figure in people's thinking as they solve everyday moral problems. Not surprisingly, models are central to this kind of moral deliberation. However, accounts of this phenomenon vary widely due to the different accounts of models adopted. This chapter surveys the main accounts of moral deliberation with their varying representations of models. The result is an improved understanding of the crucial work models do as people make moral decisions.

# Models in A

## 45. Models in Architectural Design

Pieter Pauwels

At one time, architects and construction specialists used to rely mainly on sketches and physical models as representations of their own cognitive design models. Today, they rely increasingly on computer models including parametric models, generative models, as-built models, building information models (BIM), and so forth. Of course, processes of abstraction and the actual architectural model-based reasoning itself remain in the mind of the practitioner who is in control of the design and construction process. However, this whole new array of alternative computer-based representation models has profoundly affected decision-making in architectural design and construction. In this chapter, a brief overview is first given of the state-of-the-art in design thinking research. Following this, an outline is given of how diverse data models, such as BIM and parametric models, are currently used in architectural design and construction. An indication is then given of how these models relate to the in-mind model-based reasoning on which architectural designers and construction experts rely in decision-making and creative thinking. This outline will not only review well-known theories of design thinking and architectural design practice, it will also integrate ongoing theoretical research about analogical reasoning and about abductive, deductive, and inductive reasoning.

<b>45.1 Architectural Design Thinking</b> .....	976
45.1.1 The Architectural Designer as a Practitioner .....	976
45.1.2 Where Are the Models in all This? .....	976
45.1.3 Abstraction, Sense-Making, and Framing into Mental Models .....	978
45.1.4 Accessing Background Knowledge Through Analogical Reasoning? .....	979
45.1.5 Abstraction from Representation Model to Mental Model .....	980
<b>45.2 BIM Models and Parametric Models</b> ....	981
45.2.1 New Technological Media in Design Thinking .....	981
45.2.2 BIM Models and Parametric Models .....	982
45.2.3 Features and Issues in the Usage of the New Modeling Applications .....	982
<b>45.3 Implementing and Using ICT for Design and Construction</b> .....	984
45.3.1 Pragmatic Usage of Semantic Modeling Applications .....	984
45.3.2 The Usage of Design Agents or Assistants .....	985
<b>References</b> .....	987

The word *model* is ubiquitous in the current practice of architectural design and construction. Whereas architects and construction specialists used to rely mainly on sketches and physical models as representations of their own cognitive design models, now they rely more and more on computer models (or computer representations of their cognitive design models). Parametric models, generative models, as-built models, BIM, and so forth,

are used day in and day out by any architectural design and construction practitioner. Although processes of abstraction and the actual architectural model-based reasoning itself still occur in the mind of the practitioner, who is in control of the design and construction process, of course, this whole new array of alternative computer-based representation models has its impact on decision-making in architectural design and construction.

## 45.1 Architectural Design Thinking

Understanding how designers think has been the goal of many research initiatives during previous decades. Several relevant overviews are available that describe the evolution of these research initiatives and their outcomes [45.1–3], therefore we will not elaborate here in extensive detail. With the emerging interpretation in the 1970s of the design process as a process in which *wicked problems* [45.4, 5] or *ill-structured problems* [45.6] are to be *re-solved*, over and over again, design is now more and more considered as a practice or a discipline in its own right, rather than a science that can be addressed using a rigid methodological approach (see for example [45.7, p. 11]).

### 45.1.1 The Architectural Designer as a Practitioner

The basis of this interpretation relies heavily on the theories by *Cross* [45.8], *Lawson* [45.9], *Schön* [45.10], and *Simon* [45.11]. These theories typically acknowledge the complexity of the design process and the role of design thinking within this process. An architectural design situation is not necessarily considered as a design *problem* that is defined by a well-structured set of constraints, and in which a number of adjustable parameters is available. Instead, a design situation in these theories is typically understood as a snap-shot, in terms of time, in the overall design process, in which a limited number of constraints and parameters are taken into account and adjusted by a designer, in order to *satisfice* the design situation, as interpreted at that moment, into an alternative and new design situation. The term *satisficing* refers to the attitude of architectural designers to *sufficiently*, instead of *entirely*, satisfy constraints (see also [45.6] and [45.12, p. 224]). In terms of optimization approaches to design, it is typically indicated that designers look for suboptimal (but satisfactory) solutions rather than for *the* most optimal solutions.

In the theories following the above understanding, a key role is typically taken by the designer as a decision-maker. Designers are considered to be reflective practitioners [45.10]. *Schön* hereby refers to architectural designers, baseball pitchers, and musicians as example practitioners [45.10, pp. 54–55]. These practitioners continuously decide which constraints they wish or do not wish to adhere to, and which parameters they wish to use in what way. In contrast to the earlier belief of designers having a *problem-focused* strategy, they are now believed to have a more *solution-focused* or *goal-oriented* strategy instead [45.11, 12]. They proceed forward through the design process, continuously facing new design situations and addressing them as

they see fit in order to obtain the goal they have in mind at that specific moment in time. After addressing these design situations, the goal and the architectural knowledge (that were used to rely on) are typically adjusted based on the “back-talk of the design situation” or “situational feedback” [45.10]. Continuously taking action on design situations results in a co-evolution of problem space and solution space (see Fig. 45.1 and [45.13, 14]). In other words, architectural designers learn while doing, not only about the design situation at hand, but also about architecture and design in general.

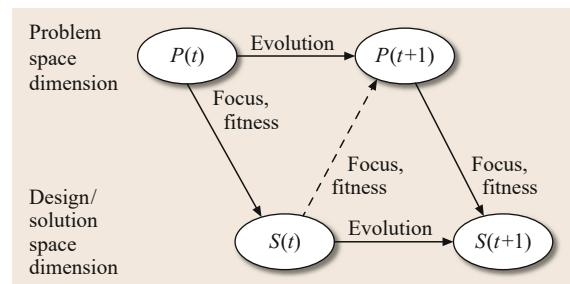
Two key features of the prevalent current understanding of the design process can be distinguished in the above paragraphs:

1. An intensive interaction exists between designer and design context, thereby resulting in a stepwise proceeding through the design process.
2. Design thinking has an important reflective, *learning-while-doing* character [45.10], enabling designers to learn from experience.

We provide a simple schematic image of this interpretation of the (architectural) design process in Fig. 45.2. It indicates how a designer forms a mental model from an observation of the external world or design situation, and uses this mental model to devise an appropriate action for altering the design situation into a new one that can ideally be considered more optimal than the previous one.

### 45.1.2 Where Are the Models in all This?

The title of this chapter suggests that models are key to the above outlined features of the interaction between designers and design situations. Although the above paragraphs did not contain the word *model* too abundantly, it is present in all stages of the design process, in each step taken by designers in the direction of a fi-

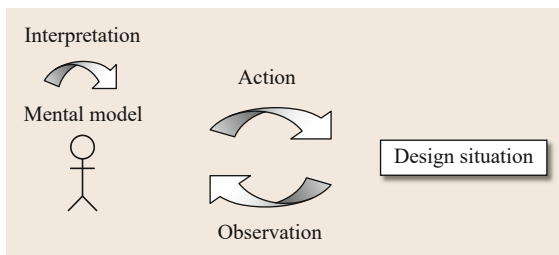


**Fig. 45.1** Maher and Poons problem-design exploration model

nal (satisficing) design situation (not solution). In each interaction, namely, architectural designers rely on their background knowledge in making the appropriate decisions. This background memory is central to the above design process. After each decision, feedback or situational backtalk is returned to the designers by the design situations with which they interacted [45.10] (see also the oil painting example given by *Simon* [45.11, p. 163]). As this situational backtalk is interpreted by the designers, it also reshapes the background knowledge of these designers.

In learning-while-doing, designers build up knowledge in direct reference to concrete experiences. This knowledge might be related to “a designerly way of knowing” [45.8, 15], which was originally put forward by *Archer* in 1979 [45.16, p. 348]. On the basis of this kind of knowledge, designers make design decisions in newly encountered design contexts. Through their ongoing interaction with new design contexts, designers continuously modify or adjust their designerly way of knowing. Obviously, these adjustments have a significant effect on future design decisions.

It is important here to consider the stepwise evolution of the design process. As each step in the design process (or each consecutive design situation) can be considered a snapshot, the evolutions in design knowledge of the designer can also be considered in a stepwise manner. Each step in the evolution of someone’s design knowledge can hereby be considered a design model. This idea follows the theory by *Schön* about the architectural designer as a reflective practitioner, continuously interacting with the surrounding context and being affected by the backtalk of that surrounding context [45.10]. As explained here, Schön also indicates that design thinking depends on the repertoire or knowledge and experience of the designer. So, the context and background of designers play a key role in all steps of the decision-making processes of those designers. This idea also explains the notion of co-evolution in design problem and design solution [45.13, 14], assuming that the design *problem* can be considered the same as the

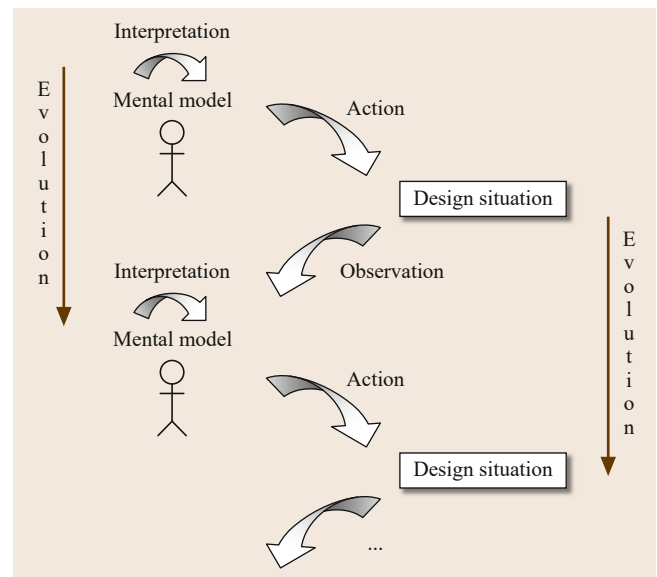


**Fig. 45.2** Schematic outline of the steps (observation – interpretation – action) that are taken by designers during the interaction with an external design situation

current design situation, and that the design *solution* can be considered the same as the internal interpretation or model of the design situation in the designer’s mind (Fig. 45.3).

The changing background knowledge of the designer has been discussed at length by *Lawson* [45.9, p. 159]. He uses the term *guiding principles*. These guiding principles can be understood as the personal background knowledge or the knowledge by experience of a designer. They consist of familiar design patterns that a designer relies upon throughout the design thinking process. A designer thus never starts a design from an empty page, never from scratch or a blank mind. Instead, a designer always relies on a lifetime of knowledge built up by experiences. It is documented in [45.9, p. 179] how these guiding principles, in combination with a mental model of the situation at hand, essentially guide practitioners (including designers) through their thinking process. They play an important role not only in framing the design situation, but also in generating solutions for a problem, devising experiments, and in learning from experiences.

According to *Lawson* [45.9, p. 159], these guiding principles include not just objective, factual information, but include much more information, involving, for instance, motivations, beliefs, values, and attitudes. Guiding principles may be very intense and clearly structured or, on the other hand, vague and unclear, but they always influence design decisions one way or another. These characteristics of guiding principles can be related to the tacit dimension suggested by



**Fig. 45.3** Schematic outline of the co-evolution of the mental model of a design situation and the external design situation

Polanyi [45.17, 18], who states that some knowledge cannot be formalized and is essentially experience-based, vague, and thus tacit. In some research initiatives, guiding principles are almost considered part of a *personal religion*, thereby implicitly redefining design as “a very complicated act of faith” [45.19, p. 3]. This refers to the sometimes profound intensity of the designer’s belief in personal guiding principles, making it *morally right* to follow them. It is similarly indicated by Ward [45.20] how imagination relies almost entirely on known concepts, and, although modifications are made, they are typically only constituted by different combinations of known elements. It is hard to entirely step outside one’s own categories and beliefs, also in imagining [45.20].

It is very unclear in what form guiding principles are stored in the mind of a designer. What is clear though, is that this background information serves as a kind of repertoire of reference models for the designer to continuously and actively reorganize and restructure new design situations in memory into new abstract mental models or understandings of those design situations. In this context, references can be made to the work on case-based reasoning (CBR) [45.21–23], in the sense that the concept of CBR captures the idea of matching new cases with previously encountered cases in order to appropriately act upon them (Fig. 45.4).

### 45.1.3 Abstraction, Sense-Making, and Framing into Mental Models

From the previous sections, we now see that there should be some mechanism or phenomenon that allows architectural designers to *link* incoming design situations to their available repertoire of reference models (their experiential background information) so that they can obtain an abstract in-mind interpretation of those design situations. Basically, this is a moment of *interpretation* or *abstraction*. It occurs when a designer is sketching and all of a sudden gains an insight in the form of recognizing a part of the sketch as something he has seen before in an alternative context. It occurs when an architectural designer is visiting architectural building sites in order to find inspiration for the issues he is struggling with in the design he is working on. It occurs when the architectural designer communicates his latest design to a client or to any related or unrelated third party and gets insight from the feedback he receives through simple conversation. While doing each of these things, the architectural designer appears to interpret or make abstraction of the incoming information, after which he relates it in his mind to what he has experienced before. Previously used names for this phenomenon are *retrieval* (in the context of

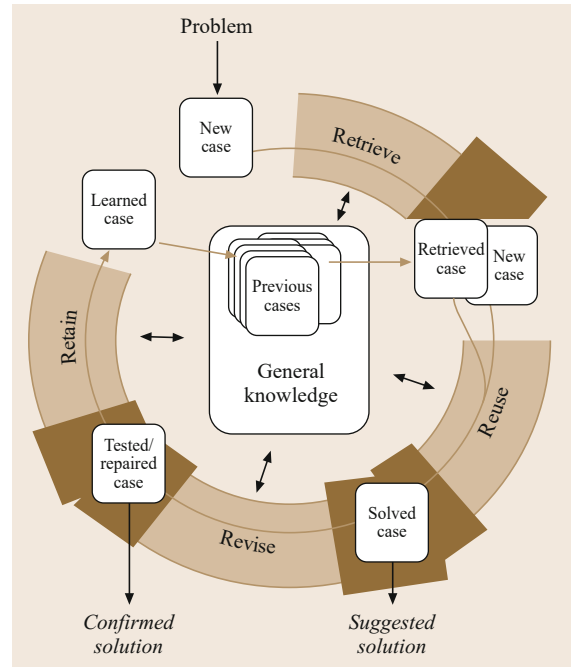


Fig. 45.4 Schematic outline of the case-based reasoning (CBR) process (after [45.21–23])

CBR – Fig. 45.4), *sense-making* and *problem framing/setting* [45.10, 24–27].

In [45.10, 26], for instance, design thinking is characterized as a specific kind of problem solving, in which the designer “must make sense of an uncertain situation that initially makes no sense”. Making sense of the situation then happens by switching back and forth between problem and solution, while continuously reframing both. Schön [45.10, pp. 39–40] refers to problem setting as:

“the process by which we define the decision to be made, the ends to be achieved, the means which may be chosen. In real-world practice, problems do not present themselves to the practitioner as givens. They must be constructed from the materials of problematic situations which are puzzling, troubling, and uncertain.”

In architecture, this often happens in the interaction between the client and the architect. This interaction usually starts with a client having an impression of what he wants to achieve, and an architect who does not have a clue of the client’s desires and needs. Often, these desires and needs are conflicting or do not seem to make sense. But, through the continuous feedback that the architect receives from the client, this design brief gets an increasingly clear structure in the designer’s mind. The architect thus sets the problem (how many floors

are requested in the building, who will be accessing and using the building, and so forth) through interaction with his surroundings. Initially, these surroundings are constituted by the feedback of the client, but later on, they will be formed by the sketch on his paper, visits to the building site, conversations with third parties, and so forth. One might say that the problem and solution are continuously reframed, resulting in a co-evolution of problem and solution. One might also say that the design situation is evolving step-by-step by the impact of an architect who decides based on his background knowledge and the context he is working in.

It is made clear by Schön [45.10] just how important the element of reframing the design situation is in his documentation of the differences between problem solving in a rational world and problem solving in the real world [45.10, p. 40]:

“When we set the problem, we select what we will treat as the *things* of the situation, we set the boundaries of our attention to it, and we impose upon it a coherence which allows us to say what is wrong and in what directions the situation needs to be changed. Problem setting is a process in which, interactively, we name the things to which we will attend and frame the context in which we will attend to them.”

In an architectural design situation, this occurs, for example, in the form of an architectural designer deciding at some point to look at the structural design in that particular design session, and leave out considerations in terms of energy or user comfort. In the following sessions, he might focus on material use, or user access, or something entirely else. But in each session, only one frame of the entire design situation is considered.

What is probably the most interesting moment in this reframing process, is the point where a solution is considered satisfactory enough and ready to be put into practice. This moment resembles the moment in which the well-known flash of insight occurs. This moment is described by [45.27] as the moment in which the two oscillating points, *problem* and *solution*, are still and close enough to be bridged by an *apposite concept* [45.27, pp. 439–440]:

“The crucial factor [...] is the bridging of these two partial models by the articulation of an apposite concept [...] which enables the models to be mapped onto each other. The *creative leap* is not so much a leap across the chasm between analysis and synthesis, as the throwing of a bridge across the chasm between problem and solution. Such an apposite *bridge* concept recognizably embodies satisfactory relationships between problem and so-

lution. It is the recognition of a satisfactory bridging concept that provides the *illumination* of the creative *flash of insight*.”

The term *apposite* makes an intended reference to the notion of appositional reasoning, which was originally coined by Bogen [45.28] and which is considered similar or the same as *abductive reasoning* by Cross [45.29].

Suppose that our architectural designer is still focusing on the structural load-bearing capacities of a building (cfr. framing of the situation). This architect might successfully relate the current design situation with a situation he encountered before and consequently decide to apply a similar structural design (for example, steel columns for load-bearing instead of concrete columns or brick walls), because the features of this structural design choice not only address issues in the overall building structure, they also appear to geniously address many other issues in terms of light penetration in the building, accessibility, fire safety, and so forth. This train of thought involves a bridge between an unsure problem and an unsure solution by an apposite concept.

#### 45.1.4 Accessing Background Knowledge Through Analogical Reasoning?

The above distinguished *apposite bridge* between current design situation and the designer’s background knowledge is often addressed and investigated as a kind of analogical reasoning or CBR (see for instance [45.30–33]). Analogical reasoning is hereby explained as the cognitive ability to think about relational patterns [45.34–37], which allows one to find a structural alignment or mapping between a base and a target pattern residing in (partially) different domains [45.34, 37–40]. During design practice, architectural designers thus continuously make alignments between the current design situation (the base pattern) and previously experienced design situations (the target pattern). Relying on such alignments, designers infer which action to take for specific design situations and hence move forward.

This understanding was formed earlier by the investigations of Douglas and Isherwood, and Cross. Douglas and Isherwood [45.41], for example, indicate that [45.41, p. viii]:

“there is a prior and pervasive kind of reasoning that scans a scene and sizes it up, packing into one instant’s survey a process of matching, classifying and comparing. [...] Metaphoric appreciation, as all the words we have used suggest, is a work of

approximate measurement, scaling and comparison between like and unlike elements in a pattern.”

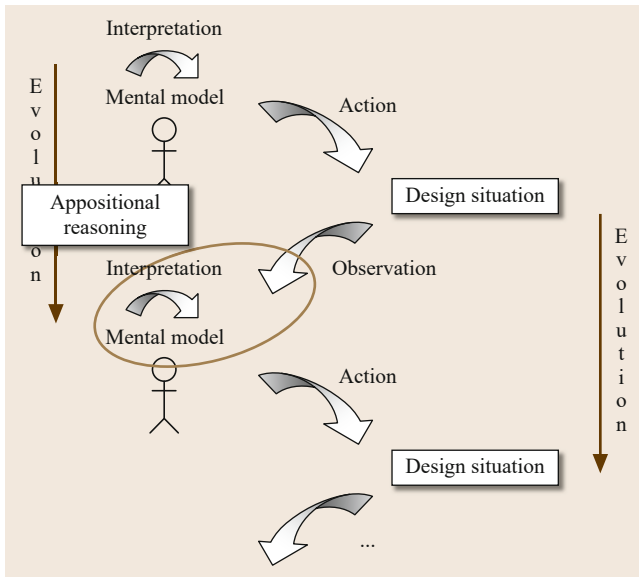
Later on, *Cross* [45.29] refers to several other research initiatives that distinguish a very similar kind of reasoning as fundamental for design thinking, thereby mentioning the terms abductive reasoning, productive reasoning and appositional reasoning as called by their respective inventors *Peirce* [45.42], *March* [45.43] and *Bogen* [45.28].

This kind of reasoning is obviously relied upon in the interpretation step which is considered in the previous sections of this chapter. This kind of reasoning is very poorly understood in general. The only thing we appear to know, is that it happens. As we are confined to behavioral studies of human design activity, and we cannot simply access the human mind during design activity, there is no real trustworthy indication of *how* it happens. When turning to the interpretation of the “work of approximate measurement” or “metaphoric appreciation” [45.41, p. viii] as a kind of analogical reasoning, we can find out that analogical reasoning often occurs between a new design-related experience (building, sketch, three-dimensional (3-D) model, conversation, and so forth) and a previous design experience as it is stored in the human mind [45.30]. But also in the very act of sketching, analogical reasoning is crucial, because it allows reinterpreting or *seeing as*, as *Goldschmidt* puts it [45.32, 33]. In *seeing as*, the designer reinterprets the sketch and, as such, adds new and origi-

nal meaning to it, thereby generating new ideas [45.32, 33]. For many student designers, who have little experience in architectural examples, *seeing as* often occurs in a more superficial way. They tend to find similarities between their sketches and other, often unrelated concepts and things based on geometrical features and shape (*Look, this looks like a ship. Maybe we can... or We are near the sea, why don't we make the building in the shape of a wave?*). Experienced architectural designers often make more abstract, meaningful and/or direct analogies, because they have a much richer set of background experiences on which they can rely.

Because analogical reasoning is guided by encountered target patterns [45.34, 37–39], the designer appears to proceed *in an unstructured and perhaps aimless way*. Therefore, the earlier mentioned definition of imagining [45.9, 44] is also closely related to analogical reasoning. A similar conclusion is given by *Boden's* research on the creative mind [45.45]. She stresses the importance of the incubation phase in creative thinking. In this phase, the conscious mind focuses on other domains, problems, or projects, thus enabling the creative mind to make diverse alternative and previously unconsidered analogies with the situation at hand [45.45, pp. 33–35].

When turning back to our initial schema of the design process (Figs. 45.2 and 45.3), we can locate the position of analogical or appositional reasoning somewhere between the design situation that is observed and the mental model resulting in the designer's mind (Fig. 45.5).



**Fig. 45.5** The approximate location of analogical or appositional reasoning in our earlier schema of the design process (Figs. 45.2 and 45.3)

### 45.1.5 Abstraction from Representation Model to Mental Model

If the bridging between the current design situation and the background knowledge of the designer occurs through analogical reasoning, the *base pattern* [45.34, 37–39] is tremendously important. Namely, this implies that what designers experience from design situations are the sole *seeds* from which they are able to make interpretations and act creatively upon.

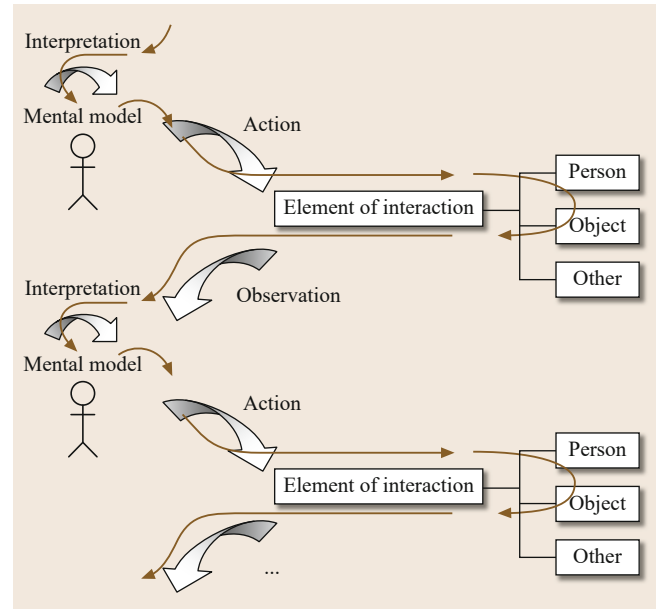
Of course, it is not realistic to assume that designers interact with design situations as a *whole*, something that might be concluded from the schemas in Figs. 45.2, 45.3 and 45.5. Instead, designers interact with a *medium* that provides a bounded interface (a frame) to the current design situation (Fig. 45.6). These media of interaction are of various kinds, including people, objects, sketches, 3-D models and so forth.

The most obvious experiences of design situations in architecture (elements of interaction in Fig. 45.6) are sketches [45.46]. As *Goldschmidt* indicates [45.32, 33],



sketches are not only visual expressions of what one wants to express, they are also elements for reinterpretation and thus for generating all kinds of new knowledge. *Cross* similarly refers to the importance of sketching as it enables a designer to explore several solutions and problems to a certain design situation at once, thereby considering several levels of detail at once [45.8, 47]. *Schön* [45.10], in turn, refers to the habit of many designers to continuously make representations of the design situation at hand in documents, plans, sketches, and so forth, thereby allowing a designer both to answer a previous design situation, and frame the design situation anew into an alternative perspective. To say it in *Schön's* words, the designer “shapes the situation in accordance with his initial appreciation of it, the situation *talks back*, and he responds to the situation’s back-talk” [45.10, p. 79].

We will not dive into all the characteristic features of sketching, but we will instead generalize among very diverse possibly available base patterns used to initiate analogical reasoning. Sketches, namely, are but one of the many possible representation media that can be used by designers to reflect on the design situation. Besides sketches, designers can use conversations with colleague architects, physical scale models, walking around on construction sites or inspiring related pieces of architecture, and so forth. Rather recently, this array of interaction media has notably enlarged through the development of all kinds of information technologies. New media are now available to the designer, among which there are parametric design models, two-dimensional (2-D) computer-aided design (CAD) models, 3-D BIM models, databases, websites on the Internet, teleconference applications, virtual game engine environments, and so forth. So, design representations can



**Fig. 45.6** Indication of the diverse elements with which a designer can interact: people, objects, other

take on many forms, including a sketch [45.32, 46, 48], a simple discussion [45.49], a CAD model, and so forth.

The main idea here is that by making alternative representations, designers aim at confirming the abstract model they have of a particular design situation, which is always to some extent unclear, wicked or unknown. By experiencing the resulting design representation, a new understanding or abstract in-mind model of the design situation thus emerges, which reframes the previous design situation and thus alters the design process.

## 45.2 BIM Models and Parametric Models

The apposite bridging, interpretation, or abductive reasoning step is a capacity that is not available in a computer. As we do not know the way in which our background information is stored in our neurological brain, we are obviously unable to replicate this. As a result, no information system exists that is able to store the target patterns that are required for analogical reasoning, let alone one that is able to match these target patterns with incoming sensory information and thus make analogies in a creative and autonomous manner, as we do as human beings. So, no information system is able to take over such a typically human capacity.

### 45.2.1 New Technological Media in Design Thinking

Nonetheless, architectural designers can still take advantage of information systems as an additional medium that allows them to make alternative representations with which they can interact in their sense-making or interpretation process. Any computer-based representation thus functions similar to how a sketch functions. Each such representation hereby represents only a limited semantic domain, only a partial reflection of the complete design situation. They are representations of the designer’s mental model, and by no means do they

come close to the original mental model which is always in the mind of the architectural designer and which is inherently ungraspable. Instead, the representations in these media are to be considered as representations that result from this mental model and that form, as such, initiators for further reflection on this mental model. In the following section, we will briefly look into the consequences in the context of BIM modeling and parametric modeling software, as reference examples.

### 45.2.2 BIM Models and Parametric Models

There have been many developments in information and communication technologies (ICT) for the domain of architecture, engineering and construction (AEC). Most ICT applications in this domain allow to build a certain representation or model of an architectural design (element). A considerable number of the developments in ICT for the AEC domain have focused on enlarging the amount of semantics that can be included within the resulting representations or models. In other words, instead of allowing designers to model a design using lines, points, boxes or surfaces, they now typically allow to model typed objects, such as walls, doors, structural columns, and so forth. Each instance of one of those object types can then automatically be represented using the properties that were predefined for these object types. These properties typically include basic features, such as height, width, and location, but they also include far more complex properties, such as relations with other objects (aggregation, decomposition, neighborhood, etc.), representational properties (texture, geometry, etc.), and so forth.

These developments have resulted in a number of modeling applications with capacities that make them stand out from the traditional CAD or computer-aided drafting applications. One can distinguish the following modeling application types.

#### Building Information Modeling (BIM) Applications

BIM applications allow to represent buildings using a hierarchical structure of typed objects, including building objects, materials, people, and so forth [45.50]. References can be made to the concept of feature-based modeling (FBM) [45.51]. The workflow in such applications results in a single 3-D BIM model, which is supposed to include all the information needed to build the building (Fig. 45.7).

#### Parametric and Generative Modeling Applications

Parametric and generative modeling applications allow to represent a design using a number of typically

geometric parameters. By moving sliders, parameter values are changed, and a design model can be regenerated from these modified parameter values. The design model is hereby formed by a network of parameter values and control functions that generate geometry using the associated parameter values (Fig. 45.8).

#### Database Applications

Many more basic applications in the architectural design and construction industry still rely on rather basic relational database systems. This includes, for instance, four-dimensional (4-D) (time scheduling) and five-dimensional (5-D) (cost scheduling) applications, facility management (FM) applications, energy performance calculation software, and so forth. Of course, such applications also use a semantic model of the architectural design, represented by tuple values in a relational database.

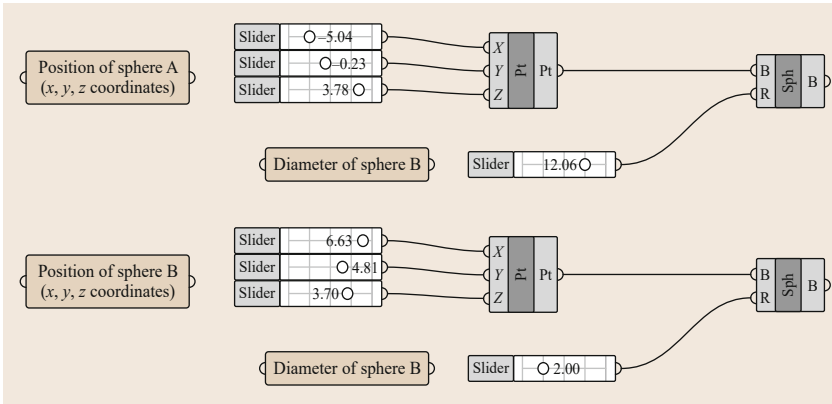
### 45.2.3 Features and Issues in the Usage of the New Modeling Applications

Obviously, for all of the three technology types mentioned above (BIM software, parametric software, database software), there are numerous interpretations and implementations. We will not dive into the details for each of these technology types in this chapter, as the focus is here on the role and function of these new interaction media within the design process. Within the scope of this chapter, it should suffice to keep in mind that each of the outlined software environments allows to build a simple or complex semantic model as a representation to interact with.

The semantic model that can be built using the outlined modeling environment typically follows the information structure that is defined by the programming code behind the corresponding modeling application.



**Fig. 45.7** Revit Architecture is one of the available BIM modeling applications that allow architectural designers to model a BIM model representation of their design



**Fig. 45.8** Rhinoceros, together with the Grasshopper plugin, is an often used environment for the parametric modeling of building geometry. This environment is typically relying on nodes and sliders to represent the semantics of the designed geometry

A Revit BIM model (Fig. 45.7) is something that cannot be captured by a parametric model in Rhinoceros and Grasshopper (Fig. 45.8), because both modeling environments deploy different programming codes and corresponding information structures. Basic 2-D CAD applications enable the user to model a design in 2-D geometric object models, using lines, points, surfaces, and so forth. Basic 3-D applications allow this in 3-D, using boxes, spheres, voids, and so forth. More advanced CAD systems typically focus on information management, and thus enable the user to model a design in more informative elements, such as walls, windows, columns, beams, and so forth.

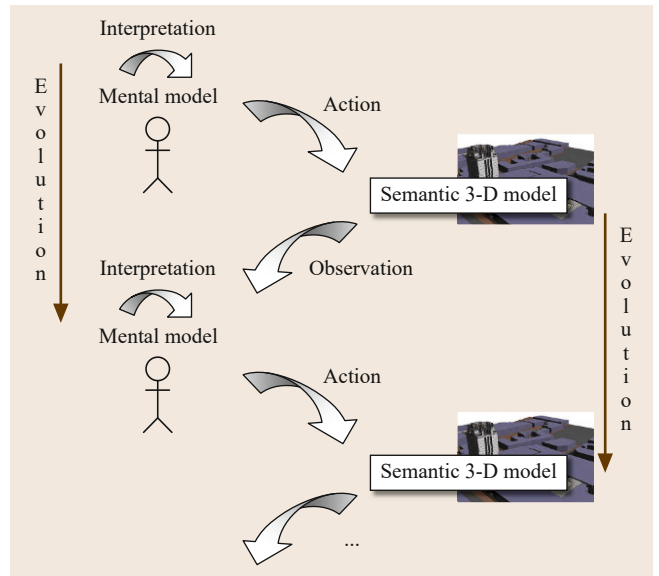
By allowing designers to model their design in a more meaningful manner (more semantic object types such as walls and doors instead of the more syntactic points and lines), designers are supposedly enabled to represent their design as a model that is much more closely related to the in-mind abstract model that they use in the design thinking process (Fig. 45.9). In concrete terms, rather than only being able to represent a design using pencil marks on a paper, semantic features in software applications allow the designers to model a semantic structure (the ontology) that reflects the in-mind structure of their designs and then use that semantic structure to represent the actual design (instantiation of the ontology). It is then easier to make decisions, as the gap between the semantic model of the design (Fig. 45.9, right) and the in-mind design model (Fig. 45.9, left) is smaller and the interpretation step that is to be made by the designer should be easier.

This functions relatively well. There are, of course, a number of issues that are commonly identified in using these modeling applications:

1. *Too much time is needed* to build the appropriate semantic structure for one's particular in-mind design model, resulting in a preference for quicker draft-

ing applications or drafting media (computer-aided drafting or sketch environments).

2. As the in-mind design model is continuously changing (cfr. co-evolution of problem and solution), one's *semantic structure is never up to date* with that in-mind design model. In other words, the semantic 3-D model (Fig. 45.9, right) is always a number of steps behind the in-mind design model (Fig. 45.9, left).
3. A need arises to share the semantic model of the design with other people, as is typically also done with sketches. However, transferring/communicating the meaning of the custom semantic structure to anyone else requires considerable effort from that other



**Fig. 45.9** The design process, as schematically represented earlier, indicating how semantic 3-D models are used by architectural designers

person as the presented semantic structure never matches with his own in-mind model. This is related to the well-known *interoperability* problem (see for reference [45.52–56]).

When considering the theories presented in the first section of this chapter, it is rather obvious and understandable that these three main difficulties emerge. If the abstract in-mind design model changes at every single snapshot of interaction with some kind of interaction medium (Fig. 45.9), of course, the representation on the medium with which is interacted is outdated at every single moment in time. Hence, it would also be a futile attempt to make a complete representation of an abstract in-mind design model in any of the available 3-D modeling applications. Note that it would likewise

be a futile attempt to make such a complete representation in one paper sketch.

Furthermore, in order to get information into another information structure (the interoperability challenge), no matter in what kind of information structure it was originally captured, one *always* requires interpretation if it is to be done properly. Thus, this requires human effort. The best option in this context of interoperability problems is to at least make flexible and intuitive tools available so that a designer can at least do the required interpretation effort in a relatively smooth and efficient manner and transform information from one semantic schema into another. Semantic web technologies might provide some of such flexible and intuitive tools, as is indicated in [45.56–59].

## 45.3 Implementing and Using ICT for Design and Construction

Considering the above writings in this chapter, modeling applications will remain to be used as alternative media for interaction by architectural designers and construction experts, no matter the amount of semantics they allow designers to represent. The semantic structures of these applications might to some extent match or resemble the in-mind model of the designer, but it will always fall short in comparison. That being considered, there is, first and foremost, a need for a flexible, intuitive, and above all, pragmatic usage of modeling applications in architectural design and construction. Many examples of such usage strategies exist in practice. Unfortunately, this pragmatic approach remains to include the effects caused by limited interoperability of information [45.53], namely an increased loss of time and an increased number of construction errors during the construction process due to the necessary remodeling of information from one environment to another.

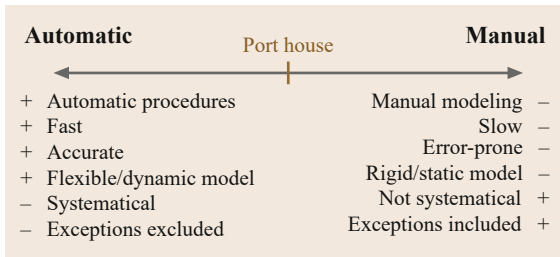
### 45.3.1 Pragmatic Usage of Semantic Modeling Applications

A first example of this pragmatic approach is documented in [45.60] and elaborates on the pragmatic usage of BIM systems in the construction of the Port House in Antwerp. Initially, an integrated BIM approach [45.50] was targeted in this project, using Revit Architecture as a central BIM environment. The BIM model would then serve as a central reference model containing and providing information for all project partners. The construction company tried to be faithful to the BIM idea, but they gradually shifted to a more pragmatic software usage approach. A BIM model was

engineered and maintained as a reference model by the construction team in charge of the whole construction process. Depending on the background and software usage approach of the project partners, the information in this BIM model was interpreted and provided to the related project partners according to the semantic structures they were using and the medium in which they were working. Communication of information thus included, for instance, Excel spreadsheets, PDF documents, and partial 3-D models in various file formats. In importing and exporting these documents to and from the modeling environment, human interpretation was required in the form of manual conversion efforts. Nevertheless, this human interpretation step produced a desirable result within a foreseeable and plannable time span.

The pragmatic software usage approach outlined here and in [45.60] requires a very good balance between automatic (technological reformatting) and manual procedures (human interpretation steps). The information structures of applications can be integrated either by implementing project-specific software components or by manual modeling. The key to a pragmatic usage of (semantic) modeling applications is finding the right balance between these automatic and manual procedures (Fig. 45.10), so that it fits the current design situation.

Similar approaches appear to be followed in other large architectural firms that concentrate on geometrically or semantically complex architectural projects. In most cases, the balance shifts towards the usage of automatic methods, as larger and more complex projects often benefit from automatic procedures in terms of



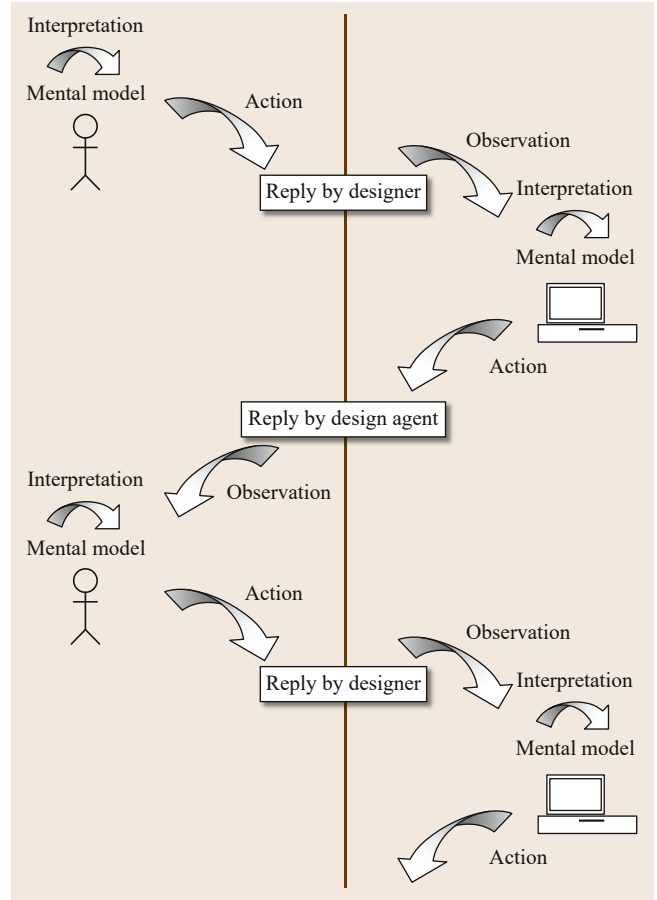
**Fig. 45.10** Indication of the balance between advantages and disadvantages of using manual and automatic procedures in information-handling in the Port House architectural design project (after [45.60])

efficiency or return on investment. A good example is the Specialist Modelling Group (SMG) in Foster + Partners, which is a group that appears to concentrate on optimizing information exchange and complex modeling for specific projects [45.61]. It appears to be confirmed in [45.60] that this not necessarily requires a standard information management approach. Pragmatically constructing a common agreement between project team members and combining manual and automatic methods with an expert group of programmers, process modelers and/or communication specialists can prove to be just as effective.

### 45.3.2 The Usage of Design Agents or Assistants

Thus, for architectural designers and construction specialists, the better option in using information technologies is to consider these technologies as yet another set of available media with which they can interact as part of their reflective practice. As with all media, there are certain rules, advantages and disadvantages that characterize each medium. One should thus carefully consider what medium to consider for what purpose. Something that all media have in common, nonetheless, is that they can all capture but a fragment of our in-mind abstract knowledge.

One might wonder whether there really is no alternative, whether information systems will really not be capable, almost as if by definition, to capture an abstract model similar to the way in which human beings do so. As indicated in the first section of this chapter, there is only one thing that stands in the way of such a development and that is the element of *interpretation*, *analogical reasoning* or *abductive reasoning* (see Fig. 45.5). This seems to be one of the main capacities that distinguishes man from computer. Key design actions require interpretation, including: the mapping of incoming semantic information to its own semantic structures, or the construction of context-specific and purposeful shape



**Fig. 45.11** Our original schema of the design process, adapted in order to communicate how the interaction between designer and computer-based agent could be taking place

grammars and creation of appropriate design decisions while relying on and continuously adapting this shape grammar, or the performance of multidimensional optimization or satisficing of design constraints. In order for a computer to perform these actions, it will first have to be able to interpret incoming information and understand it using a mechanism that involves a form of abductive, analogical or appositional reasoning.

If this is ever realized, it will result in the third of three realms of software usages in architectural design and construction support, as they were originally identified by Lawson [45.62]. We have seen the first two in the previous sections, namely, (1) the computer as a rigid problem-solving and all-knowing oracle, and (2) the computer as a draughtsman, which is simply used as yet another interaction medium while the designer remains the only decision-making and interpreting agent. In a third scenario (3), the computer is to be used as a true design agent or assistant (see also [45.57] for more in-

formation about how this relates to the earlier outlined pragmatic approach). In this scenario, designers interact with autonomously reasoning computer-based agents, as if they would interact with any other medium. Both the designer and the computer-based agent would then interpret incoming information and learn from experience. Using our schematic format of the design process again, this interaction between designer and computer-based agent would likely look as depicted in Fig. 45.11, with processes of inquiry happening in the human mind (Fig. 45.11, left) as well as in the computer agent’s information system (Fig. 45.11, right).

For this scenario to be realized, interpretation as it functions in the human mind needs to be implemented. In this regard, some effort has already been placed in the automation or implementation of abductive reasoning. It is useful to consider the number of research initiatives in the domain of abductive logic programming (ALP) [45.63]. Second, we have already looked into some pointers towards CBR in brief [45.21–23]. Unfortunately, most of these research initiatives appear to lack tangible research results. Not one of the resulting software systems appears capable of reliably simulating an interpretation step as it is produced by any human agent in an autonomous and natural manner, let alone within an architectural design context.

Some researchers have taken a different approach and have primarily considered the larger cycle that people appear to go through in any interaction with an outside world [45.64–66]. This larger cycle resembles

the original *process of inquiry* as it was outlined multiple times by Peirce [45.42]. It is hereby considered useful to not only consider abductive reasoning, but to also look for the occurrence of inductive and deductive reasoning as they were considered by Peirce [45.42]. This implementation approach makes sense as it seems to be the only valid way to let an agent learn in an autonomous manner, yet enabling it to store previous experiences so that they can be reused in interpreting future observations. Additionally, combining abductive reasoning with inductive and deductive reasoning in an iterative cycle is required if the agent’s functionality is to remain true to Peirce’s idea of a process of inquiry. When relying on this interpretation of Peirce’s process of inquiry [45.64–66], we might be able to indicate where such a cycle resides in our earlier diagrams of the design process (Fig. 45.12).

One research initiative that appeared to take on the realization of this agent approach can be found in the work of King [45.67, 68]. In this work, the goal was to build a machine that can autonomously “discover new scientific knowledge” by its capacity to “devise a hypothesis, carry out experiments, to test it and assess results” [45.68]. This endeavor to enable a machine to go through these diverse stages is very much like building a machine that is able to go through Peirce’s process of inquiry, in the sense that in the case of King’s research, the main question was also whether they were able to build a “robot scientist that can actually accomplish the entire process” [45.68]. Eventually, a machine was built that is capable of successfully constructing hypotheses, devise appropriate experiments and test them, in the domain of functional genomics, a domain in which the relations between genes and their functions are investigated. These actions are made using a core body of knowledge, resulting from a “formalization that involves over 10 000 different research units in a nested treelike structure, 10 levels deep, that relates 6.6 million biomass measurements to their logical description” [45.67].

In any case, even if this autonomous agent-based approach proves valid and a useful implementation of such an autonomous reasoning agent can be realized, it will take years before this agent takes on a form that can provide help to an architectural designer as it was theoretically outlined by Lawson [45.62]. Namely, it requires at least 25 years for brilliant human agents to learn this capacity and there is no reason to assume that computer-based agents might be able to outperform this human capacity. So, to conclude, we are best off, for now, with following the pragmatic software usage approach as it was outlined earlier in this chapter and keep relying on the architectural designers themselves as reflective practitioners and decision-makers.

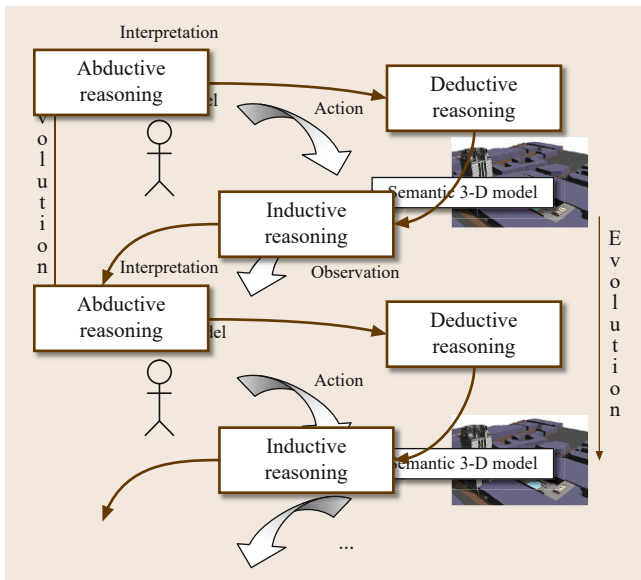


Fig. 45.12 Our original schema of the design process, with an overlay of where the abductive, deductive, and inductive reasoning might be taking place

## References

- 45.1 N. Cross: Forty years of design research, *Des. Stud.* **28**, 1–4 (2007)
- 45.2 N. Bayazit: Investigating design: A review of forty years of design research, *Des. Issues* **20**(1), 16–29 (2004)
- 45.3 C.M. Eastman: New directions in design cognition: Studies of representation and recall. In: *Design Knowing and Learning: Cognition in Design Education*, ed. by C. Eastman, W. Newstetter, M. McCracken (Elsevier, Amsterdam 2001) pp. 147–198
- 45.4 H.W.J. Rittel, M.M. Webber: Dilemmas in a general theory of planning, *Policy Sci.* **4**, 155–169 (1973)
- 45.5 H.W.J. Rittel, M.M. Webber: Planning problems are wicked problems. In: *Developments in Design Methodology*, ed. by N. Cross (Wiley, Hoboken 1984) pp. 135–144
- 45.6 H.A. Simon: The structure of ill-structured problems, *Artif. Intell.* **4**, 181–201 (1973)
- 45.7 K. Dorst: Design problems and design paradoxes, *Des. Issues* **22**(3), 4–17 (2006)
- 45.8 N. Cross: *Designerly Ways of Knowing* (Springer, London 2006)
- 45.9 B. Lawson: *How Designers Think – The Design Process Demystified*, 4th edn. (Architectural Press (Elsevier), Amsterdam 2005)
- 45.10 D. Schön: *The Reflective Practitioner: How Professionals Think in Action* (Temple Smith, London 1983)
- 45.11 H.A. Simon: *The Sciences of the Artificial*, 2nd edn. (The MIT Press, Cambridge 1996)
- 45.12 N. Cross: Designerly ways of knowing, *Des. Stud.* **3**(4), 221–227 (1982)
- 45.13 M.L. Maher, J. Poon: Modelling design exploration as co-evolution, *Microcomput. Civil Eng.* **11**, 195–209 (1996)
- 45.14 J. Poon, M.L. Maher: Co-evolution in design: A case study of the Sydney Opera House, *Proc. 2nd Conf. Comput. Aided Archit. Des. Res. Asia*, Hsinchu, ed. by Y.-T. Liu, J.-Y. Tsou, J.-H. Hou (1997) pp. 439–448
- 45.15 N. Cross: Styles of learning, designing and computing, *Des. Stud.* **6**(3), 157–162 (1985)
- 45.16 L.B. Archer: Design as a discipline, *Des. Stud.* **1**(1), 17–20 (1979)
- 45.17 M. Polanyi: *The Tacit Dimension*, 1st edn. (Doubleday, New York 1967)
- 45.18 M. Polanyi: *Personal Knowledge: Towards a Post-Critical Philosophy* (Routledge and Kegan Paul, London 1958)
- 45.19 J.C. Jones: *Design Methods: Seeds of Human Futures*, 1st edn. (Wiley, Hoboken 1970)
- 45.20 T.B. Ward: Structured imagination: The role of category structure in exemplar generation, *Cogn. Psychol.* **27**, 1–40 (1994)
- 45.21 A. Aamodt, E. Plaza: Case-based reasoning: Foundational issues, methodological variations, and system approaches, *AI Commun.* **7**, 39–59 (1994)
- 45.22 J. Kolodner: An introduction to case-based reasoning, *Artif. Intell. Rev.* **6**, 3–34 (1992)
- 45.23 J. Kolodner: *Case-Based Reasoning* (Morgan-Kaufman, Burlington 1993)
- 45.24 J. Kolko: Abductive thinking and sensemaking: The drivers of design synthesis, *Des. Issues* **26**, 15–28 (2010)
- 45.25 K. Dorst: The core of ‘design thinking’ and its application, *Des. Stud.* **32**, 521–532 (2011)
- 45.26 D. Schön: Generative metaphor: A perspective on problem-setting in social policy. In: *Metaphor and Thought*, ed. by A. Ortony (Cambridge Univ. Press, Cambridge 1979)
- 45.27 N. Cross: Descriptive models of creative design: Application to an example, *Des. Stud.* **18**(4), 427–455 (1997)
- 45.28 J.E. Bogen: The other side of the brain II: An appositional mind, *Bull. Los Angel. Neurol. Soc.* **34**(3), 135–162 (1969)
- 45.29 N. Cross: The nature and nurture of design ability, *Des. Stud.* **11**(3), 127–140 (1990)
- 45.30 A. Heylighen: Building memories, *Build. Res. Inf.* **35**, 90–100 (2007)
- 45.31 M.L. Maher, P. Pu: *Issues and Applications of Case-Based Reasoning in Design* (Lawrence Erlbaum Associates, Mahwah 1997)
- 45.32 G. Goldschmidt: The dialectics of sketching, *Des. Stud.* **4**, 123–143 (1991)
- 45.33 G. Goldschmidt: On visual design thinking: The vis kids of architecture, *Des. Stud.* **15**(2), 158–174 (1994)
- 45.34 K. Grace, R. Saunders, J.S. Gero: Interpretation-driven visual association, *Proc. 2nd Int. Conf. Comput. Creativ. Mexico City*, ed. by D. Ventura, P. Gervás, F.D. Harrell, M.L. Maher, A. Pease, G. Wiggins (Univ. Autonoma Metropolitana, Mexico City 2011) pp. 132–134
- 45.35 K.J. Holyoak, D. Gentner, B.N. Kokinov: Introduction: The place of analogy in cognition. In: *The Analogical Mind: Perspectives From Cognitive Science*, ed. by D. Gentner, K.J. Holyoak, B.N. Kokinov (The MIT Press, Cambridge 2001)
- 45.36 B.N. Kokinov: Analogy is like cognition: Dynamic, emergent and context sensitive. In: *Advances in Analogy Research: Integration of Theory and Data from the Cognitive, Computational, and Neural Sciences*, ed. by K.J. Holyoak, D. Gentner, B.N. Kokinov (NBU Press, Sofia 1998) pp. 96–105
- 45.37 T.B. Ward: Analogical distance and purpose in creative thought: Mental leaps versus mental hops. In: *Advances in Analogy Research: Integration of Theory and Data from the Cognitive, Computational, and Neural Sciences*, ed. by K.J. Holyoak, D. Gentner, B.N. Kokinov (NBU Press, Sofia 1998)
- 45.38 D. Gentner, B.F. Bowdle, P. Wolff, C. Boronat: Metaphor is like analogy. In: *The Analogical Mind: Perspectives from Cognitive Science*, ed. by D. Gentner, K.J. Holyoak, B.N. Kokinov (The MIT Press, Cambridge 2001)
- 45.39 D.R. Hofstadter: Analogy as the core of cognition. In: *The Analogical Mind: Perspectives from Cognitive Science*, ed. by D. Gentner, K.J. Holyoak, B.N. Kokinov (The MIT Press, Cambridge 2001)

- 45.40 G. Lakoff, M. Johnson: The metaphorical structure of the human conceptual system, *Cogn. Sci.* **4**(2), 195–208 (1980)
- 45.41 M. Douglas, B. Isherwood: *The World of Goods* (Allen Lane, London 1979)
- 45.42 C.S. Peirce: *Collected Papers of Charles Sanders Peirce*, Vols. 1–6 (Eds. Charles Hartshorne & Paul Weiss) (1931–1935), Vols. 7–8 (Ed. Arthur W. Burks) (1958) (Harvard Univ. Press, Cambridge 1958)
- 45.43 L.J. March: The logic of design and the question of value. In: *The Architecture of Form*, ed. by L.J. March (Cambridge University Press, Cambridge 1976) pp. 1–40
- 45.44 B. Lawson: Cognitive strategies in architectural design, *Ergonomics* **22**(1), 59–68 (1979)
- 45.45 M.A. Boden: *The Creative Mind: Myths and Mechanisms*, 2nd edn. (Routledge, London 2004)
- 45.46 A.T. Purcell, J.S. Gero: Drawings and the design process, *Des. Stud.* **19**, 389–429 (1998)
- 45.47 N. Cross: Natural intelligence in design, *Des. Stud.* **20**(1), 25–39 (1999)
- 45.48 J. Pallasmaa: *The Thinking Hand: Existential and Embodied Wisdom in Architecture* (Wiley, Hoboken 2009)
- 45.49 J. Conklin: *Dialogue Mapping: Building Shared Understanding of Wicked Problems* (Wiley, Hoboken 2005)
- 45.50 C.M. Eastman, P. Teicholz, R. Sacks, K.K. Liston: *BIM Handbook: A Guide to Building Information Modeling for Owners, Managers, Architects, Engineers, Contractors, and Fabricators* (Wiley, Hoboken 2008)
- 45.51 J.J. Shah, M.T. Rogers: Functional requirements and conceptual design of the feature-based modelling system, *Comput.-Aided Eng. J.* **5**, 9–15 (1988)
- 45.52 K.H. Veltman: Syntactic and semantic interoperability: New approaches to knowledge and the semantic web, *The New Rev. Inf. Netw.* **7**, 159–184 (2001)
- 45.53 M.P. Gallagher, A.C. O'Connor, J.L. Dettbar, L.T. Gilday: *Cost Analysis of Inadequate Interoperability in the U.S. Capital Facilities Industry, NIST Report GCR 04–867* (National Institute of Standards and Technology, Gaithersburg 2004)
- 45.54 S. Gerbino: Tools for the interoperability among CAD systems, *Proc. 13th ADM – 15th INGEGRAF Int. Conf. Tools Methods Evol. Eng. Des. Cassino* (2003)
- 45.55 T. Pazlar, Ž. Turk: Interoperability in practice: Geometric data exchange using the IFC standard, *J. Inf. Technol. Constr.* **13**, 362–380 (2008)
- 45.56 P. Pauwels: Supporting decision-making in the building life-cycle using linked building data, *Buildings* **4**(3), 549–579 (2014)
- 45.57 P. Pauwels, R. De Meyer, J. van Campenhout: Design thinking support: Information systems vs. reasoning, *Des. Issues* **29**(2), 42–59 (2012)
- 45.58 P. Pauwels, R. De Meyer, J. van Campenhout: Interoperability for the design and construction industry through semantic web technology. In: *Semantic Multimedia*, LNCS, Vol. 6725, (Springer, Berlin, Heidelberg 2010) pp. 143–158
- 45.59 P. Pauwels, D. van Deursen, J. De Roo, T. van Ackere, R. De Meyer, R. van de Walle, J. van Campenhout: Threedimensional information exchange over the semantic web for the domain of architecture, engineering and construction, *Artif. Intell. Eng. Des. Manuf.* **25**, 317–332 (2011)
- 45.60 P. Pauwels, P. Present, T. Strobbe: A pragmatic approach towards software usage in construction projects: The Port House in Antwerp, Belgium, *Proc. 9th Eur. Conf. Prod. Process Model. eWork eBus. Archit. Eng. Constr. Reykjavik*, ed. by G. Gudnason, R. Scherer (Taylor Francis, Boca Raton 2012) pp. 509–512
- 45.61 B. Peters, X. De Kestelier: The work of Foster and Partners Specialist Modelling Group, *The Bridges Conf. Math. Connect. Art Music Sci.* (2006)
- 45.62 B. Lawson: Oracles, draughtsmen, and agents: The nature of knowledge and creativity in design and the role of IT, *Autom. Constr.* **14**, 383–391 (2005)
- 45.63 D. Poole: Probabilistic Horn abduction and Bayesian networks, *Artif. Intell.* **64**(1), 81–129 (1993)
- 45.64 P. Pauwels, R. Bod: Including the power of interpretation through a simulation of Peirce's process of inquiry, *Lit. Linguist. Comput.* **28**(3), 452–460 (2013)
- 45.65 P. Pauwels, R. Bod: Architectural design thinking as a form of model-based reasoning. In: *Model-Based Reasoning in Science and Technology*, Studies in Applied Philosophy, Epistemology and Rational Ethics, ed. by L. Magnani (Springer, Berlin, Heidelberg 2013) pp. 583–608
- 45.66 A. Aliseda: *Abductive Reasoning: Logical Investigations Into Discovery and Explanation*, Synthese Library, Vol. 330 (Springer, Dordrecht 2006)
- 45.67 R.D. King, J. Rowland, S.G. Oliver, M. Young, W. Aubrey, E. Byrne, M. Liakata, M. Markham, P. Pir, L.N. Soldatova, A. Sparkes, K.E. Whelan, A. Clare: The automation of science, *Science* **324**, 85–89 (2009)
- 45.68 R.D. King: Rise of the robo scientists, *Sci. Am.* **304**, 72–77 (2011)



## 46. Representational and Experimental Modeling in Archaeology

Alison Wylie

I distinguish, by specificity and representational function, several different types of archaeological models: phenomenological, scaffolding, and explanatory models. These take the form of concrete, mathematical, and computational models (following Weisberg's taxonomy), and they exemplify what Morgan describes as the *double life* of models; they vary significantly in the degree to which they are intended to accurately represent a particular target, or are media for experimental manipulation of idealized cultural processes. At the phenomenological end of the spectrum, representational models of data include typological constructs that selectively represent variability in archaeological data on several dimensions: formal (material), spatial, and temporal. Archaeologists also build phenomenological models of data drawn from nonarchaeological sources – cultural and natural – that are relevant for interpreting archaeological data as evidence. Assemblages of these target and source models provide the necessary scaffolding for building and evaluating more ambitious explanatory and experimental models of cultural systems and processes, actual and hypothetical.

46.1	<b>Philosophical Resources and Archaeological Parallels</b> .....	990
46.2	<b>The Challenges of Archaeological Modeling</b> .....	991
46.3	<b>A Taxonomy of Archaeological Models</b>	992
46.3.1	Phenomenological Models of Archaeological Subject and Source Data .....	992
46.3.2	Scaffolding Models: Measurement Tools and Guides to Interpretation .....	995
46.3.3	Reconstructive and Explanatory Models .....	996
46.4	<b>Conclusions</b> .....	1000
	<b>References</b> .....	1000

Archaeology is nothing if not a modeling discipline. Archaeologists model the data they recover from the archaeological record, the sources on which they draw to interpret these data, the specific events and activities that produced the surviving traces, and the encompassing social, cultural, ecological contexts and processes in which these events and activities took place. And yet there has been persistent ambivalence among archaeologists about models and modeling practices. Early advocates of modeling in archaeology were aligned with the emergence of a self-consciously scientific research program in the late 1960s and 1970s, the New Archaeology. But the defining commitment of the New Archaeology – to move beyond mere description of the record and interpretive speculation about the past; to realize genuinely explanatory understanding of the

past – was most influentially articulated in terms of a vernacular logical positivism [46.1, Part 2]. The explanatory goals of scientific inquiry were characterized in terms of covering-law models, and a programmatic commitment was made to design inquiry as a program of hypothetico-deductive testing; law-like generalizations about cultural systems and processes were to be systematically tested against archaeological data or, if established on other grounds, applied as explanatory principles to archaeological cases. In practice, however, it is typically models that even the most ardent archaeological positivists build and test, not isolated theoretical or factual claims, much less systems of laws or law-like propositions. They routinely make use of models, framed at a number of different levels of specificity, to explain events and conditions in

the cultural past in terms of underlying mechanisms and historically specific processes rather than by subsuming particulars under general regularities. It is more productive, I argue, to think about archaeological practice as a genre of empirically grounded, investigative reasoning with and through models – a perspective elaborated by recent advocates of a *model-based archaeology* [46.2].

This shift in analytic frame may seem straightforward enough but things quickly become complicated when you consider the range of constructs that archaeologists count as models; these are radically heteroge-

neous on a number of different dimensions. The task I take up here is to address the question: What are archaeological models? I propose a taxonomy of the kinds of models archaeologists build and use, distinguished by specificity and representational function. In the process I address two further questions: What do archaeologists use models to do? And, how do they learn from models? I take this to be scaffolding necessary for the normative epistemic task, which lies outside the scope of this chapter, of addressing questions about what makes for a better or worse models and modeling practice in archaeology, given their diverse purposes.

## 46.1 Philosophical Resources and Archaeological Parallels

In framing a taxonomy of archaeological models I draw on a point made with particular clarity by *Morrison and Morgan in Models as mediating instruments* [46.3]: that modeling practice is not well understood if you think of models primarily as tools for operationalizing theory, derived top-down from theory or constructed as *models of theory* for application to real world systems, or as simplified descriptions of phenomena that function as tools for systematizing data built, bottom-up, from the analysis of a specific body of data. Archaeological models are no exception; they are rarely constructed in either of these ways, and even when they approximate to these types of modeling practice their content is often much more complex. In Morrison and Morgan's terms, key classes of archaeological models are autonomous; they incorporate content that is not derived from or reducible to the data they represent or the theories they interpret. As such, they put archaeologists in a position to learn things about an archaeological subject that they could not have learned either from direct empirical investigation or by manipulating – testing, refining, applying – existing theory. That said, I find it useful to think of archaeological models as falling along a spectrum of degrees of abstraction (or idealization) and empirical specificity, with resolutely descriptive, data-systematizing models at the phenomenological end of the spectrum and highly idealized, theoretically motivated models at the other.

*Morgan's* recent discussion of the *double life* of models is a second useful resource, in this case for understanding the variability of purpose evident in archaeological models [46.4]. She makes the case that models in economics figure both as objects of investigation and as tools for investigation; they support experimental as well as representational uses. As I will show, an important class of archaeological models is

quite explicitly designed to support experimental manipulation; they are objects of investigation in Morgan's sense, rather than strictly representational tools. This is a point made in especially compelling terms by *Kohler and van der Leeuw* when they argue that the value of models, as constructs that mediate between "the real world and ourselves," is that they support the "joint exploration of the model and its target system" [46.2, p. 4].

Finally, I draw on an older philosophical literature on models that anticipates, in some respects, recent philosophical thinking about models in science exemplified by *Weisberg's Simulation and Similarity* [46.5], but is particularly useful in an archaeological context because of its focus on the role of analogical reasoning in the construction and use of models. The account of modeling developed by *Harré* [46.6] and by *Hesse* [46.7, 8] brings into sharp focus the complexity of models themselves, and of the relationships that hold between models and their targets and sources. The taxonomy of archaeological models that I outline here presupposes their argument, now much expanded by Weisberg, that sentential and formal, mathematical models by no means exhaust the range of models that figure in the sciences. Iconic or *picturing* models, including what Weisberg describes as concrete models, play a crucial role in empirical inquiry: they "stand in for [...] mechanisms of nature of which we are ignorant"; they allow researchers to "picture possible mechanisms for producing phenomena" *Harré* [46.6, p. 54]. Especially relevant here is the distinction *Harré* draws between two basic types of iconic model: homeomorphic models, in which source and subject are the same; and paramorphic models, in which these are different. A key feature of paramorphic models, where archaeological practice is concerned, is that they may be *multiply connected* [46.6, pp. 47–49]; they in-

corporate elements drawn from a number of sources relevant for modeling different aspects of archaeological subjects that have no comprehensive contemporary analog.

In developing this taxonomy of archaeological models and modeling practices, I am influenced as well by the magisterial analysis, “Models and Paradigms in Contemporary Archaeology” [46.9] offered by the British archaeologist, *Clarke*, as the framework for an early and prescient collection of essays *Models in Archeology* [46.10] and by the distinctions drawn by *Kohler* and *van der Leeuw* in connection with the case they make for a *model-based archaeology* [46.2]. Both recognize the purpose-specific, partial nature of models. *Clarke* emphasizes the different functions served by models pitched at different levels of abstraction, lying on a continuum much like that posited by *Morgan* and *Morrison* for economics and the physical sciences. Archaeological models include what *Clarke* calls *mind models* that function like Kuhnian paradigms; operational models that interpret these orienting conceptual models in observational terms; and models that systematize (selectively and economically) complex bodies of data, serving as heuristic devices for visualizing, manipulating, organizing, and comparing observations [46.9, pp. 2–5]. Together, *Clarke* argues, models of these various kinds are a crucial resource for generating and articulating explanatory hypotheses. Taking up the

cause of archaeological modeling 35 years later, the explanatory function of models is primary for *Kohler* and *van der Leeuw*: “a model here is just a candidate explanation” [46.2, p. 1]. However, invoking *Levins* on the impossibility of simultaneously maximizing generality, realism, and precision [46.11], they also recognize a range of scales and degrees of abstraction in the models archaeologists devise to answer “*how* and *why* questions” [46.2, pp. 1, 7]. These are primarily marked by degree of aggregation; the key contrast for *Kohler* and *van der Leeuw* is between a new generation of agent-based models and earlier systems models (more of this ancestry shortly). *Kohler* and *van der Leeuw* also make a point that figures prominently in *Clarke*’s brief for model-based modes of practice and converges directly on *Weisberg*’s taxonomy of scientific models: that models come in a great many different forms. Their roster of mental, verbal, physical, and formal (mathematical and simulation) models is reminiscent of *Clarke*’s argument that models can be constructed as physical *homomorphic parallels* between model and target or can take the form of formal (mathematical) representations of abstract systems of relationships inherent in the target [46.9, p. 41]. In short, despite an emphasis on discontinuities, these programmatic arguments for *model-based* archaeology reflect significant continuities in evolving archaeological practice to which I hope the taxonomy proposed here does justice.

## 46.2 The Challenges of Archaeological Modeling

Some advocates of the explanatorily ambitious New Archaeology did make the case for what they described as a *systems* rather than a *law and order* approach which put modeling at the center of archaeological inquiry [46.12, 13], which in some respects anticipated arguments made with considerable force in the UK by *Clarke* [46.9]. In the discussions of modeling associated with the New Archaeology the emphasis was initially on theory-driven, *whole system* models, usually of an explicitly eco-determinist cast; these were intended to capture the essential causal and structural features of distinct types of cultural systems and the processes by which they adapted to the ecological contexts in which they took shape and evolved over time. But from the outset the constraints on *whole system* modeling and on the mathematical and simulation modeling techniques used to operationalize archaeological theories of cultural process were recognized to be all but insurmountable in explanatorily interesting cases. In a classic statement dating to 1975, echoed in a number of later assessments, *Doran* and

*Hodson* identified three pivotal problems. First, they observed, “models which are mathematically tractable are too simple for most archaeological problems” [46.14, p. 315]. This is not just a technical constraint. Although computer technology was, even then, making possible simulations that could better cope with the computational challenges of modeling whole systems, a second more fundamental problem is that these models require a level of understanding of the conditions and processes modeled that is “only rarely met in archaeological work” [46.14, p. 315]. Finally, an inescapable problem for archaeological modelers noted by *Doran* and *Hodson* and reiterated many times since is the “fundamental noisiness” of archaeological data which makes it difficult to empirically assess the descriptive and explanatory claims about the cultural past captured by or derived from these models [46.15, p. 230].

Despite this early pessimism about the prospects of ever realizing the explanatory ambitions of the New Archaeology by means of modeling approaches, models

are ubiquitous in archaeology. As Kohler and van der Leeuw put it, archaeologists [46.2, p. 3]:

“have drifted *in practice* toward what philosophers of science call a *model-based* (Giere 1999) or *semantic* (Lloyd 1988; Suppe 1977 [...]) approach to the task of explaining what happened, and why, in prehistory.”

This in part due to the proliferation of fast, cheap computer technology but, even with the promise that Doran and Hodson’s first problem might be resolved, the archaeological models that now answer the call to explanatory understanding are typically much more narrowly circumscribed than the whole system models initially advocated by allies and critics of the New Archaeology. This reflects, in part, a growing appreciation of the complexity of the human, social *ecodynamics* by which cultural systems modify as much as adapt

to their environments [46.2, p. 10]. If anything, this makes Doran and Hodson’s second two concerns even more acute. In response, archaeologists have shifted their focus to building and refining models of specific conditions and processes that are, or that could have been, responsible for specific types of event or forms of life to which the archaeological record bears witness. As Kohler and van der Leeuw describe the mandate for a *model-based archaeology* at this juncture, it is to understand “relatively small-scale” human systems, but to understand them in something closer to their full complexity: as “embedded within [...] the environments they inhabit and alter” [46.2, p. 2]. At the same time, a broad cross-section of archaeologists have embraced a wide range of more prosaic modeling practices that are resolutely descriptive and phenomenological, but that are no less crucial to the broader explanatory goals of contemporary archaeology.

## 46.3 A Taxonomy of Archaeological Models

With these conceptual resources in hand, consider some of the types of work-a-day models that abound in archaeology.

### 46.3.1 Phenomenological Models of Archaeological Subject and Source Data

One prevalent use of models in archaeology is to characterize, in systematic terms, various types of archaeological data, and the diverse experimental and ethnohistoric sources on which archaeologists rely to interpret these data as evidence. These models take a range of forms: mathematical, computational, and in some cases concrete, to use Weisberg’s categorization [46.5, Chap. 2]. They are typically homeomorphic models designed to represent variability in the target or source domain. As such, they lie at the phenomenological end of the spectrum of model function marked by Morgan and Morrison although, I will argue, they incorporate much often unrecognized theoretical and interpretive content.

#### Models of Archaeological Data

Models of archaeological data typically represent variability on three dimensions: formal, material variability; the spatial distribution of artifacts and features within sites, or of sites and assemblages of artifacts across a region; and chronological trends in the appearance, frequency, and disappearance of artifact types, architectural styles, and cultural formations over time.

Material, formal variability in archaeological data is captured by descriptive typologies, ranging from highly specific artifact typologies aimed at systematizing local variability in material culture to expansive classification schemes that delineate trans-historical cultural formations and trans-regional cultural horizons. At the artifact-specific end of the spectrum, ceramic and lithic (stone tool) typologies have been especially crucial in many contexts as chronologically and spatially sensitive markers that came to anchor the characterization of *archaeological cultures* [46.16]: distinctive assemblages of archaeological material – for example, stylistically distinctive artifacts, house forms, burial rites and subsistence practices that consistently co-occur – that were presumed to be the expression of distinct cultural configurations. So, for example, the late Neolithic culture(s) of western Europe that came to be known as the *Beaker people* were characterized archaeologically by a *package* of artifacts associated with a characteristic type of pottery, and the hunting-intensive Paleo-indian cultures of central North America were named for the distinctive Clovis and Fulsom projectile and spear points in terms of which they were first identified in the 1920s and 1930s. Broader syntheses of archaeological cultures, of the kind posited by Willey and Phillips for the Americas [46.17], and by Childe for Europe [46.18], characterize broad cultural horizons based on the sequence of appearance and distribution of these co-occurrent classes of archaeological evidence.

The presumption that formal (material) variability of this kind has inherent cultural significance has

been a matter of sharp contention within archaeology since at least the 1940s and 1950s, when *Brew* and *Ford*, and later *Spaulding*, articulated opposing views about what these models of data represent: that they are selective, purpose-specific impositions by the analyst [46.19, 20], as opposed to culturally salient features of the archaeological record that archaeologists *discover* (Spaulding [46.21, 22] and see *Adams* and *Adams* [46.23] and *Wylie* [46.1, pp. 42–51] for analysis of this debate). To illustrate the contingency and purpose-specificity of typological systems, Ford offered a thought experiment: a variety of house forms on the fictional Island of Gama Gama characterized by a range of different traits (e.g., roof style, construction on stilts, size, layout) whose variability is continuous across time and space. Although regularities in the distribution and association of these material traits can certainly be identified empirically (indeed, *statistically discovered*, as Spaulding had insisted), it is possible to carve material culture at different joints. Shifting the selection of traits will yield different patterns of association and spatial/temporal distribution, and often enough their variability is continuous so that different boundaries can be drawn between types [46.24]. Ford's point was that archaeological typologies are tools of analysis, constructed as needed to address particular archaeological questions. The typologies that served archaeologists initially in modeling spatial and temporal relations within and between classes of archaeological data may not be a plausible proxy for cultural identity, or useful in tracking shifts in technologies, subsistence practice, trading relations or social status, to name just areas of archaeological interest.

Spatial distribution models vary dramatically in scale, target, form, and purpose. They include, for example, spatial auto-correlation models that delineate artifact drop-zones around hearths and in activity areas, and a range of other models that capture the spatial relations between key features within archaeological sites. Classic examples are models that represent regularities observed in the orientation of burials and associated grave goods in mortuary sites (see the example of a Roman period cemetery in the UK discussed below), and the patterned clustering of functionally distinct rooms in Southwestern pueblos that was the basis for *Hill's* posit of generationally stable households at Pueblo [46.25], an early demonstration project for the New Archaeology. They include, as well, models of an architectural *grammar* of the kind developed by *Glassie* for Middle Virginia folk housing: an inventory of geometric forms structured by a basic unit of measurement (the diagonal constitutive rectangles and squares) and a set of grammatical rules for assembling these into canonical house forms [46.26]. At a regional

scale archaeologists develop formal and computational models of the distribution sites or visible features on the landscape, now facilitated by widespread use of geographical information systems (GISs). For example, spatial packing models (imported from quantitative geography) were developed to capture the proxemics of settlement hierarchies which, in turn, were the basis for positing regional chiefdoms in Neolithic Europe [46.27]. More recently, landscape archaeologists have developed richly interpretive spatial models of the sacred (rather than political) landscapes in which Neolithic and Bronze Age monuments like Stonehenge are embedded [46.28, 29]. A related example that incorporates experimental elements (of which, more below) is *Llobera's* delineation of corridors of movement between Neolithic Galician mamoas, identified both in terms of ease of movement, given regional topology, and the viewscape afforded travelers along these pathways [46.30]. In these cases, digital repositories of spatial data and the analytic power of GIS analyses are a crucial resource; predictive modeling of where archaeological sites of various kinds are likely to occur is now a key component of cultural resource management [46.31].

Chronological models represent the appearance and disappearance, and related changes in the form and frequency of specific types of material culture over time, and at scales ranging from individual artifact types, cross-type styles and assemblages, to broad cultural formations. The locus classicus for such models is *Kroeber's* decidedly nonarchaeological seriation of changes in fashion in which he determined that, despite the perception of rapid and dramatic change, the proportions that define what is fashionable change very slowly and predictably; his test case was the evolution of styles in women's formal wear from 1845 to 1915 [46.32]. Influential examples developed to illustrate these seriation principles in archaeological terms include *Deetz* and *Dethlefsen's* classic analysis of changes in the frequency of decorative styles in New England tombstones; they demonstrated the same regular *battleship curve* in stylistic changes over time as had *Kroeber* [46.33]. Another example that continues to be used as a basis for building finegrained chronologies in historical archaeology is *Binford's* formal model of a regular pattern of change in the mean bore hole diameter of clay tobacco pipes produced in Europe and North America between 1600 and 1900, described as *deterministic and mathematical* by *Clarke* [46.9, p. 18]. Although physical dating techniques are now predominantly the basis for archaeological chronologies, tradition-specific seriation models continue to be a key resource in many contexts. Indeed, although it was widely assumed that local, *relative* chronologies

would automatically be replaced when radiocarbon dating was introduced – the *first radiocarbon revolution* (initiated by Libby in the 1950s) – in fact, discrepancies between these systems have been pivotal in raising questions about the accuracy of absolute chronologies that generated the painstaking, 60 year process of calibrating radiocarbon dating curves – the *second radiocarbon revolution* [46.34, pp. 130–140].

The challenge of establishing spatial-temporal control dominated the initial construction of typological systems in most contexts of archaeological research, but this by no means exhausts the purposes for which they are used. As *Boozer* observes in a discussion of the “tyranny of typologies” [46.35], once these phenomenological models were developed, practitioners often lost sight of the purposes they were designed to serve. They became entrenched as the dominant medium of communication within archaeology; they configure reporting conventions and set the framework for the comparative analyses that were the basis not only for regional models of cultural diversity and evolution, but also the fine-grained analogical comparisons that underpin interpretive claims about the evidential significance of archaeological data. Often they persist even when accumulating data undermines the distinctions they draw and as focal questions change, requiring analysis in terms of traits that track other dimensions of variability. As the disconnect between these models of data and evolving research agendas becomes increasingly strained, the central point made by Brew and Ford in the mid-1950s is more relevant than ever: that in constructing typologies archaeologists must choose among a great many observable, measurable traits, so any one selection necessarily reflects specific investigative purposes. In analysis of a problematic typology of domestic Romano-Egyptian house forms, *Boozer* draws attention to the ramifying downstream consequences of failing to keep the contingency of these models of data clearly in view, reifying them as representations of a fundamental cultural reality and treating them as the framework within which all subsequent research must be conducted [46.35, pp. 104–106].

#### Models of Nonarchaeological Sources

A second important genre of phenomenological modeling in archaeology is of the data drawn from nonarchaeological sources on which archaeologists rely to interpret archaeological data as evidence. These are also typically homeomorphic models, in this case of natural or cultural processes that are presumed to be responsible for (or that could have been responsible for) the production, deposition, and preservation or degradation of the types of material that make up the archaeological record: *N-transform* and *C-transform* models, to use

language introduced in the 1980s by *Schiffer* in connection with widely influential account of archaeological inference [46.36].

Archaeologists rely on an enormously broad range of other fields, from ethnography and history to biomedicine, ecology, and physics for the background knowledge necessary to build these models. But as useful as these resources are, often archaeologists find that the cultural and/or natural processes of interest to them have not been intensively studied, or not studied at scales or in contexts relevant to archaeological questions. The fields of experimental archaeology and ethno-archaeology have grown up in response to these limitations. At the C-transform end of the spectrum, the Kalinga Ethnoarchaeological Project is one example of a long-running research program in which archaeologists have undertaken their own ethnographic research with the aim of documenting methods of production, exchange networks, and patterns of cultural transmission, in this case, of ceramic technology [46.37]. A recent report on this project includes a directional graph of household pottery exchange: a phenomenological model of ethnographic data relevant to the question of whether shifting patterns in ceramic production and exchange can serve as a proxy for intensification in a craft-based agricultural economy [46.37, p. 43]. N-transform modeling includes, for example, the uses archaeologists make of well-established geological models of soil formation and erosion processes to understand archaeological deposits. But here again archaeologists often develop their own models of the impact that, for example, the activities of burrowing animals and insects can have on archaeological features and stratigraphy: “bioturbation” and “faunalturbation” [46.38, pp. 271–276]. A classic example is *Stein’s* model of the rate at which earthworms can completely turn over an archaeological midden, obscuring archaeological features and redistributing cultural material [46.39]. A number of experimental archaeologists have taken this a step further, building concrete models designed to provide insight into the processes by which particular classes of artifacts could have been produced or transformed over time into distinctive types of archaeological deposit. Bell [46.40] describes a number of experimental projects in England and Europe that involve full-scale recreations of key archaeological features, like earthworks and mounds, house structures and middens, which are then monitored, sometimes over decades, for patterns of collapse and erosion. The identification of weed complexes that are diagnostic of different types of early Neolithic farming practices in Europe (described below) depends on phenomenological models of bioecological conditions under which weed and food crop species co-occur, and the results of agri-

cultural experiments designed to model the impact on these plant assemblages of different plant husbandry regimes [46.41, 42].

Recent developments in archaeometallurgy, dietary studies, and radiocarbon dating, among other areas, also illustrate the complexity of putting external resources to work in archaeological contexts, particularly when this requires modeling physical or biochemical processes that are affected by and that reciprocally shape human activities. For example, *Pollard and Bray* [46.43] make the case that provenience studies of European Bronze Age metal artifacts has run aground; the complexity of the chemical composition of these artifacts undermines a long-running program of analysis aimed at linking individual artifacts or assemblages to particular sources of raw material. Rather than persist in the attempt to disentangle a signal linked to origin from the noise of degradation – an approach they describe as narrowly scientific – they argue for an alternative that takes as its point of departure the assumption that the chemical components of these objects are themselves dynamic, the product of jointly social and technological/material histories of circulation, reuse, repurposing. To this end they identify distinct types of copper based on the presence or absence of trace elements that reflects the differential effects on them of oxidation and interaction under conditions of repeated melting, mixing, and recycling: a phenomenological model of variability in the chemical composition of this class of material [46.43, pp. 118–120]. Similarly, complex phenomenological modeling is required to make use of stable isotope and trace element analysis of skeletal material as a basis for reconstructing dietary profiles. In an example discussed below, this involved modeling the clines in the chemical composition of groundwater across England and Europe in order to estimate the geographical origins and lifetime travels of individuals buried in a late Romano-British cemetery. Finally, the process of refining radiocarbon dates likewise depends on integrating evidence relating to physical, climatic, ecological conditions that can affect the ratio of radioactive to stable carbon in organic matter recovered from archaeological contexts: for example, fluctuations in atmospheric carbon levels, carbon sinks, patterns of carbon uptake, sources of contamination. While these N-transform models focus on factors affecting the radiocarbon signal itself, the characteristic approach of the *third radiocarbon revolution* has been a *pragmatic Bayesianism* ([46.34, pp. 140–141], [46.44, pp. 217–218]): a strategy of modeling the probability distributions for a range of radiocarbon dates that could have been produced by an organic sample. This approach takes into account not only N-transform processes that affect the measured ratio of stable to radioactive carbon in a sample,

but also multiple lines of archaeology-source evidence including, for example, stratigraphic superposition and seriation.

These examples of phenomenological models – models of data associated with archaeological subjects and sources – illustrate the now well-established point that seemingly straightforward descriptive, representational models are actually quite complex conceptually. Even when, on the face of it, they seem to be abstracted directly from the phenomena, and their source and subject is ostensibly the same, they incorporate substantial purpose-specific theoretical and interpretive, as well as descriptive, content.

### 46.3.2 Scaffolding Models: Measurement Tools and Guides to Interpretation

The complexity of phenomenological models arises, not just because their targets and sometimes their sources are complex, but because their purposes are complex; they are intended to serve a number of inferential and investigative purposes beyond systematizing the data they represent. Models of source data are intended to capture projectable relations between the physical traces that survive in the archaeological record and the antecedent events, conditions, and processes that produced them. Well constructed, they are mediators in a rather different sense than that introduced by Morgan and Morrison; they function in archaeological interpretation as auxiliary hypotheses that mediate the interpretation of archaeological data as evidence relevant for positing and testing hypotheses about the archaeological target of interest: cultural events and activities, conditions of life, systems, and processes. Here are two examples in which archaeologists make use of phenomenological models of source and subject data in this scaffolding sense, as *measurement tools or interpretive guides*.

*The Roman Diaspora Project.* In this project archaeologists make sophisticated use of an array of N-transform and C-transform models to specify the likely origins and lifetime travel of individuals buried in a late period Roman cemetery in Winchester and York, UK [46.45, 46]. The catalyst for this study was an interpretation, dating to the 1970s, of the formal traits of burials in this cemetery – skull morphology; epigraphy, statuary, and associated artifacts; and patterns of spatial orientation and distribution – that were taken to be markers of cultural affiliation and status associated with degree of *Romanization*, resistance to Roman rule, and North African or Eastern European origins as *incomers*. The Diaspora project team undertook to develop dietary profiles based on isotope and trace element analysis of

bone marrow and dental cores that allowed them to determine where individuals were likely born and spent their early years, as well as where they had lived and traveled. This analysis depends crucially on two types of phenomenological models mentioned earlier: models of cross-continent clines in the mineral composition of groundwater, and of the isotopic signatures of various types of diet. The upshot was that several individuals who had been identified originally as incomers had most likely been born and raised in the vicinity of the cemetery where they had been buried; others who most likely originated in North Africa were buried in graves that had been interpreted as elite; and several children proved to have originated outside the region where they were buried. These scaffolding models were, then, the basis for characterizing the status and mobility of individuals buried at Winchester and York in terms that pose a substantial challenge to the earlier interpretation of their remains and the canonical, text-based accounts of population diversity and mobility in the Roman Empire that had informed this interpretation [46.45, 46].

*Farming practice in Neolithic Eastern Europe.* The objective of this project was to adjudicate between competing models of the farming practices that had been adopted in various locales as agricultural subsistence patterns were taking shape in Eastern Europe through the Mesolithic to Neolithic transition 10 000 years ago [46.41]. Each of these models had some support, and each had different explanatory implications for understanding the impact of this major transition in subsistence practice on settlement patterns, material culture, social relations, and population mobility. It had proven difficult to discriminate between these competing models not least because the contemporary analogues for each type of practice involve suites of plants – cultigens and weeds – now adapted to ecological settings that have been continuously reconfigured through millennia of intensive human activity. To determine which types of farming practice were adopted at various junctures and in different locales Bogaard developed a series of scaffolding models of functional plant ecology that incorporate the phenomenological models of experimental and bioecological data mentioned earlier. These scaffolding models represent the distinctive complement of weeds associated with each type of crop and crop management, for example, intensive rather than extensive agriculture, shifting rather than fixed-plot cultivation, and spring rather than winter cropping [46.41, pp. 154–159]. The background knowledge from plant science and archaeobotany provided an initial set of posits about these weed complexes, refined through a program of experimental archaeology designed to recreate hypothesized Neolithic farming practices and document their ecological viability, labor

demands, yield, and archaeological signatures. Bogaard thus constructs archaeological proxies for the major crop husbandry models on offer and uses these lower level scaffolding models as the basis for systematically assessing the representational plausibility of each of them in specific prehistoric contexts.

In these two cases archaeologists build an assemblage of homeomorphic phenomenological models of source as well as archaeological data that, together, serve as scaffolding for the interpretation of archaeological data as evidence of specific past events and practices. As interpretive scaffolding, these models serve as the basis for analogical arguments that make possible systematic comparisons between the sources of interpretation (natural and/or cultural processes observed in the present) and the archaeological targets or subject of interpretation [46.47, 48]. While no one line of evidence based on scaffolding models is likely to be decisive, they can be used very effectively, in combination, to build and test broader reconstructive and explanatory claims about the past. The principle at work here is that such *cables* of argument will be compelling to the extent that the scaffolding models used to construct distinct lines of evidence are causally and epistemically independent of one another; this “vertical independence” is the key to ensuring that they have the capacity to be mutually constraining [46.49, p. 387], [46.50].

### 46.3.3 Reconstructive and Explanatory Models

As these examples suggest, assemblages of scaffolding and phenomenological models are the basis for building explanatory models of the cultural past of the kind that are identified as the central goal of archaeology. These last are *complex paramorphic models* either of particular archaeological targets (specific past cultures) or of generalizable types of cultural system or cultural process. Consider three reconstructive models that address explanatory questions, and that bring into sharp focus two key dimensions on which archaeological modeling varies: in degree of idealization as opposed to representational fidelity to a specific subject past; and in the nonrepresentational use of models in an experimental mode, as objects of investigation.

*Representational models: the Desert Archaic simulation.* This is a classic whole-system simulation of prehistoric subsistence practice in the Great Basin (US) known as the “Desert Archaic” that was developed by *Thomas* [46.51], in the spirit of the New Archaeology, to determine whether the Shoshone seasonal round documented in the 1930s could be projected back in time: whether it could be treated as, in effect,



a homeomorphic model of the subsistence practices of antecedent, archaeologically identified cultures in the region. Thomas' strategy was, first, to develop a computational model of the source: *Steward's* 1938 ethnographic account of the seasonal round of Shoshone foragers in the Great Basin [46.52]. He then "reduce[d] the activities [modeled] to their correlative tool assemblages", ran this single year model a thousand times, corrected for the impact of less frequently available resources, and in this way generated aggregate patterns of artifact deposition for the region. He then tested the model output against the results of archaeological surveys in the region, establishing that drop patterns for most artifact classes did conform to expectations, most strikingly in case of sites located in open areas where they had not previously been archaeologically documented. This process also threw into relief several empirical and inferential weaknesses inherent in his model, for example, the tool types differentiated by edge angle proved not to be reliably diagnostic of the functionally different types of site posited by the model. In short, one result of testing the expectations generated by his simulation of the Shoshone ethnohistoric seasonal round was to make it clear that the lithic typologies on which archaeologists conventionally relied in this region – phenomenological models of this class of archaeological data – were too coarse-grained with respect to tool function to be reliable scaffolding for the interpretation of these data as evidence relevant to questions about subsistence practice.

This reconstructive, explanatory model is explicitly representational and homeomorphic, at least aspirationally; it is intended to capture *how actually* Shoshone foragers exploited the resources afforded by the Great Basin over the 1000 year period precontact. It is credible to the extent that the models of data (source and subject) used to generate test outcomes are themselves well established and fit for purpose, and to the extent that they are also causally and epistemically independent of the overarching model they are meant to test. The principle at work here is that the material signature posited for each element of the Shoshone seasonal round should not nepotistically ensure that archaeological data will conform to expectation; I describe this elsewhere as a requirement of "vertical independence" [46.49, p. 381].

*Hybrid representational and experimental models: Gila Naquitz* (the early Mesoamerican village). This is a more sophisticated computational model of the evolution of the foraging and farming practices of a hypothetical microband developed by *Flannery* and *Reynolds* in the mid-1980s [46.53]; it answers *Flannery's* earlier call for attention to modeling approaches when the New Archaeology was taking shape [46.12]. One goal

of this modeling exercise was to simulate the process of incremental change in subsistence practices evident in the archaeological record of a cave site in the Oaxaca valley (8700–6600 BC). The simulation developed was, in this respect, a representational model built up from an assemblage of subsidiary homeomorphic scaffolding models; *Flannery* and *Reynolds* report that they did "everything we could think of to make the model realistic" with respect to the climate and paleoecology of Gila Naquitz, and the "wide spectrum" repertoire of foraging resources exploited by its late Holocene occupants that had been documented archaeologically [46.53, p. 436]. In addition, however, this model incorporates a crucial experimental component; *Flannery* and *Reynolds* manipulate key elements of the model to test the impact of different intergenerational processes of community learning from trial and error in face of fluctuating climatic and ecological conditions [46.53, p. 441]. *Flannery's* larger purpose is to assess the credibility of competing explanatory accounts of how and why agriculture developed, apparently independently, in a great many locales around the world at roughly the same time (10 000–5000 BC). He argues that archaeological and paleo-ecological evidence calls into question conventional appeals to exogenous forcing factors like environmental crisis, and urges archaeologists to consider the explanatory potential of accounts that posit more gradual processes by which incipient agriculture emerged as an extension of foraging practices, driven as much by internal social processes as by pressures to adapt to climatic variation in the early Holocene.

*Flannery* and *Richard's* strategy was to develop a computational model in two stages, simulating first the evolution of wide spectrum foraging in the area of Gila Naquitz, and then the emergence, in this context, of incipient agriculture. To make these models as realistic as possible, the repertoire of subsistence activities represented in each of these two stages was based on archaeological data that establish what resources were being exploited when the cave at Gila Naquitz was occupied; the climate was modeled as generating wet, dry, and average years randomly, based on paleoclimatic data; and the assignment of values to such variables as availability, yield, labor requirements and dietary return for the dozen key sources of food exploited by the microband was based on scaffolding models of region-specific archaeological and paleoecological data. In addition, *Flannery* and *Reynolds* developed several hypothetical subroutines to model the information-sharing and decision-making practices by which the hypothetical foragers could learn from trial and error experimentation with different resource collecting schedules and modifications to their repertoire of collecting strategies. This jointly realistic and

experimental simulation was initially run for foraging strategies alone and showed rapid improvement in efficiency until, after about 500 iterations, it proved hard to improve on the established pattern; at this point positive feedback for change shifted to negative feedback encouraging conservatism. Then they introduced several archaeologically documented incipient agricultural strategies to the repertoire – for example, clearing thorn forest to allow weedy plants (beans, and squash) to colonize, and deliberately planting maize and squash seeds – and simulated another learning process. In this second stage simulation the foraging strategies of the hypothetical microband gradually shifted, incorporating the full suite of agricultural strategies until they reached stable performance in 550 iterations.

The adequacy of the model as a representation of the real system should be evaluated in two ways, Flannery and Reynolds argue. First, as with Thomas' model, it should be assessed in terms of the correspondence of model outcomes with actual outcomes documented archaeologically – specifically, outcomes not built into the original simulation. Key measures of success were congruence in the relative emphasis on each plant species exploited for both models; the order in which changes in practice and shifts in emphasis emerge in the case of incipient agriculture model; and in the time frame for stabilization in both models. Second, Flannery and Reynolds add an assessment of model robustness that depends on experimental manipulation of model parameters and inputs. For example, to assess the role and plausibility of the multigenerational learning processes they had built into the foraging model they disabled the information feedback loop and found that performance peaked early but then oscillated in a manner quite unlike anything suggested by the archaeological record. They also changed the environmental conditions and population density under which agricultural strategies were adopted and found that the random alternation of wet, dry, and average years is a crucial stimulus for the experimentation and learning processes that, in the simulation, give rise to incipient agriculture. Under conditions of substantially greater climatic or populational stress the hypothetical band proved to be more conservative, while under conditions of lower stress the band's subsistence strategies fluctuated without the directional intensification of practice observed archaeologically.

This, then, is a computational paramorphic model poised between modeling *how actually* and *how possibly* incipient agriculture took shape in the Oaxaca valley. It incorporates a number of subsidiary homeomorphic models – analytic and descriptive models of climate, ecology, subsistence strategy – but reaches beyond them to model archaeologically enigmatic socio-

cognitive factors. As such, this model is autonomous in the sense outlined by Morgan and Morrison [46.54] and, given this autonomy, it manifests the double life of models discussed more recently by Morgan [46.4]. The simulation developed by Flannery and Reynolds serves both as a tool for investigating the archaeological subject, for which representational adequacy is key, and as an object of investigation in its own right. Experimental manipulation of the model generated a number of insights into causal dynamics of the system that could not be directly investigated, and suggests that intergenerational learning from trial and error can result in extensions of foraging practices that ultimately transform them into agricultural practices. In short, exploration of the hypothetical world of the model provides at least preliminary support for their more general contention that you do not necessarily need to posit a prime mover external to the system to account for major cultural transformations; these may well be explicable in terms of incremental changes in a number of interlinked social practices and ecological conditions.

Sophisticated models designed to simulate complex, path-dependent interactions between multiple causal and ecological factors, including decision making processes and social dynamics, have since been developed in a number of connections. In a recent optimal foraging model of the Pleistocene colonization of Sahul (Australia-New Guinea), O'Connell and Allen are explicit in rejecting *minimalist* models that downplay the cognitive and technological sophistication of these incoming foraging populations [46.55, p. 5]. Their model incorporates ethnographically and ecologically informed submodels of decision-making that had reciprocal impact on the complex environments they entered, under conditions of short-term climatic instability [46.55, p. 12]. Contributors to *Model-Based Archaeology* [46.56] likewise emphasize the complexity of the processes by which human populations modify their environments and, in turn, reconfigure their practices and technologies in response to environments they have in part created [46.57, p. 61]. Their agent-based models are built up from a great many scaffolding models that are as realistic as possible, given available archaeological and paleoecological data, but also incorporate what Kohler et al. refer to as “cultural algorithms” [46.57, p. 89]. These models are then a platform for simulating the impact of various types of stress and shifts in social organization or learning process. For example, Wilkinson et al. [46.58] develop a baseline model of an Early Bronze Age Mesopotamian settlement that they describe as a plausible, but “static view of settlement and land use” [46.58, p. 192]. They then build agent-based simulations that incorporate a number of key behavioral patterns (reciprocal exchange, kinship and subsistence

activities) in order to explore the effects of chronic or acute labor shortages and disease on settlement population and household viability. These simulations are not representational, but they provide an insight into factors that affect settlement sustainability “from the standpoint of the individual household agents” rather than at the level of the settlement as a whole and its “aggregate properties” [46.58, pp. 201, 203]. They illustrate “different evolutionary trends that households can follow” within the same socio-ecological environment, and in the process bring into focus conditions under which aggregate system behavior can abruptly change as a consequence of agent-level decisions that push the system toward “a hidden resource threshold” [46.58, p. 206].

*Experimental models: Hopi agriculturalists in the US Southwest.* This is a suite of even more hypothetical “how possibly” models developed by *Hegmon* [46.59] and *Robertson* [46.60] to explore the impact of different food sharing practices on the survival rates of households in a small-scale farming community, and the potential for (some of) these practices to generate stratification [46.60, p. 13]. Although they rely on well established phenomenological and scaffolding models of the paleoecology, settlement patterns, and social organization in the prehistoric Southwest, their purpose is not to model the dynamics of any particular ancestral Hopi farming community. Rather it is to investigate various properties and dynamics of the model itself. *Hegmon*’s initial model is a highly idealized computational model designed to simulate the survival rates of a dozen households in a hypothetical farming community that practices traditional ethnographically documented Hopi-style maize farming on three different kinds of fields under typical Southwest conditions of crop yields in wet and dry years. She asks what the survival rates for individual households would be if, rather than sharing food in dry, low-yield years, each household kept its own produce to itself, if they shared only in years of scarcity, or if all households consistently pooled their produce. For multiple runs of 20-year simulation cycles she found that households had only a 45% survival rate if they relied exclusively on their own produce. By contrast, on a *restricted sharing* scenario the survival rate was 80% for households in communities of four or more households. Pooling all produce generated equivocal results.

A related model developed by *Robertson* [46.60] relies on the same basic set-up but simulates the effects, over time, of two different sharing arrangements: an egalitarian, *credit-dispersing* strategy by which the shortfall of individual households is met through redistribution of a pool of total community surplus, and a *credit concentrating* procedure by which the household with the largest surplus has the first opportunity

to redistribute, starting with households with the smallest shortfall and meeting the needs of as many deficit households as its surplus permits. *Robertson* finds that [46.60, p. 13]:

“restricted sharing practices not only enhance household survival rates but also have the potential to lead to the growth of rather high levels of both debt and credit without any overt political maneuvering.”

For 100 runs of 40-year simulation cycles most households canceled out their credit or debt to one another, but some households did significantly better than others. Crucially, *Robertson* reports that these results were not tightly correlated with differences in the quality of the fields allocated to a household, and that they are robust even under the credit dispersing strategy, with some amplification under the credit concentrating strategy.

Despite their reliance on realistic, if highly idealized, baseline models of the regional ecology and of ancestral Hopi farming practices, these models are constructed primarily for purposes of experimentation, not to simulate the dynamics of any actual archaeological community as in the case of the models developed by *Thomas* [46.51], *Flannery* and *Reynolds* [46.53], or *Wilkinson* et al. [46.58]. The value of these models is heuristic; they allow *Hegmon* and *Robertson* to test hypothetical claims about the cumulative effects that different social arrangements could potentially have on the distribution of wealth in Southwestern communities that cannot be directly tested archaeologically. In the process, they show that significant social stratification can emerge without having to introduce the mechanisms of a chiefdom-style political formation. *Kohler* et al. [46.57] describe similar goals in connection with an agent-based simulation of the performance of farming households in the context of a suite of highly realistic, archaeologically constrained resource models of the environment in which prehispanic settlement patterns would have evolved in southwestern Colorado (600–1300 AD). Their ultimate goal is to understand archaeologically documented cycles of colonization, settlement concentration, and depopulation in the region [46.57, p. 63] but their primary interest in the simulation is in “abstract properties of the simulated exchange systems” [46.57, p. 96]. They defer assessment of the archaeological plausibility of these simulations, noting that their discussion of factors that make a difference in this simulation “is purely hypothetical”; the value of the simulation is its “power [...] to show us alternative worlds” which, even if they did not exist, “may be able to tell us many things about the worlds that did” [46.57, pp. 99–100].

## 46.4 Conclusions

I draw three conclusions from this taxonomy of archaeological models.

First, the diversity of archaeological modeling practices reinforces analyses developed in other contexts, most pointedly, in philosophical terms, by *Weisberg* in *Simulation and Similarity* [46.5], and by *Kohler* and *van der Leeuw* in their brief for model-based archaeology [46.2]. What counts as adequacy in model construction depends fundamentally on what the model in question is meant to do, and this is an irreducibly pragmatic issue: a matter of research priorities, technical capabilities, empirical and interpretive resources.

Second, reconstructive and explanatory models of the cultural past are assemblages of smaller scale phenomenological and scaffolding models that, together, represent specific factors, variables or processes presumed to constitute the archaeological target, whether this is a particular event, a local set of practices, or large-scale cultural systems and long-term processes. Taken as a whole, these assemblages are multiply connected paramorphic models; they are constructed analogically, and their content derives from homeomorphic models of subject-domain archaeological data and of source-domain data drawn from a diverse array of other fields.

Finally, models at one scale, or models of one dimension of a cultural system or life-world, are the basis for testing and refining models pitched at other scales or that represent other dimensions of the target. Claims about the empirical, theoretical credibility of an explanatory account of the past typically concern the credibility of model components, themselves narrowly specified models of particular aspects of the past cultural context or process or system under study. On a modeling approach, evidential constraints are thus diffuse, impinging on archaeological understanding of the cultural past at a number of points; testing model outputs against source data or archaeological data may suggest the plausibility of the model as a whole, but more immediately it establishes the credibility of specific elements of the assemblage. The hypothetico-deductive account of confirmation and testing that was vigorously advocated by New Archaeologists and still influences programmatic debate in archaeology captures little of what matters in this process of building, refining, manipulating, and assessing explanatory models in archaeology. When these models are compelling, their credibility arises from mutually constraining and reinforcing relations among subsidiary models rather than from any one self-warranting epistemic foundation.

## References

- 46.1 A. Wylie: *Thinking from Things: Essays in the Philosophy of Archaeology* (Univ. California Press, Berkeley 2002)
- 46.2 T.A. Kohler, S.E. van der Leeuw: Introduction: Historical socio-natural systems and models. In: *The Model-Based Archaeology of Socio-natural Systems*, ed. by T.A. Kohler, S.E. van der Leeuw (SAR, Santa Fe 2007) pp. 1–12
- 46.3 M. Morrison, M.S. Morgan: Models as mediating instruments. In: *Models as Mediators: Perspectives on Natural and Social Science*, ed. by M.S. Morgan, M. Morrison (Cambridge Univ. Press, Cambridge 1999) pp. 10–38
- 46.4 M.S. Morgan: *The World in the Model: How Economists Work and Think* (Univ. Cambridge Press, Cambridge 2012)
- 46.5 M. Weisberg: *Simulation and Similarity: Using Models to Understand the World* (Oxford Univ. Press, Oxford 2013)
- 46.6 R. Harré: *The Principles of Scientific Thinking* (Univ. Chicago Press, Chicago 1970)
- 46.7 M. Hesse: *Models and Analogies in Science* (Notre Dame Univ. Press, Notre Dame 1970)
- 46.8 M. Hesse: *The Structure of Scientific Inference* (Macmillan, London 1974)
- 46.9 D.L. Clarke: Models and paradigms in contemporary archaeology. In: *Models in Archaeology*, ed. by D.L. Clarke (Methuen, London 1972) pp. 1–60
- 46.10 D.L. Clarke (Ed.): *Models in Archaeology* (Routledge, London 1972)
- 46.11 R. Levins: The strategy of model building in population biology, *Am. Sci.* **54**(4), 421–431 (1966)
- 46.12 K.V. Flannery: Cultural history versus cultural process: A debate in American archaeology, *Sci. Am.* **217**(1), 119–122 (1967)
- 46.13 J.A. Sabloff: *Archaeology Matters: Action Archaeology in the Modern World* (Left Coast, Walnut Creek 2008)
- 46.14 J.E. Doran, F.R. Hodson: *Mathematics and Computers in Archaeology* (Edinburgh Univ. Press, Edinburgh 1975)
- 46.15 M. Aldenderfer: The analytical engine: Computer simulation and archaeological research, *Archaeol. Method Theory* **3**, 195–247 (1991)
- 46.16 V.G. Childe: *The Danube in Prehistory* (Oxford Univ. Press, Oxford 1929)
- 46.17 G.R. Willey, P. Phillips: *Method and Theory in American Archaeology* (Univ. Chicago Press, Chicago 1958)
- 46.18 V.G. Childe: *The Dawn of European Civilization*, 6th edn. (Kegan, Paul, London 1957)

- 46.19 J.O. Brew: The use and abuse of taxonomy. In: *The Archaeology of Alkali Ridge Utah*, ed. by J.O. Brew (Harvard Univ. Press, Cambridge 1946) pp. 44–66
- 46.20 J.A. Ford: Comment on A.C. Spaulding, "Statistical techniques for the discovery of artifact types, *Am. Antiq.* **19**(4), 390–391 (1954)
- 46.21 A.C. Spaulding: Review of measurements of some prehistoric design developments in the southeastern states, by J.A. Ford, *Am. Anthropol.* **55**, 588–591 (1953)
- 46.22 A.C. Spaulding: Statistical techniques for the discovery of artifact types, *Am. Antiq.* **18**(4), 305–313 (1953)
- 46.23 W.Y. Adams, E.W. Adams: *Archaeological Typology and Practical Reality. A Dialectical Approach to Artifact Classification and Sorting* (Cambridge Univ. Press, Cambridge 2008)
- 46.24 J.A. Ford: On the concept of types, *Am. Anthropol.* **56**, 42–57 (1954)
- 46.25 J.N. Hill: Broken K Pueblo: Patterns of form and function. In: *New Perspectives in Archaeology*, ed. by L.R. Binford, S.R. Binford (Univ. Arizona Press, Tucson 1968) pp. 103–142
- 46.26 H. Glassie: *Middle Virginian Folk Housing* (Univ. Tennessee Press, Knoxville 1975)
- 46.27 C. Renfrew, S. Shennan (Eds.): *Ranking, Resource and Exchange: Aspects of the Archaeology of Early European Society* (Cambridge Univ. Press, Cambridge 1982)
- 46.28 M. Parker Pearson, Ramilisonina: Stonehenge for the ancestors: The stones pass on the message, *Antiquity* **72**, 308–326 (1998)
- 46.29 A. Whittle: Remembered and imagined belongings: Stonehenge in its traditions and structures of meanings, *Proc. Br. Acad.* **92**, 145–166 (1997)
- 46.30 M. Llobera: Working the digital: Some thoughts from landscape archaeology. In: *Material Evidence: Learning from Archaeological Practice*, ed. by R. Chapman, A. Wylie (Routledge, London 2015) pp. 173–188
- 46.31 P. Verhagen, T.G. Whitley: Integrating archaeological theory and predictive modeling: A live report from the scene, *J. Archaeol. Method Theory* **19**(1), 49–100 (2012)
- 46.32 A.L. Kroeber: On the principle of order in civilization as exemplified by changes of fashion, *Am. Anthropol.* **21**(3), 235–263 (1919)
- 46.33 J.F. Deetz, E.S. Dethlefsen: Death's head, cherub, urn and willow, *Nat. Hist.* **76**, 29–37 (1967)
- 46.34 S.W. Manning: Radiocarbon dating and archaeology: History, progress and present status. In: *Material Evidence: Learning from Archaeological Practice*, ed. by R. Chapman, A. Wylie (Routledge, London 2015) pp. 128–158
- 46.35 A.L. Boozer: The tyranny of typologies: Evidential reasoning in Romano-Egyptian domestic archaeology. In: *Material Evidence: Learning from Archaeological Practice*, ed. by R. Chapman, A. Wylie (Routledge, London 2015) pp. 92–109
- 46.36 M.B. Schiffer: *Formation Processes of the Archaeological Record* (Univ. New Mexico, Albuquerque 1987)
- 46.37 W.A. Longacre, T.R. Hermes: Rice farming and pottery production among the Kalinga: New ethnoarchaeological data from the Philippines, *J. Anthropol. Archaeol.* **38**, 35–45 (2015)
- 46.38 V.T. Holliday: *Soils in Archaeological Research* (Oxford Univ. Press, Oxford 2004)
- 46.39 J.K. Stein: Earthworm activity: A source of potential disturbance of archaeological sediments, *Am. Antiq.* **48**(2), 277–289 (1983)
- 46.40 M. Bell: Experimental archaeology at the crossroads: A contribution to interpretation or evidence of 'Xeroxing'? In: *Material Evidence: Learning from Archaeological Practice*, ed. by R. Chapman, A. Wylie (Routledge, London 2015) pp. 42–58
- 46.41 A. Bogaard: *Neolithic Farming in Central Europe: An Archaeobotanical Study of Crop Husbandry Practices* (Routledge, New York 2004)
- 46.42 A. Bogaard: Lessons from modeling neolithic farming practice: Methods of elimination. In: *Material Evidence: Learning from Archaeological Practice*, ed. by R. Chapman, A. Wylie (Routledge, London 2015) pp. 243–254
- 46.43 M. Pollard, P. Bray: The archaeological bazaar: Scientific methods for sale? Or: 'putting the "arche-" back into archaeometry. In: *Material Evidence: Learning from Archaeological Practice*, ed. by R. Chapman, A. Wylie (Routledge, London 2015) pp. 113–127
- 46.44 A. Bayliss, A. Whittle: Uncertain on principle: Combining lines of evidence to create chronologies. In: *Material Evidence: Learning from Archaeological Practice*, ed. by R. Chapman, A. Wylie (Routledge, London 2015) pp. 213–242
- 46.45 H. Eckardt, C. Chenery, P. Booth, J.A. Evans, A. Lamb, G. Müldner: Oxygen and strontium isotope evidence for mobility in Roman Winchester, *J. Archaeol. Sci.* **36**, 2816–2825 (2009)
- 46.46 S. Leach, M. Lewis, C. Chenery, G. Müldner, H. Eckardt: Migration and diversity in Roman Britain: A multidisciplinary approach to the identification of immigrants in Roman York, England. *Am. J. Phys. Anthropol.* **140**(3), 546–561 (2009), doi: 10.1002/ajpa.21104
- 46.47 C. Shelley: Multiple analogies in archaeology, *Philos. Sci.* **66**(4), 579–605 (1999)
- 46.48 A. Wylie: The reaction against analogy, *Adv. Archaeol. Method Theory* **8**, 63–111 (1985)
- 46.49 A. Wylie: Critical distance: Stabilising evidential claims in archaeology. In: *Evidence, Inference and Enquiry*, Proceedings of the British Academy, Vol. 171, ed. by P. Dawid, W. Twining, M. Vasiliki (Oxford Univ. Press, London 2011) pp. 371–394
- 46.50 R. Chapman, A. Wylie: *Evidential Reasoning in Archaeology* (Bloomsbury, London 2016)
- 46.51 D.H. Thomas: A computer simulation of Great Basin Shoshonean subsistence and settlement patterns. In: *Models in Archaeology*, ed. by D.L. Clarke (Methuen, London 1972) pp. 671–703
- 46.52 J.H. Steward: *Basin-Plateau Aboriginal Sociopolitical Groups*, Vol. 120 (Bureau of American Ethnology, Washington DC 1938)

- 46.53 K.V. Flannery, R.G. Reynolds: Simulating foraging and early agriculture in Oaxaca. In: *Gila Naquitz: Archaic Foraging and Early Agriculture in Oaxaca, Mexico*, ed. by K.V. Flannery (Academic, New York 1986) pp. 433–508
- 46.54 M.S. Morgan, M. Morrison (Eds.): *Models as Mediators: Perspectives on Natural and Social Science* (Cambridge Univ. Press, Cambridge 1999)
- 46.55 J.F. O'Connell, J. Allen: The restaurant at the end of the universe: Modelling the colonisation of Sahul, *Aust. Archaeol.* **74**, 5–16 (2012)
- 46.56 T.A. Kohler, S.E. van der Leeuw (Eds.): *The Model-Based Archaeology of Socionatural Systems* (SAR, Santa Fe 2007)
- 46.57 T.A. Kohler, C.D. Johnson, M. Varien, S. Ortman, R.G. Reynolds, Z. Kobti, J. Cowan, K. Kolm, S. Smith, L. Yap: Settlement ecodynamics in the prehispanic central mesa verde region. In: *The Model-Based Archaeology of Socionatural Systems*, ed. by T.A. Kohler, S.E. van der Leeuw (SAR, Santa Fe 2007) pp. 61–104
- 46.58 T.J. Wilkinson, M. Gibson, J.H. Christiansen, M. Widell, D. Schloen, N. Kouchoukos, C. Woods, J. Sanders, K.-L. Simunich, M. Altaweel, J.A. Ur, C. Hritz, J. Lauinger, T. Paulette, J. Tenney: Modeling settlement systems in a dynamic environment: Case studies from Mesopotamia. In: *The Model-Based Archaeology of Socionatural Systems*, ed. by T.A. Kohler, S.E. van der Leeuw (SAR, Santa Fe 2007) pp. 175–208
- 46.59 M. Hegmon: The risks of sharing and sharing as risk reduction: Interhousehold food sharing in egalitarian societies. In: *Between Bands and States*, ed. by S.A. Gregg (Southern Illinois Univ. Press, Carbondale 1991) pp. 309–329
- 46.60 I.G. Robertson: Sharing, debt, and incipient inequality in small-scale agricultural economies: A computer simulation, *Proc. 62nd Annu. Meet. Soc. Am. Archaeol.*, Nashville (1997)

# Models and Ideology

## 47. Models and Ideology in Design

Cameron Shelley

Models play a number of roles in design. Models may assist designers in the solution of technical problems. In addition, models may assist designers in achieving ideological goals. Ideological goals of designers could include respect for cultural norms, such as the distinction between masculine and feminine, or adherence to a design paradigm, such as modernism. In this latter role, design models could be compared to model citizens, that is, community members of exemplary character. Use of such models helps designers to produce solutions that fit with the prevailing norms of good design and to promote the standards of design paradigms. For example, the Ville Savoye house was designed by Le Corbusier using ships as models both to solve technical problems of accommodation but also to visibly promote the modernist design paradigm. The purpose of this chapter is to review examples of models that serve this last ideological function. Design ideologies reviewed include revivalism, modernism, industrial design,

47.1	<b>Design and Ideology</b> .....	1003
47.2	<b>Models and Ideology</b> .....	1004
47.3	<b>Revivalism: Looking to the Past</b> .....	1005
47.4	<b>Modernism: Transcending History</b> .....	1006
47.5	<b>Industrial Design: The Shape of Things to Come</b> .....	1009
47.6	<b>Biomimicry</b> .....	1011
47.7	<b>Conclusion</b> .....	1013
	<b>References</b> .....	1013

and biomimicry. Each of these paradigms is characterized by a set of values that designers seek to reflect and promote through their works. There is no finite or canonical list of design ideologies but this set is widely known and acknowledged. So, these examples illustrate how models may serve ideological functions in various design disciplines.

### 47.1 Design and Ideology

In his history of modern design, *Forty* [47.1] considers several explanations for the increasing differentiation in design of consumer products from the nineteenth century forwards. For example, product design increasingly varied according to the gender of the intended purchaser. This trend was manifested in various articles, including wristwatches [47.1, p. 65]:

“In wristwatches, the disparity in size between those for gentlemen and those for ladies exceeded that between male and female wrists, and a lady’s watch usually had more delicate features and face. Being smaller, ladies’ watches have generally been more expensive, but when they can be compared to men’s watches of a similar price, the ladies’ models are still more ornamented. In the 1907 Army and

Navy Stores catalogue, the men’s watches were all calibrated with Roman numerals, while the ladies’ watches all had Arabic numerals, whose form – curvilinear rather than angular – may be judged more delicate.”

These differences in wristwatch design have little to do with any physical differences between the arms and eyes of men and women and more to do with differences in their social roles. As the social spheres of men and women diverged, so did prevalent ideals of the masculine and feminine. These ideals then came to be embodied in the design of consumer goods that, in turn, served as confirmation that the ideals reflected an objective reality. In other words, as *Forty* [47.1, p. 66] points out, the design of objects such as men’s and

ladies' wristwatches reflects an ideology in which *masculine* versus *feminine* is a crucial distinction. Design, along with fiction, education, and religion, helped to make this ideology manifest, thus both making it visible and material and also reinforcing its veracity.

## 47.2 Models and Ideology

Given the importance of ideology in accounting for designs, there remains the issue of how ideology exerts its influence. One way in which ideology participates in design is through the use of models. Designers frequently rely on models in order to address design problems.

For example, in the original version of the Calculator app for its iPod and iPhone, Apple designed the interface to recall the signature Braun ET-44 calculator created by legendary designer Dieter Rams [47.2]. The rounded buttons with convex, raised centers were a distinctive design element of the original calculator. Even though the flat, touch-sensitive surface of the iPhone could not accommodate raised buttons, the crisp and candy-like appearance made the interface attractive and inviting.

Apple's designers had at least two reasons for using the ET-66 as a model. First, it was a borrowing from a proven design, aimed to make the calculator app straightforward and pleasant for users, regardless of their familiarity with the original. Second, it was an homage to a design classic; Steve Jobs' admiration for Rams is well known [47.3]. By imitating the ET-66, Apple's designers connected their work to an honored design tradition. In brief, Rams' ET-66 stood as a model for the Calculator app in two respects, as a guide to resolving an interface challenge and as an assurance of right-mindedness of their design philosophy.

The term *model* itself exhibits an ambiguity between these two senses:

1. A model can be a *model solution*, that is, something that exhibits how a particular problem can be solved correctly or optimally.
2. A model can be a *model citizen*, that is, someone who embodies right thinking and good conduct and is thus worthy of emulation.

A model solution provides a solution to a problem that can be applied analogically to other problems of

The role of design in both manifesting and reinforcing ideology is demonstrated in other social conceptions as well, such as *juvenile* versus *adult*, *middle class* versus *working class*, and *master* versus *servant* [47.1, p. 67ff].

a similar nature. A model citizen embodies the values of a community and inspires others to act in accord with those same values.

Rams' calculator served as a model for the Calculator app in both senses. It saved Apple designers the trouble of trying to design a calculator interface from first principles. It also provided them with a standard of design excellence that their own work could strive to live up to.

The role of models as model citizens in design more clearly reflects their role in the ideology of design. That is, model citizens represent an ideal of social standards and behavior that designs tend to reify and reflect. So, to further understand the place of ideology and models in design, we must examine models in design as model citizens.

To explore this subject further, it will be convenient to look at design as practiced in a number of ideologically informed, modern design movements. Each movement exemplifies a different ideology and illustrates how models have been used as model citizens in design. The design movements discussed are as follows:

1. *Revivalism*: The Gothic Revival of the nineteenth century looked to the European Middle Ages for models of good design.
2. *Modernism*: The Modernist movement of the early twentieth century looked to contemporary heavy industry for models of good design.
3. *Industrial design*: Consultant designers of the mid-twentieth century looked to forecasts of future industry for models of good design.
4. *Biomimicry*: Engineers in the later twentieth century looked to biological organisms for models of good design.

In each case, the ideological aspects of models, in the sense used by *Forty* [47.1], are explained and discussed. From this examination, the importance of ideology in model selection will become more clear.



## 47.3 Revivalism: Looking to the Past

In general, revivalism refers to the use of elements from an historical design style in contemporary designs. One of the best-known revivalist design movements was the Gothic Revival, which reached its height in Britain in the nineteenth century. The Gothic Revival began in the eighteenth century, after the antiquarian Horace Walpole wrote the medieval romance, the *Castle of Otranto* in 1764, and another book about his old house, Strawberry Hill in 1774. The latter book described how Walpole renovated this house with elements from medieval buildings, such as pointed arches, crockets, quatrefoils, and so on. This book helped to raise general interest in the architecture and design from medieval Europe [47.4].

In the years after the Napoleonic Wars, the government Church Building Commission undertook to subsidize the construction of hundreds of churches throughout Britain. The general idea was to knit back together the social fabric that had unraveled somewhat under the pressures of the prolonged and ideologically charged conflict. Many of these churches were built in a Gothic style. That is, they often applied design elements from Gothic structures, in the spirit of Walpole's version of decoration [47.5]. See the Gothic Revival St. Peter's church, built under the auspices of the Commission in Fig. 47.1 for an example.

Some architects criticized the Commission's approach to architecture. Among them was Augustus Welby Northmore Pugin (1812–1852), most famous for his work on the Palace of Westminster in the 1840s. Pugin objected that, oftentimes, Gothic elements such

as pointed windows and buttresses were simply tacked on to structures that were essentially classical in design, having the basic form of Greek temples. The resulting hybrid structures were not truly Gothic at all in his eyes. Compare St. Peter's to the neoclassical St. John's church, built under the auspices of the Commission, in Fig. 47.2. Note how St. Peter's imitates the roof profile and basic layout of the neoclassical model, except with Gothic-style buttresses and accoutrements in place of classical ones.

Pugin had formed a strong attachment to medieval buildings and churches in particular. He had converted to Catholicism in 1834, in part as a result of his experiences studying medieval churches in England and northern France. He saw a strong connection between medieval architecture and proper, Christian faith. As his biographer puts it, for Pugin, “the Catholic Church is the true church, Gothic architecture its revealed form, true in the sense of absolute, a divine, revealed form” [47.6].

For Pugin, the ideology of the Gothic Revival contained at least two social ideals, those being *authenticity* and *conservatism*. Being authentic did not mean, of course, that a building had to be a genuine medieval structure. Instead, it meant that a building should be a close facsimile of genuine buildings of the earlier era. Studies of genuine structures in England and northern France served as models that could guide the revivalist architect in this matter.



**Fig. 47.1** St. Peter's Church, Blackley, UK. Designed by E. H. Shellard, ca. 1845. The church has the basic form of a Greek temple but is dressed up with Gothic features such as pointed windows and superfluous buttresses. Photo by David Dixon



**Fig. 47.2** St. John's church, London, UK. Designed by Francis Octavius Bedford ca. 1824. Photo by The Voice of Hassocks

An important aspect of authenticity, then, was *localism*. That is, revivalist structures should use local materials and building methods that imitated the methods used in the Middle Ages by builders in the vicinity where the new building was to stand. In the Middle Ages, the transportation infrastructure of northern Europe was neither efficient nor developed enough to allow for the shipment of large volumes of materials over long distances. As a result, medieval buildings tended to be made of materials acquired in the local area. Similarly, poor infrastructure meant that the builders hired to construct buildings were also recruited from the local area. This situation facilitated the existence of local idioms in design. That is, each region tended to see the rise of design traditions within the region and different from the traditions that arose in other regions. For authenticity in a building to be situated in a given region, Pugin thought it best to observe the local building traditions that characterized genuine, medieval buildings nearby.

From an instrumental standpoint, this emphasis on localism was not always optimal. In the nineteenth century, it would often be more economical to ship materials and workmen from other areas of the country by rail or canal. Pugin's emphasis on localism was motivated on ideological and not instrumental grounds.

Beyond authenticity, Pugin's careful imitation of medieval design was motivated by religious conservatism. Contemporary Protestant churches tended to be spatially simple in the sense that their interiors were relatively undifferentiated spaces in which the congregation and minister gathered together. A Catholic church from the Middle Ages was a microcosm of the medieval worldview, a hierarchical arrangement of separate spaces, each with its own appropriate functions and occupants. Pugin's church designs persisted in this traditional divided arrangement of spaces. For example, Pugin's church designs usually contained a *rood screen* to separate the nave from the chancel, thus keeping the altar and choir apart from the congregation. This

separation had an important function in the medieval ceremony of mass but had fallen out of favor among Anglicans and Catholics by the nineteenth century. Pugin designed the screens for his churches as a way of advocating for the return to medieval forms of worship, which he regarded as worthier. A controversy ensued that was settled, in the end, by an appeal to the Vatican, which sided with Pugin's opponents. As a result, many of the rood screens in his churches were subsequently removed [47.7].

So, authenticity and religious conservatism were important social ideals in Pugin's version of the Gothic Revival. First, the continued presence of medieval Gothic architecture in the country supported the view that such architecture was an authentic expression of Englishness. Study of regional variations in medieval Gothic architecture in England only served to reinforce its authenticity. By reviving the Gothic style, then, Pugin was not introducing a foreign element into English life.

Second, Pugin could point to the models as evidence of the Englishness of the Catholic Church. Britain was officially Anglican and had only recently passed a law tolerating Catholicism and allowing Catholics to hold public offices. Pugin hoped to turn this tolerance into broader acceptance of Catholicism, even reconciliation with the Anglican Church. People's attachment to these medieval buildings, reinforced through Pugin's own works, might lead them to reconsider their separation from the Church of Rome.

The Gothic Revival shows how models can play a role in the ideological side of design. Design is undertaken not merely to solve a given problem but to reflect a worldview. In the Gothic Revival, historical authenticity and religious propriety were dominant ideological values that informed design. Practitioners of the Gothic Revival, such as Pugin, naturally looked to surviving instances of Gothic architecture as models of solutions to technical design problems and also as embodiments of their social ideals.

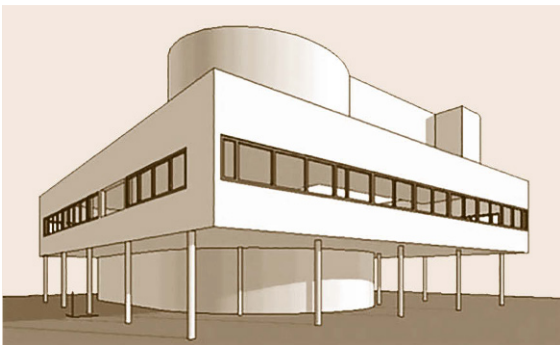
## 47.4 Modernism: Transcending History

Historical structures are an obvious source of models within a design ideology that seeks to reinstate or reinvigorate past modes of living. By the outset of the twentieth century, some designers had decided that revivalism was not a tenable ideology. Technological advances had brought with them new challenges and those challenges called for new approaches. Thus, appropriate design ideology had to be divorced from attachments to the past.

The designer who perhaps most embodies this form of modernism is Le Corbusier. Born as Charles Edouard Jeanneret (1887–1965) in Switzerland, Le Corbusier trained in engraving but made his mark in architecture and urban planning. Le Corbusier set up his practice in Paris towards the end of World War I and advocated modernist housing as a means of quickly rebuilding the housing stock destroyed in the war. He published a collection of essays, *Vers une*

architecture [47.8], discussing his views on architecture.

In this book, Le Corbusier famously promoted his view that the design of contemporary industrial objects formed the best source of models for architecture. In particular, he described a house as “a machine for living”, meaning that houses should be designed just like cars, boats, or other mass-produced objects. His most famous house, the Villa Savoye built in 1928–1931 near Paris, was designed in this manner (Fig. 47.3). It was made of modern materials, reinforced concrete, steel, and glass, with no attempt at disguise or decoration. The use of steel and concrete structure meant that there was no need for load-bearing walls, so interior walls were eliminated or minimized. Thus, the interior was largely open space. Each of the four exterior walls contained long ribbon windows, reducing the distinction between inside and outside. The roof was flat, not peaked, and contained a garden for the occupants to use, rather like the recreation deck on an ocean liner. In fact, the design of the house was inspired by the design of the decks of ocean liners such as RMS Aquitania, a picture of which was featured in Le Corbusier’s book (Fig. 47.4).



**Fig. 47.3** La Ville Savoye, Poissy, France. Designed by Le Corbusier ca. 1930



**Fig. 47.4** Deck of the RMS Aquitania. Detail of photo by Library of Congress

In urban planning, Le Corbusier’s early ideas are conveyed by the *Ville Contemporaine* exhibit that he mounted at the Salon d’Automne in Paris in 1922 [47.9]. This ideal city contained a central district consisting of glass skyscrapers and apartment buildings arranged in a grid pattern. Connecting the buildings with each other and the surrounding countryside were a network of highways. At the center where the highways met would be a seven-level transportation hub including levels for railways, highways, subways, with an airport on the top layer. The buildings were to be raised off the ground on stilts so that the ground level of the entire city could be a large green space.

Each component of the design is dedicated to the fulfillment of a single function. The buildings in the central core were for white-collar workers who would run the city. Blue-collar workers and industrial facilities would be placed in areas outside the central district. Some buildings in the core were for work, others for housing, others for commerce. Some spaces were dedicated to transportation, for example highways and airports, whereas others were dedicated to recreation, for example rooftops and the ground-floor green space. In order to transition from one activity to another, a person would drive a car along the highway to the space designed for that activity.

This initial design was generic. However, Le Corbusier later exhibited a design of this type specifically for Paris at the *Exposition des Arts Decoratifs* in 1925. It was called the *Voisin* plan after the automobile and aircraft manufacturer that sponsored Le Corbusier at the event. In this plan, Le Corbusier suggested razing the central district of Paris and replacing it with a grid of eighteen sixty-story skyscrapers connected by highways. Important monuments like Notre Dame would be retained but, otherwise, the core of the city was to be entirely rebuilt. The traditional but chaotic layout of Paris, with its maze of medieval streets, was to be eliminated in favor a functional and efficient grid of massive buildings and roads.

Le Corbusier’s ideas become influential particularly after World War II, when building and rebuilding projects took off all over the industrialized world. Highways were built, old neighborhoods were bulldozed to make way for expressways and apartment blocks, and the downtowns of big cities filled with slabs of concrete, glass, and steel. Note the resemblance of the Co-op City buildings in Fig. 47.5, constructed ca. 1970, to those envisioned by Le Corbusier.

A great virtue of modernist design is that it made efficient use of industrial materials, allowing infrastructure to be built rapidly and affordably. A great problem of the modernist approach is that it could be overwhelming and inflexible, treating people somewhat as



**Fig. 47.5** Co-op City, the Bronx, USA. Photo by Jules Antonio

goods to be stored and moved about as required by the design of their infrastructure [47.10].

The main social ideals of Le Corbusier's modernist view are *functionalism* and *conformity*. Functionalism identifies the analytic approach to design adopted by the modernists, embodied in the expression *form follows function*. In brief, design should be approached from perspective independent of history. To a functionalist designer, it is of no relevance how buildings or anything else were designed in antiquity or the Middle Ages. All that matters is the problem to be solved and the means available to solve it with. Le Corbusier took a Platonic view on which the activities of living and the architectural forms for building are regarded as a set of timeless forms. His buildings exhibited a preference for simple geometric figures and solids in combination.

In terms of analysis, modernist designers sought to disaggregate the various activities that would occur in the use of the design. When designing a house, or machine for living, the process of living should be broken down into its subprocesses and some space designated for each. The remaining task was to place the spaces for each activity in the correct relation to each other. In the Villa Savoye, the ground floor contained spaces for auxiliary functions, including the entrance, the garage, and rooms for chauffeurs and maids. The main floor contained the bedrooms for sleeping, the kitchen for cooking, and the salon for interacting. The roof contained the garden for relaxation and recreation. In the Voisin plan, activities such as resting, recreating, working, and moving were each assigned a separate space. Highways were used to allow people to move from one activity to another with as little hindrance as possible.

Besides disaggregation, functionalism also implies a kind of universalism. Just as the relevance of historical antecedents is minimized, so is the relevance of region-

alism. Functionalist design tends to focus on the basics. Living in a house, for example, is analyzed to an almost biological level: eating, sleeping, exercising, interacting. These functions are universal ideals that apply to all people. In contrast, cultural preferences, such as having a porch or enlarged foyer to lend importance to the front entrance, are minor considerations.

In addition, another important component of functionalism is honesty [47.11]. That is, the aesthetic value of a modernist design should come from its construction instead of from add-on decorations that serve no basic purpose. In the Victorian era, designers might use iron as a structural element of Gothic buildings, as Pugin did for the Palace of Westminster. However, the iron would be hidden from view. From a modernist perspective, this practice is dishonest. A building made of steel, glass and concrete should display these materials. Furthermore, its construction should be such that nothing further is required to make the building look good.

Besides functionalism as such, modernist designs also tend to require *conformity* from their users. A modernist building, for example, does not invite later modification. If the building's form resulted from a correct application of timeless ideals, then no modifications should be necessary. In the Voisin plan, residents were not to be invited to customize their spaces to suit themselves. In some cases, this attribute of modernist design may be put down to the narcissism of the designers. In many cases, however, conformity was an implication of the modernist view of industrial production as the ultimate state of civilization. The efficiencies of industrial design and production would drive out suboptimal or idiosyncratic architecture, replacing it with universal design. In that event, people would have to accommodate themselves to their designed surroundings, rather than the reverse. This view is admirable in its egalitarianism, that everyone should enjoy the same, amenable standard of living. However, it is also objectionable in the sense that it aims to achieve this end through imposition of a rigid mode of life.

In any event, models were important to the development of modernist design. The influence of ships, in particular, on Le Corbusier was noted above. The ability of passenger ships like the Aquitania to efficiently accommodate hundreds of people on long voyages clearly impressed him. In his architecture, he sought to apply the lessons of ship design, as he saw them, to the design of houses and even cities. The efficiency with which modernist designs could provide accommodation for large number of people, using modern, mass-produced materials such as concrete, steel, and glass, made it a highly suitable building regime for the post-World War II construction boom.

Of course, modernist models had their limits. After all, ships (and cars and airplanes, other sources much admired by Le Corbusier and contemporary modernists) are not themselves machines for living but machines for transportation. The arrangements made in such vehicles for people to travel in them temporarily are not necessarily appropriate for structures where people expect to live permanently. The conformity required in arrangements for air travel, for example, is more of an imposition in a house or a neighborhood.

However, such limitations might be overlooked because industrial models also served their ideological function. That is, living in industrial surroundings would accommodate people to the industrialized world that they inhabited. In his way, then, Le Corbusier was just as concerned as Pugin for the authenticity of his architecture. However, cars and ocean liners served him as guarantors of authenticity and the good life in place of the medieval cathedrals favored by his predecessor.

## 47.5 Industrial Design: The Shape of Things to Come

Besides looking to the past for inspiration or to the present, it is also conceivable to look to the future for models suitable for the purposes of good design. Although such a perspective may sound paradoxical or even impossible, it was characteristic of a third design movement that I wish to examine, that being industrial design.

Strictly speaking, industrial design is not a movement but a profession. It arose as a result of the industrial revolution and the mass production of goods. Before the revolution, household goods were typically produced by craftsmen who worked in local design traditions, producing a given item from raw materials. With industrialization, craftsmen were replaced by semiskilled laborers who did not participate in the design process. Goods were designed either in imitation of previous craft traditions, or they were designed by their inventors. By the early twentieth century, both these approaches had proven inadequate for the novel technologies that were being mass produced. A group of professionals arose whose occupation was giving proper form to these new technologies. These were the industrial designers.

In spite of the fact that industrial design was, and continues to be, a profession, its first practitioners shared a set of social ideals that informed their work. Thus, the profession also constituted a design movement with a characteristic ideology. Perhaps the key values of this ideology were *progress* and *consumerism*.

The value of progress in industrial design was most clearly captured by Raymond Loewy, perhaps the most famous industrial designer of his era. Loewy (1893–1986) was born in France and received an education in technology in a preparatory school in Paris. He served in the French Army Corps of Engineers during the World War I. After the war, he emigrated to the United States where he made a living applying his artistic talents as a window dresser for New York department stores and as an illustrator for fashion magazines. His

first break in industrial design came with the commission to redesign the Gestetner duplicating machine in 1929. Afterwards, Loewy established a successful design consultancy and participated in the design of a variety of industrial objects, from cigarette packages to cars, locomotives, and refrigerators.

In his autobiography, *Loewy* tries to capture some of the lessons he had learned over the course of his career [47.12]. One of the key lessons is embodied in what he calls the MAYA principle. *MAYA* is an acronym for the phrase *most advanced yet acceptable*. In his view, a well-designed product should appear to its users to be technologically advanced but also comfortably familiar. Loewy had observed a tension in people's minds about what they expect from the things they use: On the one hand, people expect technology to improve over time, so that a newer product should outperform older ones. As a result, they expect the design of their gear to change over time. On the other hand, people like to stick with what they know or are used to. Thus, change in design can be discomforting or unwelcome. The *MAYA* principle suggests that industrial design has to balance people's expectation of innovation with their need for stability.

Consider Loewy's redesign of the Gestetner mimeograph machine [47.13]. The Gestetner was an industrial contraption with an exposed mechanism and perched on an ungainly metal frame. Loewy enclosed the mechanism in a streamlined case and streamlined the machine's appearance and footprint. By enclosing the machine's workings, Loewy made it less dangerous, for example, the user's tie and fingers were not likely to get caught in its gears and its toner was less likely to splash the user's skin and clothing. The new appearance also made the machine more approachable.

Loewy's redesign of the Gestetner provides a good illustration of the *MAYA* principle. The new design was advanced in the sense that it brought the productivity of an industrial device into the office space. Beforehand,

the ungainly and mechanical look of the Gestetner had caused users to categorize it as industrial equipment. Thus, it was treated like a furnace or a boiler and hidden away from the office spaces where its duplicating function was most useful. Afterward, by making the Gestetner look and feel much like a file cabinet, Loewy caused office managers to think of it as a piece of office furniture, to be kept in the workplace itself. Thus, the new design was advanced in the sense that it brought industrial productivity to the office, and acceptable in the sense that it looked right at home next to the file cabinets and desks already situated there. After the introduction of the new design, the Gestetner sold well as a piece of office furniture.

The MAYA principle illustrates the importance of progress to industrial design of this era. Advancement, on this principle, is an indispensable part of the design of new goods. That is, one of the jobs of a good designer is to provide customers with goods that will outperform previous designs, thus making the work of customers more productive, and their lives more pleasant. As a practical matter, the MAYA principle also instructs us that progress is best served up in moderate doses. This view stands somewhat in contrast with the view among some current designers that advancement should come in the form of game-changing or disruptive designs.

Consumerism is also part of this picture, although not one that is explicitly noted in the MAYA principle. If there is to be advancement in design, then existing designs must become obsolete. New designs can make old ones obsolete in at least two ways. First, new designs may perform a given job better than old designs. A new engine, for example, may burn fuel more efficiently than older designs. Second, new designs may appeal to people more than old designs. The practice of changing the style of cars each year provides a good example: people may get rid of an old car in favor of a new one not because the new one is technically superior but because the appearance of the new car makes them feel unhappy about the appearance of the old one. This mental phenomenon is known as *psychological obsolescence* [47.14].

One example of how industrial design could be applied to psychological obsolescence is provided by Loewy. One of Loewy's best-known designs was a streamlined pencil sharpener. The sharpener was designed in the shape of an aerodynamic tear drop, with the hole for insertion of the pencil tip at the round end and the handle to turn the mechanism at the pointed end. *William Gibson* describes the sharpener as follows [47.15]:



**Fig. 47.6** Petipoint streamlined iron, made by the Weverly Tool Co. of Sandusky, Ohio, ca. 1941. Detail of photo by Tomislav Medak

“The Thirties had seen the first generation of American industrial designers; until the Thirties, all pencil sharpeners had looked like pencil sharpeners; your basic Victorian mechanism, perhaps with a curlicue of decorative trim. After the advent of the designers, some pencil sharpeners looked as though they'd been put together in wind tunnels. For the most part, the change was only skin-deep; under the streamlined chrome shell, you'd find the same Victorian mechanism. Which made a certain kind of sense, because the most successful American designers had been recruited from the ranks of Broadway theater designers. It was all a stage set, a series of elaborate props for playing at living in the future.”

The point is that the mechanism of the sharpener has not changed. Loewy has simply made the casing more up-to-date.

This application of industrial design encourages consumerism in the sense that it invites users to confuse technological innovation with stylistic innovation. In the case of the pencil sharpener, this confusion could lead users to dispose of their existing goods in order to purchase new ones that do not sharpen pencils any better.

*Gibson* also observes that industrial design of the era allowed people to *play at* living in the future. This point is key to see how the use of models fits into this version of industrial design. Designers like *Loewy* could not, of course, actually see into the future and take from there the models they needed for the present. They could, however, take current trends in technology and extrapolate them. One trend they could extrapolate involved *streamlining*, that is, the use of aerodynamic shapes. Industrial designers of that era felt that air travel was the transportation of the future for all [47.16]. Thus, industrial designers took existing aircraft as models, imagined how they would look in the future, and then

applied these ideas to the design of various contemporary goods, even such slow-moving objects as pencil sharpeners or irons [47.17] (Fig. 47.6). Thus it was that industrial designers could look to the future, as it were, for models to apply to contemporary design problems.

As with revivalists or modernists, industrial designers of the early twentieth century used models in order to address design problems. Their models were selected not merely for their ability to answer questions of function but because they embodied the ideals of the movement. Central to the ideology of that movement were the ideals of progress and consumerism.

## 47.6 Biomimicry

Designers typically look to artifacts to find models for their work. Of course, it makes sense to seek one artifact to be a model for another. Yet, the natural world also provides models for designers. On the face of it, this observation seems odd since the natural world is not an artifact. However, the process of evolution has produced organic forms that can be treated as designs, that is, as solutions to problems posed by the environment. For example, many early airplane designers looked to birds and other flying organisms in order to work out their designs [47.18]. As is the case with artifacts, organisms such as birds can serve as model solutions but also as model citizens.

In fact, several design movements have taken organisms as models in different ways. In art, the Art Nouveau movement often used plants as models for its curvaceous forms. Designers interested in sustainability often look to organic systems for inspiration, such as the *Cradle-to-Cradle* paradigm espoused by *McDonough* and *Braungart* [47.19]. Others adopt organic models based on the biophilia hypothesis of *Edward O. Wilson* [47.20], on which human beings simply have a profound need of, and liking for, natural forms and systems. Each movement has a distinct ideology associated with it.

Another design movement taking organisms as models, known sometimes as *biomimicry*, views organisms as marvels of engineering. The efficiency and optimality found in organic forms serves as a model for engineering designs [47.21]. Biomimicry is the paradigm explored here. In particular, the exposition of biomimicry given by *French* [47.22] is used because its ideological content is presented clearly.

Like the modernist architects, an important ideal of biomimicry is *functionalism*. That is, design should be dispassionate and analytic. Like *Le Corbusier*, who

sought to separate architecture from nostalgic attachments to historical forms, *French* holds that good engineering is the result of cool calculation, for which evolution provides an instructive example [47.22, p. xii]:

“Living organisms are examples of design strictly for function, the product of blind evolutionary forces rather than conscious thought, yet far excelling the products of engineering. When the engineer looks at nature he sees familiar principles of design being followed, often in surprising and elegant ways.”

The forces of evolution are blind in the sense that they respond only to the problems of the present and are not sentimental about the past or, for that matter, concerned with their legacy for the future.

The anti-sentimental stance of biomimicry focuses the attention of the designer on the products of evolution through natural selection. In many treatments, natural selection is the only kind of evolutionary scheme discussed, for example [47.18].

Other evolutionary forces are set aside. *Darwin* [47.23] argued that, in addition to natural selection, the form of organisms could be explained by sexual selection. The extravagant tail of the peacock, for example, could not be explained by a struggle for existence as it hardly improves the peacock’s ability to fight, fly or feed. However, it could be explained by the need of male birds to impress choosy females. The focus on aesthetics involved in sexual selection is not compatible with the functionalist outlook of biomimicry.

Biomimicry also exhibits the other aspect of functionalism shown by modernism, namely an emphasis on efficiency. *French* notes how organisms display adherence to a principle of economic efficiency, namely the

division of labor. In classical economics, an economy functions most productively when each of its members simply does what they do best, and nothing else. A similar principle holds in the organic world [47.22, p. 3]:

“The division of labour is but a special case of a more general principle of functional design, the *separation of functions*. Thus, simple single-celled organisms have to provide all their functions in one cell, whereas higher animals and plants have many different kinds of cell for special purposes, carrying sap, extracting water and minerals from the soil, transmitting signals, secreting digestive juices, etc. The early steam-engines, following Newcomen’s design of 1712, had a cylinder in which the steam did work and in which it was also condensed. Watt’s engine, fifty years or so later, separated the functions of working cylinder and steam condenser, so greatly increasing the efficiency.”

Organisms evolve to display this economy of organization, thus providing models for engineers.

As before, this characterization of the natural world diverts attention from organic structures that serve multiple functions. The feathers on a bird’s wings, for example, may help that structure to create lift, but also play a role in creating thrust, and supplying a platform for decorative features important in the competition for mates.

Also like modernist architects, although biomimicry involves suspicion of ornament, it shows an aesthetic concern through attention to elegance. A design, like an organism, can achieve elegance simply by being thoroughly functional [47.22, p. 14]:

“One characteristic of functional design is elegance. Most people find a buttercup beautiful, and many would say that the locomotive was at least pleasant to look at. However, the buttercup has an essential elegance, much more fundamental than its mere appearance. It is an elegant solution to a difficult problem in functional design.”

This approach to aesthetics is *reductionist* in the sense that it holds that a kind of beauty, namely elegance, can be achieved by adhering to nonaesthetic values such as dispassionate analysis and economy. This kind of beauty is exemplified by organisms like buttercups but also, to a lesser degree perhaps, by artifacts like locomotives. French’s treatment of elegance is thus much like the modernists’ treatment of honesty in functionalism.

A salient difference between biomimicry and modernism is that the former allows an ideal of *plurality* instead of *conformity*. Modernists like Le Corbusier held that their views on design applied fundamentally

and universally. As such, there is no room for other design paradigms.

Biomimicry, however, admits the (limited) applicability of other approaches to the use of natural models in design. French contrasts the field of engineering with that of architecture, which has a different ideology and occasionally produces good results [47.22, pp. 5–6]:

“However, in much architecture the functional aspects are very secondary to aesthetic ones, and, moreover, rather readily met (or indeed, neglected altogether, as in some badly-designed buildings which have nonetheless won awards). Another defect of architecture as a training ground for functional design was that often the economic constraint, so powerful throughout nature and engineering, was virtually absent, the patrons caring more for glory than the public good (which was the public’s loss then, but is sometimes our gain now).”

Clearly, French’s pluralism is graded and grudging. That is, he places engineering above architecture in terms of the typical quality of its results. Nevertheless, architects sometimes produce works that are both artistic and sound. A medieval cathedral, for example, may be ornamented in beautiful sculptures of people and animals that no engineer would ever produce. Yet, it also includes a broad roof and sturdy buttresses that make it sound and lasting.

One other ideal of biomimicry of the type described by French may be termed *masculinity*. Hofstede [47.24] defines masculinity in a culture as a gender role that emphasizes self-orientation, assertiveness, and ambition over feminine values such as relationships, communication and caring. The preference for masculine features of evolution is suggested by the emphasis placed on natural selection as a competition for survival amongst individuals, a kind of war of all against all. Attention is paid mostly to features that animals have for obtaining food, fleeing predators or fighting rivals. Features that animals have for more feminine tasks such as forming groups, cooperating with others and raising offspring are less often examined. Consider this description of the severity of conditions to which natural designs respond [47.22, pp. 265–266]:

“The difficulties of a hostile environment are added to by those of growth; many organisms must fend for themselves from an early stage; for example, fish may be all on their own when only a centimeter or so long, though they may eventually reach a meter in length. The caterpillar-pupa-butterfly and tadpole-frog metamorphoses are familiar. If the caterpillar is eaten there will be no butterfly;



each stage must be viable. Insects and other arthropoda, such as crabs and shrimps, have a hard outer skin which cannot grow with them, and must be moulted periodically as it becomes too tight; until the soft new armour hardens, they are relatively defenceless.”

Adaptations that animals possess for feminine functions are mentioned briefly and as an afterthought [47.22, p. 266]:

“One of the devices used to cope with the extreme severity of the design problem of living creatures is a very high production rate, so that out of millions of embryos a handful may survive. But some less wasteful approaches have appeared in the course of evolution, principally, the protection of the young by adults among the higher animals and termites, ants and bees.”

## 47.7 Conclusion

Design paradigms exhibit ideological characteristics. That is, they adhere to a set of ideals about what good design is and is not. The selection and use of models in a design paradigm is duly affected by the prevailing ideology. Within a design paradigm, models are selected both because they are model solutions, providing answers to design questions, and because they are model citizens, properly reflecting and reinforcing the ideals of the paradigm.

The ideologies of four design movements have been explored and exhibited. Revivalism, modernism, industrial design, and biomimicry have all been important

approaches to design problems. Each exhibits a characteristic ideology. Each ideology is reflected in the selection and use of models in each case. In this way, each ideology assists in the process of innovation of solutions to design problems, and produces artifacts that bear out the soundness of the ideology.

It seems fair to say that biomimicry concerns itself primarily with those aspects of animal bodies that are most closely associated with culturally masculine activities. As with any ideology, biomimicry tends to be self-reinforcing. That is, models for artifacts may be located in nature according to the ideals described above. Subsequently, the success of artifacts designed in this way attests to the validity of the ideology. For example, perhaps the most famous instance of biomimicry is velcro [47.18, pp. 268–270]. Swiss engineer George de Mestral investigated how burrs stuck so tightly to his coat and his dog’s fur. He went on to devise an artificial equivalent made of nylon that is still a popular fastener. Burrs are the seed pods of burdock plants, which use their adhesive function in order to spread the seeds around as a part of the struggle for survival. Thus, the continuing fame of this example attests to the functional and masculine ideals of biomimicry.

approaches to design problems. Each exhibits a characteristic ideology. Each ideology is reflected in the selection and use of models in each case. In this way, each ideology assists in the process of innovation of solutions to design problems, and produces artifacts that bear out the soundness of the ideology.

**Acknowledgments.** This chapter is based upon C. Shelley: Models and ideology in design. In: *Model-Based Reasoning in Science and Technology: Theoretical and Cognitive Issues*, ed. by L. Magnani (Springer, Berlin 2013) pp. 609–623

## References

- 47.1 A. Forty: *Objects of Desire: Design and Society from Wedgwood to IBM* (Thames and Hudson, London 1986)
- 47.2 D. Tweney: *iPhone’s design tribute to a 1977 Braun calculator* (2007 20–July) Retrieved 2011, 5–Dec. from Wired, <http://www.wired.com/gadgetlab/2007/07/iphones-design/>
- 47.3 W. Isaacson: *Steve Jobs* (Simon and Schuster, New York 2011)
- 47.4 M. Aldrich: Gothic sensibility: The early years of the Gothic Revival. In: *Master of Gothic Revival*, ed. by P. Atterbury, A.W.N. Pugin (Yale Univ. Press, New Haven 1995) pp. 13–30
- 47.5 A. Saint: Pugin’s architecture in context. In: *A. W. N. Pugin: Master of Gothic Revival*, ed. by P. Atterbury (Yale Univ. Press, New Haven 1995) pp. 79–102
- 47.6 R. Hill: Augustus Welby Northmore Pugin: A biographical sketch. In: *A. W. N. Pugin: Master of Gothic Revival*, ed. by P. Atterbury (Yale Univ. Press, New Haven 1995) pp. 31–44
- 47.7 D. Meara: The Catholic context. In: *A. W. N. Pugin: Master of Gothic Revival*, ed. by P. Atterbury (Yale Univ. Press, New Haven 1995) pp. 45–62
- 47.8 L. Corbusier: *Vers Une Architecture* (Vincent, Fréal & Cie, Paris 1923)
- 47.9 W.J. Curtis: *Le Corbusier: Ideas and Forms* (Phaidon, Oxford 1986)
- 47.10 W. Rybczynski: High hopes. In: *City Life: Urban Expectations in a New World*, ed. by W. Rybczynski (Scribner, New York 1995) pp. 155–172

- 47.11 H. Conway, R. Roenisch: *Understanding Architecture: An Introduction to Architecture and Architectural Theory* (Routledge, New York 2005)
- 47.12 R. Loewy: *Never Leave Well Enough Alone* (Simon and Shuster, New York 1951)
- 47.13 S. Barmak: *A pioneer of user-friendly*, (2007; 15–July) Retrieved 2011; 5–Dec. from Toronto Star, <http://www.thestar.com/sciencetech/Ideas/article/238172>
- 47.14 G. Slade: *Make to Break: Technology and Obsolescence in America* (Harvard Univ. Press, Cambridge 2006)
- 47.15 W. Gibson: The Gernsback continuum. In: *Universe 11*, ed. by T. Carr (Doubleday, New York 1981) pp. 81–90
- 47.16 J.L. Meikle: *Twentieth Century Limited: Industrial design in America, 1925–1939* (Temple Univ. Press, Philadelphia 1979)
- 47.17 E. Lupton: *Mechanical Brides: Women and Machines from Home to Office* (Princeton Architectural Press, New York 1993)
- 47.18 S. Vogel: *Cats' Paws and Catapults: Mechanical Worlds of Nature and People* (W. W. Norton & Co., New York 1998)
- 47.19 W. McDonough, M. Braungart: *Cradle to Cradle: Remaking the Way we Make Things* (North Point Press, New York 2002)
- 47.20 E.O. Wilson: *Biophilia* (Harvard Univ. Press, Cambridge 1984)
- 47.21 W. Nachtigall: *Biological Mechanisms of Attachment: The Comparative Morphology and Bioengineering of Organs for Linkage, Suction, and Adhesion* (Springer, Berlin, Heidelberg 1974), transl. by M. A. Beiderman-Thompson
- 47.22 M.J. French: *Invention and Evolution: Design in Nature and Engineering* (Cambridge Univ. Press, Cambridge 1988)
- 47.23 C. Darwin: *The Descent of Man and Selection in Relation to Sex* (John Murray, London 1871)
- 47.24 G. Hofstede: *Culture's Consequences: International Differences in Work-Related Values*, 2nd edn. (SAGE Publications, Beverly Hills 1984)

## 48. Restructuring Incomplete Models in Innovators Marketplace on Data Jackets

Yukio Ohsawa, Teruaki Hayashi, Hiroyuki Kido

Innovators Marketplace, a market-like workshop where cards showing existing pieces of knowledge in various domains are combined to create ideas of services/products and thrown into demand-driven communication to choose practical ideas, has been extended to a setting of the market of data. This extension is called Innovators Marketplace on Data Jackets, a workshop in which each prepared card called a data jacket represents the digest knowledge about a dataset, that is, a kind of metadata. Data jackets are disclosed, whereas the corresponding data are not, and participants of the workshop create ideas for combining and analyzing data using the visualized correlation of data jackets. In this chapter, this workshop is described as a systematic process for reasoning on incomplete models, where each data jacket is regarded as an incomplete local model in the domain of the data, and communication is launched for satisfying requirements in the market (regarded as incomplete global models) by restructuring and combining local models. The data jacket may initially include atoms and terms in the domain, not connected via complete causal relations. Via the communication, however, links including causal relations appear and are revised toward obtaining a *glocal* model corresponding to a solution to satisfy requirements in the marketplace. In this process, the local model corresponding to each element is also revised to obtain useful knowledge digesting the corresponding data.

48.1	<b>Chance Discovery as a Trigger to Innovation</b> .....	1016
48.2	<b>Chance Discovery from Data and Communication</b> .....	1016
48.2.1	Chance Discovery as a Problematic Child of Data Mining .....	1016
48.3	<b>IM for Externalizing and Connecting Requirements and Solutions</b> .....	1020
48.4	<b>Innovators Marketplace on Data Jackets</b> .....	1022
48.4.1	Marketplaces of Data .....	1022
48.4.2	The Procedure of IMDJ .....	1022
48.5	<b>IMDJ as Place for Reasoning on Incomplete Models</b> .....	1023
48.5.1	Grounding Incompletely Defined Models Into Well-Defined Models .....	1023
48.5.2	Abductive Reasoning for Thoughts and Communications in IMDJ .....	1025
48.6	<b>Conclusions</b> .....	1029
	<b>References</b> .....	1029

One century passed since Schumpeter suggested the concept *innovation* as a creative activity by which the economy jumps up to a new state, and that this creativity is to be realized by a novel combination of industrial resources [48.1]. After half a century, *Rogers* pointed out that leading consumers play the role of *innovators* [48.2]. That is, an important idea cannot become an innovation without innovators who are not only those who produce products/services but also consumers who discover new values of products in using them and diffuse the discovered values to the majority. Here lead-

ing consumers can also invent, not only use and diffuse technologies [48.3]. Thus, innovation came to be a term referring to the thoughts and the interaction of stakeholders in the market, involving inventors in companies, sensitive and communicative diffusers, and also consumers. As a result of their interaction, novel dimensions are to be introduced for evaluating and improving the performance of humans' activities in the real life, as in *Drucker's* redefinition of innovation "a change that creates a *new dimension of performance*," which urged him to discuss *do's* are essential paths to innovation [48.4, 5].

## 48.1 Chance Discovery as a Trigger to Innovation

In this chapter, we aim at showing the potential contribution of model-based reasoning to modeling the process of data-driven innovation. For this, we introduce *chance discovery* that means to discover a chance event – an uncertain event significant for making a decision, as in the definition since 2000 [48.6]. Especially if the chance event is rare or novel, the attention to the event may trigger the creation of a novel performance dimension to the human life, as Drucker's above definition of innovation. In fact, users of methods for chance discovery went beyond their previous achievements in data-based decisions, as exemplified in the next section. In successful cases [48.7, 8], chance discovery has been understood as innovation by enhancing humans' activities with value sensing [48.9] and sense making [48.10, 11], taking advantage of tools for data visualization positioned and used in the process for decision making. Areas such as evidence extraction and link discovery (EELD) [48.12] shared this basic idea, but the point of chance discovery was unique in the sense we focused on the effect of communication with sharing a graph visualizing the correlation of events in data for creating scenarios of actions that satisfy stakeholders' intentions and constraints, by connecting frequent patterns via chance events.

In this chapter, first we review some basic approaches to chance discovery. Then Innovators Marketplace (IM) [48.13] will be introduced as a methodology of workshop for chance discovery, where participants are aided to communicate from individual viewpoints reflecting their roles in the market, reason with combin-

ing pieces of basic knowledge as elements for creation, and communicate to introduce various aspects and knowledge for improving presented ideas and choosing practical ideas. Then in the latter half of this chapter IM will be extended to IM with Data Jackets (IMDJ), recently developed in order to fulfill intentions of business people by social data sharing without violating constraints of stakeholders.

From the aspect of model-based reasoning, each element to be combined in IMDJ initially includes a set of atoms and terms in its *local* domain, which are not yet connected via causal relations. Links including causal relations are to be given via reasoning and communication to combine the elements, to embody and realize consumers' requirements casted as a *global* model, in the sense the requirements are not restricted to a definite relevant domain. In this process, the local model corresponding to each element is revised, with constructing and reconstructing *global* models that mean the links between local and global models, via the idea-revising communication as shown later. Here stakeholders externalize intentions and constraints via communication, so that latent values are externalized and scenarios of actions for realizing those values are created. Strategies to combine and/or analyze data are obtained, as solutions for the requirements, as a result of choice from these scenarios. A logical framework for guiding this process of reasoning in IMDJ is shown in this chapter, on which we propose a new procedure of IMDJ where participants shall obtain feasible solutions to satisfy requirements of data analysis users.

## 48.2 Chance Discovery from Data and Communication

Data mining recently tends to be regarded as a method for showing objectively useful knowledge, as far as people read application cases of data mining superficially. This is a reasonable trend, because large amount of data are collected via sensing systems such as POS (position of sales) registers, RFID tags, events in a networked system, etc., without interruption of subjective thoughts of humans. Furthermore, the importance of attention to the volume, the variety, and the velocity of changes in data is coming to be widely and concretely conceived in businesses and sciences.

### 48.2.1 Chance Discovery as a Problematic Child of Data Mining

These features of recent data are, however, really casting a serious problem. A number of variables in data

increases with the growth in the variety, and the volume of data embracing changes reflecting the changes in the real world is making it hard to choose essential attributes (which may be called variables). For example, suppose the features of animals have got collected in huge data corresponding to the overall history of evolution. Via evolution, anus and urethra have been developed and came to be separate for almost all mammals, not for birds or reptiles. Thus, the number of holes at the bottom of the body may be regarded as a useful attribute for defining mammals as a separate class from others. To define mammals, however, this attribute cannot be regarded as the most essential because other attributes may take place. For example, the stability of blood temperature, that is, monothematic or not, is a feature regarded as typical for mammals. Attribute selection methods have been developed for explaining

classes simply and correctly, on selected essential attributes without such a redundancy [48.14, 15].

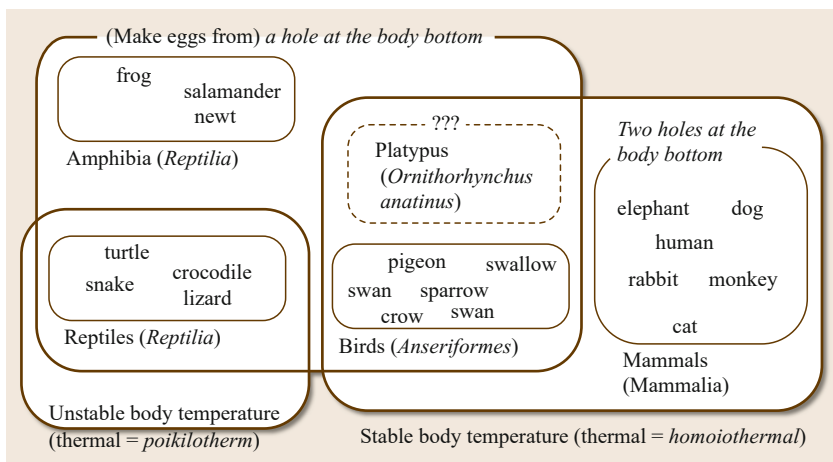
On the other hand, recently the attention of users (rather than researchers) of data mining came to be paid to a new problem – explanation of the boundary between classes. For example, some rare species are hard to be classified into mammals or reptiles or birds, because they have features that cannot be covered in a single class as far as we apply such attributes as above. Platypus has one hole at the body bottom that works both as anus and urethra, but has a stable blood temperature, that is, homoiothermal. This makes platypus be positioned at the boundary between mammals and reptiles/birds. Here we should now restore the importance of the number of body bottom holes for classifying animals into two large classes (mammals, and reptiles/birds) and also explaining the outstanding feature of the rare samples of homoiothermal animals positioned between the two classes. Although platypus in Fig. 48.1 has been finally classified as a member of mammalia, such a rare sample in data has been ignored as a noise in quite classical methods of data analysis, whereas other studies highlighted predicting rare events [48.16, 17], extracting exceptional patterns [48.18, 19].

On the other hand, in the approaches to chance discovery under the definition of a *chance* as an event significant for human’s decision, we stood on the principle that a decision is to choose one from multiple scenarios of actions/events that can be taken in the future. Based on this principle, a chance can be regarded as an event at the cross point of multiple scenarios, may be transient and rare, after which the forthcoming sequence of events is uncertain. The cross point is an important candidate of chance if it means a trigger of contextual shift, that is, an event that occurs in transition from one established scenario to a scenario that occurs

in a new context. This is understandable by relating the dimension of performance in Drucker’s definition of innovation to the *context* here. That is, a noteworthy chance should be at the boundary of contexts, such as platypus for biologists desiring to explain the history of evolution. Symptoms at the boundary of disease progress and recovery [48.20] and products at the boundary of different contexts of consumptions [48.7] have been obtained in cases of chance discovery. Chance discovery can be positioned as a data-driven guidance of humans’ attention to black swans [48.21].

An important point here is that not all cross points of scenarios are necessarily noteworthy chances. Thus we need a method to choose a useful event, even if we can be informed about all events at contextual bounds based on data available. For doing this, one way may be to count up all possible sequences including candidates of chances, and compare their evaluated utilities. This, however, is inefficient even for a computer if we have a large number of observations corresponding to the large number of events and their attributes in data. Considering this point, we focused efforts to developing methods to take advantage of human’s sense to choose high-utility scenarios and events, borrowing computer’s power to visualize a map showing the complex structure of event–event co-occurrences (actions of humans have been counted as events).

However, even if one has got a sophisticated software for visualizing big data, it is not easy to learn to sense the utility of a sequence of events without embedding one’s body in the real-world situation and acquire experiences of gain/loss due to various sequences of events and actions. In other words, the process of acting on/for thoughts with manipulation [48.22, 23] is required for enabling data collection and interaction with the environment, toward opening one’s mind to inconsistencies and questions. As a result, hidden events are

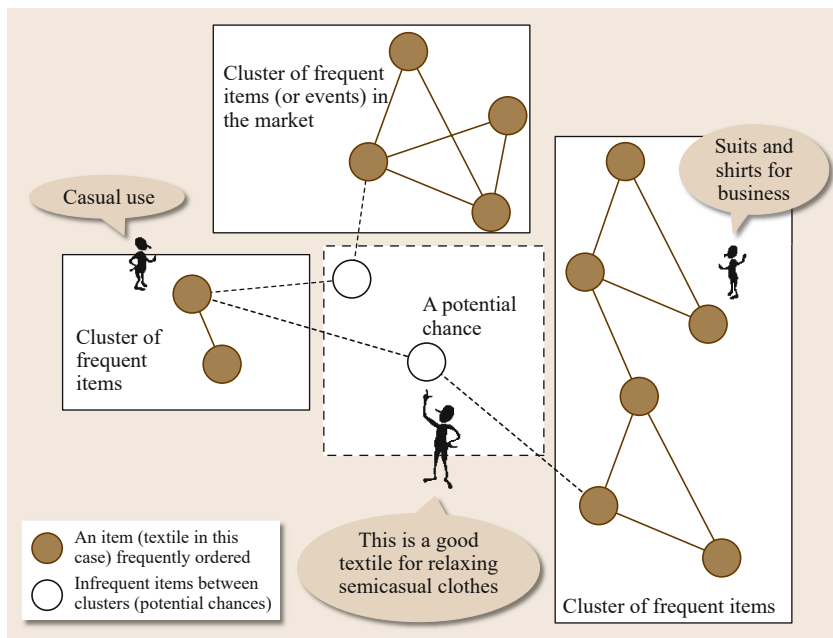


**Fig. 48.1** A novel/rare, event or item, guiding to the discovery of useful variables for explaining the boundary of classes. Such variables give insights for modeling hidden causalities

revealed and *variables* – identical to *attributes* above such as the number of holes in the animal body in the example above – are highlighted. From the collected data, scenario communication of humans [48.8, 24] aided by data visualization works in externalizing latent criteria for evaluating an event as a candidate of chance. Such externalized criteria may come to be regarded as latent variables, of which the contribution to the utility of an event (or of the scenario including an event) had been discounted. In this step, analogy [48.25, 26], insight with reframing [48.27, 28], and metacognition came to be argued as bases for realizing chance discovery as a systematic process toward perceiving useful information for decision making. In projects we conducted with industries, teams of marketing or product developing acquired novel awareness of valuable parts of their markets they had not taken into consideration. For acquiring the awareness, such tools for visualization as KeyGraph [48.7, 20, 29–32] assisted decision-making users by showing a map of the market having (1) a set of clusters of items frequently co-selected, that is, items in each cluster have been chosen by the same customer at the same time, and (2) rare items at the boundary of these clusters, which may imply a latent market coming up in the future. In other words, we regarded such a rare item as a candidate of chance, because a scenario that is a sequence of actions and events to occur in a certain context is represented by events or items in each cluster. Items in (2) can be regarded as bridges between scenarios, at their cross points, or a significant chance to switch from/to sce-

narios as in the case of a biologist looking at animals in Fig. 48.1 with the desire to discover an explanatory scenario of history of evolution.

Here let us refer to a case presented in [48.7], where a map obtained by KeyGraph assisted chance discovery in business, which will be linked to IM in the next section. For this example, KeyGraph showed clusters in item (1) above, and rare items in (2) above which meant products in the niche not yet popular. The map obtained in a textile manufacturing firm from data of customers' choices of textile products, was as outlined in Fig. 48.2. Here, the black nodes linked by solid lines show clusters above. The white nodes and the dotted lines show items corresponding to (2) above and their co-occurrence with items in clusters respectively. Ten marketers in this textile firm attached real textile samples corresponding nodes in the printed map. The marketers, looking at the figure and touching the textile items on it, discussed about their new scenarios, that is, plans of business. First, they noticed the clusters corresponding to popular item sets mentioned in (1) above: The large cluster in the right means an established market of textiles for business clothes (suits, shirts/blouses for under suits, etc.), and the one in the left came to be interpreted as of casual (called *worn-out*) wear by multiple marketers. Then, a few marketers pointed out that consumers desire to change from one cluster to another. For example, when one moves from the place to work at daytime to a restaurant for dinner after working, one may like to change clothes from business suits to casual for relaxation. Interested in such consumers,



**Fig. 48.2** A successful use case of chance discovery. Each node represents a textile (*black*: frequent textile, *white*: infrequent), links co-occurrences in customers' ordering (*solid*: between frequent items, *dotted*: between a frequent item and an infrequent) (after [48.7])

the marketers came to pay attention to an item between the clusters, because they noticed and agreed that the item is suitable as a material of a new jacket to wear after working, without changing trousers, as desired by consumers. In this case of Fig. 48.2, their new scenario was to sell the new item at the center of the map that was not yet popular by that time, to suit manufacturers as a material of casual jacket office workers can wear fitting the trousers of suits to go out for dinner.

Such a creative communication in business can be regarded as a manner for externalizing latent knowledge, of marketers and developers of textile in the case of Fig. 48.2. As studied in requirement engineering, externalizing latent intentions and constraints of stakeholders are an essential step for designing acceptable products/services [48.33–35]. And, the latent (tacit dimension of) knowledge presented by *Polanyi* [48.36] behind activities of humans should be and can be externalized for and by enabling a creative process of collaboration in businesses [48.37]. However, we should also note that externalized or created knowledge cannot be always evaluated highly or accepted by others.

This point can be linked to sticky information [48.3, 38, 39], meaning information is to be localized to individual people who may be either industrial inventors or consumers. The information about consumers' requirements and the knowledge of inventors are hard to be transferred from/to each other, which may disturb the mutual understanding of all stakeholders including consumers. This happens even between developers of a product, so sometimes teamwork in a firm comes to be disturbed. In successful cases, even if useful information in a firm is sticky, users of a product may propose new designs from their own utility-intensive viewpoints whereas manufacturers may design products to improve the power efficiency from the solution-intensive viewpoints. This might be good enough if the proposal of user is easy to implement with techniques available for users or if users get satisfied by the technical improvement by manufacturer. However, this does not stand if users' ideas are not feasible or if technologies of inventors are not easy to understand or available for users.

A new approach of chance discovery to this problem can be found in the invention of *tsugology* [48.40], where the three-tuple hidden behind an action – *intention*, *preconstraint* (pre-existing constraint), and *postconstraint* (the constraint to be made after and due to the action) – is called a *tsugo* after a Japanese daily use word. People, living in social relationships, should externalize one's and others' *tsugoes* for choosing actions that are feasible, that is, for realizing an intention under constraints that sometimes emerge from the real life. A point of *tsugology* (studies for externalizing and connecting *tsugoes* of stakeholders of a problem

in face) is that constraints emerge dynamically. That is, post-constraints may emerge from one's action and may make a preconstraint on others' actions, in contrast to static constraints considered in design methodologies in the literature. For example, if Mr. X sells out data on consumption of foods and drinks, the preconstraint of consumers, such as their requirement to protect their privacy, may be violated. In this case, the leakage of information about privacy is a post-constraint for Mr. X that had not been considered before but got externalized in discussing the plan to find business opportunities from consumption data. By noticing such a constraint, Mr. X should think of a new action, such as using data without customers' IDs but just linking his data with other's data about weather, via date and time for realizing his original intention to sell beer efficiently to suitable consumers. This intention itself may also get externalized via speaking out his own thoughts in a workshop with neighbors and colleagues. As in [48.40], we found that the essential part of sticky information is relevant to underlying *tsugoes*, that is, intentions and constraints of stakeholders.

We can summarize the recommendable process to chance discovery [48.6], learned from successful and unsuccessful cases so far, as follows. The process should start from data collection based on user's (or users') interest. Then, one can visualize the data using suitable tools for chance discovery, and regard it as a map of the market. The visualized result aids in collecting stakeholders as participants of communication to discuss novel scenarios leading to a successful business. This process can be reinterpreted by the four-step spiral below:

1. *Sensing external events*: Keep sensing events in the environment, and collecting data on events that are beyond individual human's sensing capacity.
2. *Recollection*: Try to recollect and explain scenarios from the past. These are sequences of events and actions, explained with the background context. For this step, *subject data*, that is, text about one's and teammates' thoughts of actions and *tsugoes*, should be collected to visualize thoughts as an itemized list of sentences, drawn images, or a graph of word-to-word correlations. Individual's meta-cognition of one's own awareness is thus accelerated.
3. *Scenarization*: Extend the scenarios about the past into scenarios as plans for the future, applying analogy, and explain the scenarios as done in step (2). Data visualization here aids participants' building of scenarios based on the visualized connections of events and items.
4. *Co-evolution of scenarios*: In creating, by combining, scenarios in (3) participants speak out to

externalize variables, that is, embodied criteria to evaluate the utility of created scenarios. Variables relevant to tsugoes of stakeholders are to be externalized here from awareness of inconsistencies between scenarios of participants. Crosses of scenarios are obtained as a result of this step.

Via these steps, the following three effects are expected:

*Effect I: Representing basic knowledge*, that is, the preparatory knowledge to be explained in step (2) above and combined for innovation mainly in step (4). The knowledge should be represented by individuals in order to enable the reasoning for novel combination and explanation of scenarios. This requirement for step (2) urges individuals to discuss latent contexts behind events in step (1).

*Effect II: Reasoning with combinations and analogy*: The extension with analogical transplantation and combination in steps (3) and (4) of basic knowledge, possibly with changing minor parts, are executed. For this, participants are urged to find similarities between the bases (problems for which basic knowledge worked previously) and the target (the present

problem) and explain them as common features of the past and the future, which work as candidates of common variables of scenarios to be combined. For example, *relaxing* in the center of Fig. 48.2 can be a variable to explain an obtained use scenario of the new item which is a candidate of potential chance.

*Effect III: Communication with presenting users'/inventors' conditions*: Individuals speak, reflecting each one's living condition, facing conflicts. Such a communication is relevant to step (4) above, urges awareness of latent conditions which should be considered for reaching consensus about the scenario to choose in decision making. For example, the requirement to fit the trousers of suits could be considered in modifying the color of the new textile in the center of Fig. 48.2.

Note that items above show our viewpoints to highlight the link between chance discovery and our methods of workshops for innovation. For more general principles of innovative collaboration, reader is referred to references on collaborative design approach [48.41–43], to find processes to evaluate and discover values of existing and emerging items.

### 48.3 IM for Externalizing and Connecting Requirements and Solutions

Here again let us confirm that this chapter has been written to show the utility of model-based reasoning in data-driven innovation. For this purpose, the method IMDJ to be introduced in Sect. 48.4 plays a role of environment setting for communication toward data-driven innovation based on the methodology of chance discovery, by which we propose a model-based explanation of reasoning to combine and analyze data. And, in this section, we introduce IM in preparation for Sect. 48.4.

IM is a gamified workshop where ideas are created by *inventors* who combine elements, that is, existing technologies and pieces of knowledge, and evaluated by players who play their roles as *consumers*. IM can be regarded as a method for chance discovery. That is, IM is a workshop where each inventor presents one's own basic knowledge from the viewpoint of a member of the market (e.g., a provider/inventor of products), reason with combinations and/or analogy, and communicate with consumers, focusing and shifting contexts. The case in Fig. 48.2 can be regarded as an origin of IM, where marketers knowing different parts of the market and relevant technologies communicated to present emerging demands for creating a strategic scenario of business.

IM starts with a given set of *basic cards*, on which summaries of existing pieces of knowledge are printed. The cards are put on a game-board that is a graph, visualizing the correlations among contents of basic cards in order to aid the creation of inventors' ideas. As in Fig. 48.3, the inventor's tasks are to buy a preferable number of basic cards and combine some of those cards to present an idea for new business. Other inventors may propose the presenting inventor to start collaboration or share the created idea, with negotiating the dealing price. Here each can use a presented idea as a new element to combine with basic cards or other ideas. The inventor having got the largest amount of money at the halting time (a fixed length of time after starting) wins the game. On the other hand, each consumer plays one's own role, which is an occupation chosen from among a variety such as *housekeeper, farmer, medical doctor, power industry, aged people, students, manufacturer, transportation*, etc., and buys a preferable idea from some inventor for the price determined by negotiation, for improving the life quality of his/her own role. A consumer becomes the winning consumer if he/she obtains the highest value as a result of other participants' evaluation about his/her presenta-



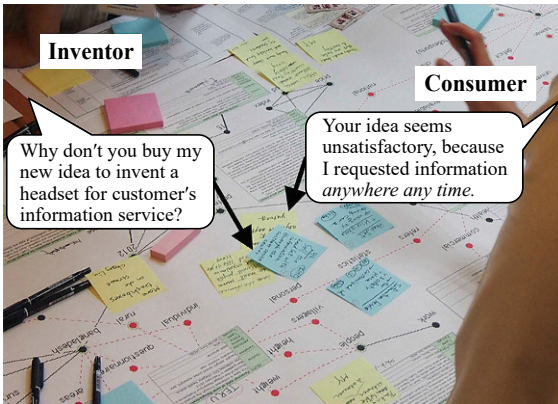


Fig. 48.3 A scene of Innovators Marketplace (IM)

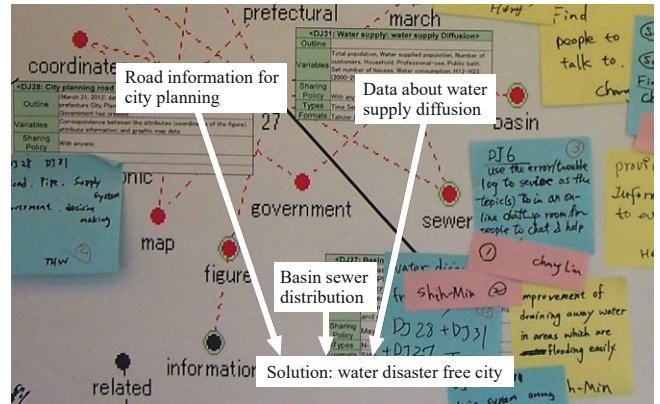


Fig. 48.4 Innovation game on the game board, made of KeyGraph

tion at the halting time. This presentation is about how the quality of life will be improved by the ideas the consumer bought in IM. IM has been introduced in industries and researches, for example, for strategies in businesses and for sustainable safety of nuclear power plants [48.13].

To make the game-board, the contents of basic cards are collected to make a text file and visualized with KeyGraph [48.29] or its extension such as data-crystallization [48.44]. Here KeyGraph shows the positions of not only existing knowledge but also of latent possibility that a new idea may emerge by combining pieces of existing knowledge connected or closely located in the graph. For example, the basic cards showing existing data in Fig. 48.4 (*DJ* here will be explained later in Sect. 48.4) are *DJ27: Basin sewer distribution* that shows the location of basic sewers in a city, *DJ28: City planning road information* that enables to investigate the traffic conditions on each road of the city, and *DJ31: water supply diffusion* that means data about quantity of water supplied at each point of the city. By combining these pieces of information, an inventor proposed the idea of *water disaster free city* that is to be realized by enabling to extinguish flooded polluted water and to supply clean water to citizens. Since this idea was evaluated highly by participants who attended as consumers, hereafter the participants came to be encouraged to apply data to reinforce disaster protection.

Reader interested in the details and the novelty of IM are advised to see [48.13], but let us show a very brief comparison of IM with another methods

of idea creation using stickers and/or cards. That is, the Kawakita Jiro (KJ) method (as introduced *Hambleton* [48.45], called Affinity Diagram) has been world widely used for half a century since it was invented. In KJ, collected pieces of information are put on cards, arranged on a two-dimensional space and classified hierarchically into clusters and subclusters, reflecting the concepts participants externalize. Then, lines between clusters are drawn with titles, corresponding to participants' ideas about relationships among the clusters. Finally the location of all cards and lines among them are observed. As a result, concepts and scenarios are acquired by tracing the lines and closely positioned cards. However, because visualization such as KeyGraph has not been employed in KJ, the bird-eye view of the associations of ideas, is not easily available in the early stage. As shown by *Ohsawa* and *Nishihara* [48.13], we published technical features of IM such as the effects of KeyGraph. It is more important, however, to learn the manner of communication in IM, borrowed from the market of products and services, where people interact competitively and co-creatively with evaluations of ideas using toy money. And, the example above (Figs. 48.3 and 48.4) is of a specific IM for combining and reusing data that is, of IMDJ to be mentioned below. That is, the novelty of Innovators Marketplace is not a point in this chapter where we focus on the logical structure of users' thoughts in IMDJ that is a specification of IM. In other words, we use IMDJ as a setting of collective intelligence for data-driven innovation, in order to discuss the meaning of model-based reasoning.

## 48.4 Innovators Marketplace on Data Jackets

The IMDJ method has an additional novelty to IM, which is the data jackets (DJ). The point of DJ is that it enables participants to show out the digests of data they own for evaluating the potential use value of data, without opening the content. DJ can be used as a knowledge unit as a card in IM to be used as an element to be combined with others in the workshop. Advising reader to see [48.46, 47] for some details, below let us briefly introduce IMDJ.

### 48.4.1 Marketplaces of Data

As in the previous work such as IM, we define stakeholders as people involved in the process to solve a problem. In order to realize creative communications for externalizing and solving problems they potentially share, stakeholders should share data to the extent they can. The market of data is a useful concept for realizing such a creative communication of stakeholders seeking innovation, in that the market is essentially a social environment where items are created and exchanged for reasonable conditions. That is, data in the market of data are created, valued, sold, opened free, shared after negotiation or based on some external guiding force such as the governmental control. In the on-line market of data [48.48, 49], a number of data digests are exhibited, similar to the catalogue of free-access open data (e.g., [48.50, 51]) except that prices are assigned to each dataset in the market. The effect of thus pricing data is significant. That is, potential users and providers of data are enabled, respectively, via pricing, to provide and choose suitable data to share. Furthermore, by estimating the utility of each dataset via negotiation to discount/raise the price, stakeholders can even discuss for improving use scenarios of data to fulfill new values in their businesses. For this purpose, data should be priced reasonably on fair negotiation of stakeholders – owners, analysts, users, and brokers of data.

Thus, the market of data is expected to be a place where the value of data are communicated with externalizing and sharing potential scenarios for combining and using data. In addition, analysts often need to learn techniques from each other. In order to analyze data in hand, it is an essential step for an analyst to import structural models of causality from domains of other analysts, and to choose a suitable model for explaining the latent causality in his/her target domain. This, in other words, is analogical learning of the basic knowledge structure [48.52] or of knowledge-use process from other analysts. In comparison with the diverting use of previous analysis, called transfer learn-

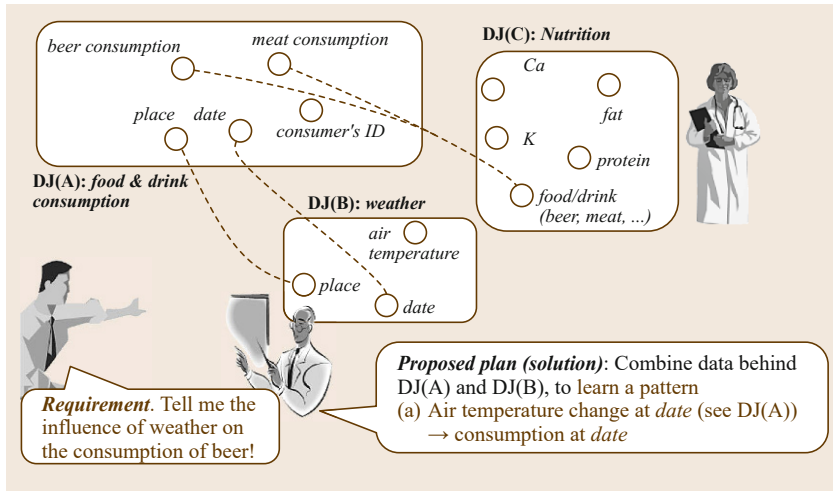
ing [48.53], the effect of the marketplace is to have humans, rather than machines, learn generic ideas from other analysts' experiences.

The thoughts for data-driven innovation in IMDJ, where values of data are externalized and shared in the wide society, can be explained as a model-based reasoning (MBR) as in this chapter. The basic idea of IMDJ comes from what customers and salesclerks do in shopping stores of media such as movie DVDs, where only quite superficial pieces of information in disk jackets are shown for advertisement or for exhibition to customers approaching shelves. On the other hand, the content of DVD – movies, music, etc. – should be hidden in order to reduce the risk that the details may be copied and used free by anyone who does not pay or by rivals. Such a policy may look like an attitude of data closure that may suppress data-driven innovation, but is a useful idea as the basis of IMDJ where each data owner takes part with filling in and disclosing only DJs, which are just small pieces of information describing the digest of existing data. These pieces work as the unit of knowledge in the logical reasoning.

### 48.4.2 The Procedure of IMDJ

IMDJ [48.46, 47] is a specification of the IM. IMDJ is a process where participants propose ideas to combine data, even if the data are confidential, by disclosing DJs that are digests of existing data or data to be collected in the future, and to analyze the combined data. In contrast to data contents, hard to collect or manage, DJs are easy to collect (just enter via [48.54], the entry sheet of DJs), and to describe links between data. DJs here correspond to elements to be combined in IM. For example, a DJ corresponding to weather data and another to data on consumptions in a restaurant can be disclosed, although the data contents should not be, and used for understanding the potential relevance of data about weather and about food consumption. Such an understanding will be urged by presenting *time* and *place* as common variables described in the two DJs. By communication with sharing attention to these variables, participants in IMDJ may propose a scenario to combine and analyze datasets via links of DJs, to discover reasons for new activities.

IMDJ is a gamified workshop, where winners are to be determined in the same manner as in IM. That is, one who earns the largest amount of money becomes the winning *inventor*, and one becomes the best *user* if having made the best presentation about how the solutions one bought in the game can satisfy the own requirements in the market, as customers in IM did. Each



**Fig. 48.5** The core step in Innovators' Marketplace on Data Jackets (IMDJ) (after [48.47])

participant listens to and evaluates others' utterances in order to buy practical and useful solutions in order to become the winning user, and tries to present solutions in order to be evaluated highly and become the winning inventor as in Fig. 48.5 [48.47], in the communication phase. For example, by combining two datasets, one about weather and the other about liquor consumption, a solution such as *one tends to drink one more bottle of beer if the air temperature is higher by 3° than the day before*, can be proposed. This idea may be interesting to a user working for a brewery, but may be criticized if not yet satisfactory to others. Criticisms are not easy to listen to, but urge improvements of solutions in IMDJ because participants consider them rather than being upset, in a gamified mood.

The utility of data may not be evaluated directly in IMDJ, because data are not necessarily disclosed. However, the utility can be evaluated reflecting the eval-

uation of ideas in DJ. That is, the evaluation of an idea reflecting its feasibility and utility, that is, coincidence with participants' requirement, will be reflected to the evaluation of DJs used for creating the idea. As a result, a user interested in a solution pays both to the inventor for the solution and to the expert, that is, provider of the DJ used, for the data used for inventor's creating the idea. This indirect evaluation of data during the playing time of IMDJ will be finally reflected to the condition in which the data get provided to analysts or users. Data corresponding to DJs used in creating an idea (solution), bought for a high price in IMDJ, are expected to be purchased for a high price after the session of IMDJ. On the other hand, even if the DJ has been used in many solutions, the owner of the corresponding data may decide to open the data free to the public in spite of the high evaluation in IMDJ, for the benefits of people in the society.

## 48.5 IMDJ as Place for Reasoning on Incomplete Models

### 48.5.1 Grounding Incompletely Defined Models Into Well-Defined Models

IMDJ can be expressed as a process to revise models of actions and events by coupling three types of information, that is, (1) of requirement from users, (2) of data from experts or data owners, and (3) of proposals from inventors. (1), (2), and (3) can be, respectively, regarded as a *global*, *local*, and *glocal* models, as follows:

1. A *global* model: The representation of desired knowledge that a participant playing the role of data user (corresponding to consumer in IM) expresses, as a requirement without restriction of the technical domain for solution
2. A *local* model: A set of atoms and/or terms in a domain of technology, which may be used for solving the requirement above but not yet connected via causal relations that are certain to be valid. These models are to be given initially as DJs
3. A *glocal* model: The representation of a connection between local and global models, represented by using elements in multiple local domains. Such a model is expected to be created as the result of inventors' reasoning

Requirement models of (1) above form an incomplete structure of demanded causality or relations among events. For example, a brewery's requirement to know the causality between weather and beer consumption can be put in natural language as *if weather changes, the consumption of beer may change* that is incomplete in that it is unknown which variables of weather should be included in the condition *if weather changes*. Such a causality can be described as an incomplete and uncertain model as

$$beer\_consumption \leftarrow ? weather\_change, \quad (48.1)$$

where the LHS of the relation represented by  $\leftarrow ?$  shows a tentative conclusion of the condition given in the RHS. Let us call this as a declared requirement, or goal  $G$ . Yet, the causality is not convinced or numerically evaluated in such a form as conditional probability  $P(beer\_consumption|weather\_change)$  because neither  $beer\_consumption$  nor  $weather\_change$  are defined as variables in data. The causality may be modeled as in (48.2), on the other hand, via the analysis of data combining data about beer consumption and about weather. That is, some model may have been described by the expert of each dataset, as the *local model*, over functions such as  $air\_temp(date)$  and  $beer\_percons(date)$ , where  $date$  means the variable in the form of  $dd/mm/yyyy$  appearing both in data  $beer\_percons$  and  $air\_temp$ . If so, the proposed model in (48.1) can be concretized as in (48.2) via data analysis and become a candidate of solution that embodies the requirement represented. This is a glocal model because this connects the goal in (48.3) and local models that is, data given in local domains such as beer consumption and of weather. The local model is made complete by data supporting.  $l$  and  $K$  respectively mean liter and Kelvin.

$$\begin{aligned} beer\_percons(date) &> 0.31 \\ \leftarrow air\_temp(date) - air\_temp(date - 1) &> 2 K \end{aligned} \quad (48.2)$$

The grounding from the incomplete model as in (48.1) to the well-defined model as in (48.2) should occur based on the knowledge of expert about data, via the communication between experts and inventors in IMDJ and analysis of data if applicable. Below we aim at formally describing the systematic process where the requirement, that is, the user's goal as in (48.1), is to be externalized and the solution to the requirement as in (48.2) can be obtained based on experts' local models corresponding to provided data. The three stakeholders, that is, experts, inventors, and users of data should interact for restructuring the local, glocal, and global models, respectively, in IMDJ in order to realize such

a process. In this section we aim at describing the constraints we should realize as a social system where the process is to be realized.

*Innovators Marketplace on Data Jackets, described as a list of constraints*

$DJ_i (i \in [1, N])$ : The  $i$ -th DJ, among those collected before or created in IMDJ

$DJ_i := \{F_i, P_i, V_i, \}$ , where elements are defined as follows:

$F_i$ : the set of variables in  $DJ_i$ , expressed as functions over other variables in  $DJ_i$

$P_i$ : the set of predicates (relations of variables) in  $DJ_i$

$V_i$ : the set of other variables in  $DJ_i$  than  $F_i$  or  $P_i$ .

$G$ : The goal, that is, the requirement incompletely defined as the relation over informal terms corresponding to events/features as in (48.1).

$T$ : The theory, that is, a model described by a set of clauses represented by a combination of defined predicates in  $P_G$  below.  $T$  is given by  $T(P_G, F_G, V_G)$ , that is, a relation over elements  $P_G, F_G$ , and  $V_G$ , of a set of DJs in  $DJcom(G)$  defined as in (48.3), which satisfies (48.4) where  $[v]$  for variable  $v$  means the defined domain range of the value of  $v$  and also satisfies *Conditions I and II* mentioned below. Below, we list a part, rather than all clauses, of  $T$ , in exemplifying  $T$ .

$$\begin{aligned} DJcom(G) &:= \{DJ_a, DJ_b, \dots, DJ_L\} \\ &\subseteq \{DJ_1, DJ_2, \dots, DJ_N\} \text{ where} \end{aligned}$$

$$\begin{aligned} P_G &:= P_a \cup P_b \cup \dots \cup P_L, F_G := F_a \cup F_b \cup \dots \cup F_L, \\ V_G &:= V_a \cup V_b \cup \dots \cup V_L. \end{aligned} \quad (48.3)$$

$$\begin{aligned} \exists v \in V_G [\forall V_x \in \{V_a, V_b, \dots, V_L\}, \\ \exists v_x \in V_x ([v] \cap [v_x] \neq \emptyset)]. \end{aligned} \quad (48.4)$$

*Condition I*:  $\exists G' [G \supseteq G' \Leftarrow T]$ , which means goal  $G$  subsumes a conclusion  $G'$  derived by theory  $T$ . Here, subsuming means there is a substitution to elements in  $G$  that implies  $G'$ , and substitution here means to rewrite terms in  $G$  using defined functions in  $F_G$ , variables in  $V_G$ , and predicates in  $P_G$ . This is a generalization of substitution to variables, for the grounding of informal (in natural language) expressions of a goal.

*Condition II*:  $T$  is *completely defined* (this concept is defined just after here) and consistent with data  $D_a, D_b, \dots, D_L$  corresponding to DJs in  $DJcom(G)$ .

$\{DJ_a, DJ_b, \dots, DJ_L\}$ , put  $DJcom(G)$ , is the DJ set, the combination of which satisfies  $G$ .

*Completely defined  $T$*  above means a completed glocal model, which is a logically consistent relation between  $G'$  defined in *Condition I* and theory  $T$ . Let us

here restrict  $G'$  and  $T$  to be a set of Horn clauses described by one head (conclusion) and body (conditions) for simplicity. The example of  $G'$  and  $T$  below is used only for this explanation of *completely defined*  $T$  in this subsection.

$$G' \text{ beer is consumed (city, date)} \\ \leftarrow \text{air temperature rises (city, date)}. \quad (48.5)$$

$$T: \text{beer is consumed (city, date)} \\ \leftarrow \text{air temperature rises (city, date),} \\ \text{people are rich (city)}. \quad (48.6)$$

$$\text{air temperature rises (city, date)} \\ \leftarrow \text{Typhoon is coming (city, date)}. \quad (48.7)$$

Then the relation  $T$  is *completely defined* for  $G'$  is defined as follows:

- i)  $G'$  is derived by  $T$  and existing data, that is, the head of  $G'$  should be included in the head of some clause in  $T$ , and all predicates in the body of  $G'$  should be in the body of some clause in  $T$ , with satisfying (ii) through (iv).

*Example:*  $T$  above derives (i. e., resolves)  $G'$  above if it is believed to be true that people are rich and Typhoon is coming as in (ii) below.

- ii) All clauses and bodies of clauses in  $T$  are supported by data if any, corresponding to DJs in  $DJcom(G)$ , and other clauses in  $T$ .

*Example:* *people are rich* and *Typhoon is coming*, as well as (48.6) and (48.7), must be supported by some data for deriving  $G'$ .

- iii) In each clause in  $T$ , all variables in the head must appear in some predicate in the body. This is the range restriction constraint in clause logic.

*Example:* *city* and *date* are included in the bodies in all clauses above.

Combining this condition with condition (ii), the theory comes to be based on data. That is, all variables appearing in predicates derived from  $T$  are certified to be in some DJ(s) in  $DJcom$ , as in (48.4).

- iv) All predicates in the body of each clause must share some variable(s) with some other predicates in the same clause, for the resolution in deriving a predicate in the head of a clause.

*Example:* *city* is shared between the two predicates in the first clause in  $T$  above. This variable should be also shared with *Typhoon is coming* in the second clause, to be resolved with *people are rich (city)* in deriving  $G'$ .

This condition is formalized as in (48.4), a point of which is that the names variables may differ across data, but the existence of common values in the

defined domain ranges is regarded as a reason for believing the two variables can be unified. Such a common variable is essential for enabling comparison or combination of data. However, the value common to variables, for example, a variable  $v_x$  in  $V_x$  and  $v_y$  in  $V_y$ , does not have to really exist in neither  $D_x$  nor  $D_y$ . This is because the analysis may result in finding sheer difference in the values of  $v_x$  and  $v_y$  rather than a pattern linking variables in  $D_x$  and  $D_y$  to be learned using common values of  $v_x$  and  $v_y$ . In this sense,  $[v]$  in (48.4) is defined as the domain range of variable  $v$ , not the values really taken in data.

Note we have at least two alternative subconditions for *Condition II-(ii)* as follows

$$\text{Condition II-ii(a):} \\ \exists D \subset D_a \cup D_b \cup \dots \cup D_L \{T \Rightarrow D\} \quad (48.8)$$

$$\text{Condition II-ii(b):} \\ D_a \cup D_b \cup \dots \cup D_L \Rightarrow T \quad (48.9)$$

*Condition II-(a)* means  $T$  explains some part of the data, and *Condition II-(b)* means  $T$  is a necessary theory for the data in  $DJcom$ . A noteworthy point is that the contents of data are not necessarily available in advance or in the initial state of IMDJ. In such a case, the data are analyzed after the data trading after workshop. And in the analysis, statistical validation for evaluating the support, that is, the extent of satisfying *Condition II-ii(a)*, or the confidence that is, the extent of satisfying *Condition II-ii(b)*, of the theory proposed in IMDJ is to be executed. If a theory – regarded as a glocal model in this chapter – is not acceptable according to the evaluation, the revision of the model will be considered, and new rules will be explored by data mining applied to data in  $DJcom$ .

### 48.5.2 Abductive Reasoning for Thoughts and Communications in IMDJ

Let us return to the example of (48.1) and (48.2), in order to present the process of reasoning to obtain satisfactory models. The breaking of (48.1) into (48.2), is to be realized here via the substitution below.

*Transforming goal  $G$  to  $G'$ :* put *beer\_consumption*  $\leftarrow ?$  *weather\_change* as in (48.10) by substituting *beer\_consumption* with *beer\_consumption (date)* and *weather\_change* with *air\_temperature\_change (date)*. This and the next step to describe (48.11) and (48.12) are to be executed by considering words' meaning.

$$G' : \text{beer\_consumption (date)} \\ \leftarrow \text{air\_temperature\_change (date)} \quad (48.10)$$

Definition of predicates in  $G'$ : for undefined constants  $\alpha$  and  $\beta$ ,

$$\begin{aligned} \text{air\_temperature\_change}(\text{date}) \\ := \text{air\_temp}(\text{date}) - \text{air\_temp}(\text{date} - 1) > \alpha K \end{aligned} \quad (48.11)$$

$$\begin{aligned} \text{beer\_consumption}(\text{date}) \\ := \text{beer\_percons}(\text{date}) > \beta 1 \end{aligned} \quad (48.12)$$

$G'$  is thus interpreted as (48.13)

$$\begin{aligned} \text{air\_temp}(\text{date}) - \text{air\_temp}(\text{date} - 1) > \alpha K \\ \leftarrow \text{beer\_percons}(\text{date}) > \beta 1. \end{aligned} \quad (48.13)$$

And, parameters  $\alpha$  and  $\beta$  in (48.11) are embodied using data below, brought in by experts of weather and of beer, respectively, modeled by functions as

$\text{air\_temp}(\text{date})$  :  
average daily air temperature for each day  $\text{date}$   
 $\text{beer\_percons}(\text{date})$  :  
average consumption of beer per consumer for date

As a result of validation on these data, parameters are substituted with values as  $\alpha = 2$  and  $\beta = 0.3$ , so that (48.2) is obtained as the solution, which is a completion of the theory as glocal model in (48.13).

As shown above, the initial and informal model  $G$  of requirement *globally* rules the overall knowledge processing from a general viewpoint, that is, without fixing variables in any data (such as  $\text{date}$ ) to use for reasoning. Then, the *local* models in (48.11) and (48.12) corresponding to data in local domains are chosen and applied, to connect  $\text{beer\_consumption}(\text{date})$  and  $\text{air\_temperature\_change}(\text{date})$  in  $G'$  of (48.10) via variable  $\text{date}$  and the clause created in the glocal model. As a result, the incompletely defined goal  $G$  is put into an expression of first-order logic (FOL) with defined predicates.

Here we should note the goal could not have been translated into such a well-defined FOL without having experts' knowledge for modeling their data. Also here we find inventors' reasoning to combine experts' models as in (48.11) and (48.12) worked in substituting words in the global model, that is, the informally suggested goal  $G$ , with predicates to be defined on variables in available data. Without this contribution of experts' knowledge, (48.2) and (48.13) may have lost the *completely defined* nature especially *Condition II-(iii)*, and put as in (48.14) due to choosing data on the superficial

similarity to terms in  $G$  and variables in data, without checking the meaning or values of variables.

$$\begin{aligned} \text{beer\_consumed}(\text{month}) \\ \leftarrow \text{weather\_condition}(\text{week}) \end{aligned} \quad (48.14)$$

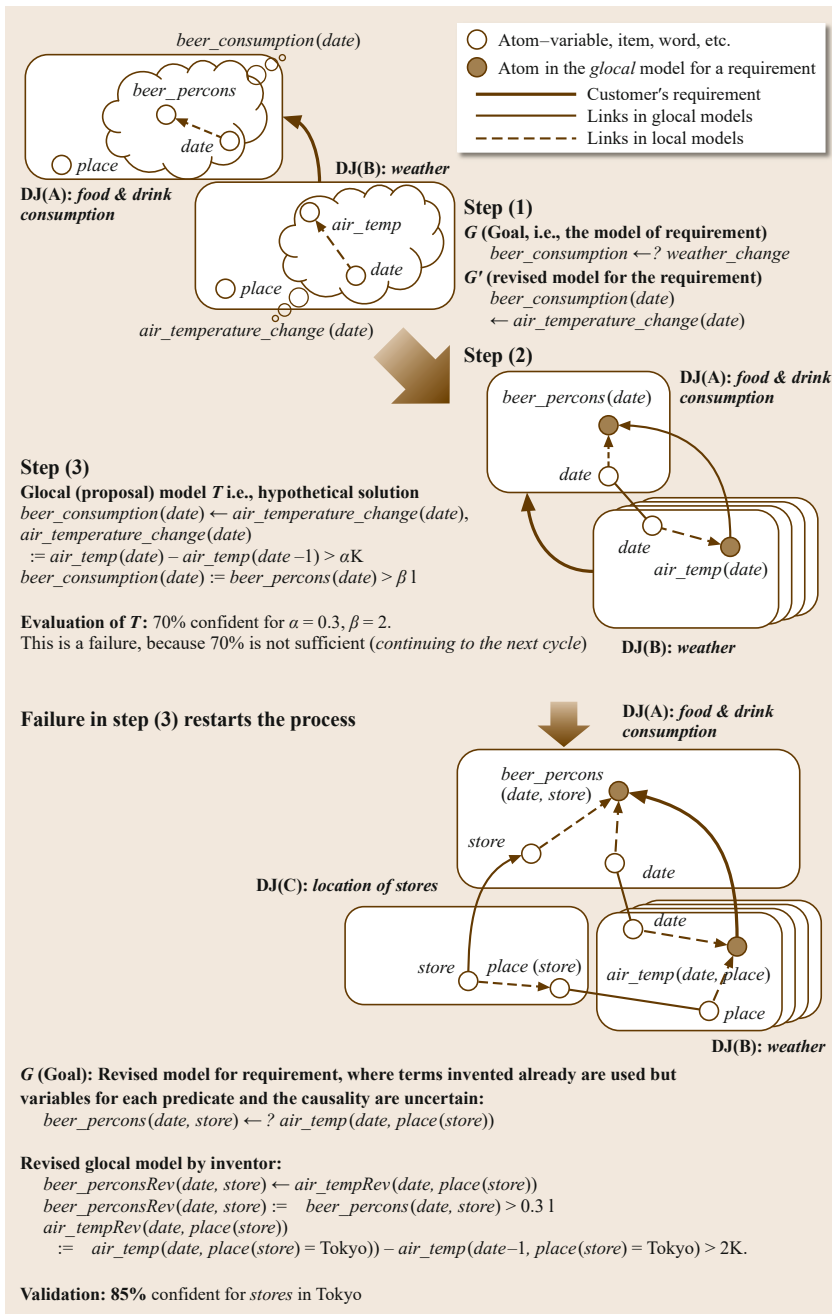
In (48.14), variables are not shared between the head and the body. Thus we can regard the inventors' models as *glocal*, that is, the result of interaction between the global model and local variables in  $V_G$  in data. On the other hand, in the successful example to derive (48.2) above,  $\text{date}$  worked as the common variable in  $V_G$  between the used two datasets, playing the role to connect data on weather and on beer consumption. The utility of this variable is evaluated and supported by comparing the values in  $\text{date}$  in the data including  $\text{air\_temp}(\text{date}1)$  and the data including  $\text{beer\_percons}(\text{date}2)$  to find the existence of common values between  $\text{date}1$  and  $\text{date}2$ , as meant by (48.4). As a result, it is confirmed that  $\text{date}$  can be in  $V_G$  and used for resolution of the two predicates.

As in this case, the requirement should be presented by a data user. Then, in order to translate the requirement into FOL, the elements in the requirement should be put into predicates or functions corresponding to variables in existing data. A set of data, including these variables and share variables such as  $\text{date}$  in the above example, should be collected here. For this, experts having models of data may be invited for presenting local models for these data, for using/collecting data corresponding to local models that may derive the global requirement model if combined by the glocal model of inventors' proposal. An alternative set of data may be proposed here, described as  $DJcom$ , if a set is not satisfactory for describing the requirement. This goal-driven process of reasoning can be regarded as abduction, and itemized as steps below. It is clear from this list that the organization of IMD provides with the reasoning and communication required for satisfying constraints proposed in the previous section.

### A Process of Abductive Reasoning with Communication for IMDJ

*Step 1* (goal setting): Users of data or of analysis results express requirement  $G$  as a model  $G'$  in FOL, with predicates invented with the assistance of experts. Here experts describe local models for DJs with inventing predicates (functions are translated into predicates as in (48.11) and (48.12)), reflecting the interpreted relations of attributes in each DJ, and make  $DJcom$  a set of DJs including attributes or predicates in  $G'$ .

*Step 2*: Inventors check the existence of common values of variables in different DJs in  $DJcom$  (for this,

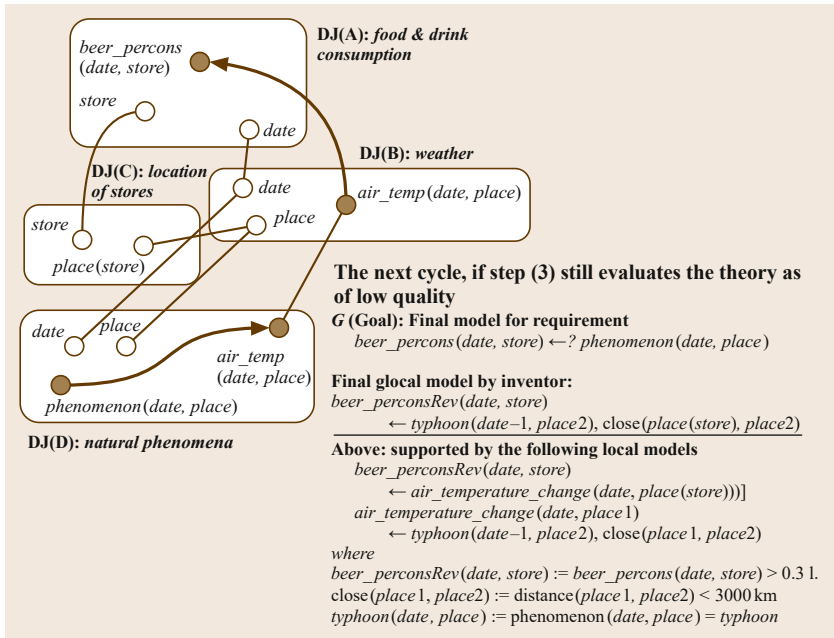


**Fig. 48.6** The abductive reasoning based on incomplete models in IMDJ

it is recommended that experts declare the defined domain range of variables in describing DJs in advance).

*Step 3* (abduction as the main step): Inventors explore hypothetical theory *T*, which derives *G'* and include only predicates and attributes obtained in steps so far.

If such a hypothetical theory *T* exists, *T* is evaluated subjectively on users' interest, or objectively on data, if available, corresponding to DJs of which attributes and predicates are in *T*. If the evaluation of *T* is higher than a predefined threshold, it is regarded as the solution. If the evaluation is lower, discard *T* and restart *Step 3*.



**Fig. 48.7** An abductive reasoning, where local models are added (continuing from Fig. 48.6)

Otherwise (if no hypothetical theory  $T$  derives  $G'$ ), call experts for adding DJs (including attributes missed for deriving  $G'$ ), and restart from *Step 1* since the failure to derive  $G'$  means the global model is not supported by the current local models.

Thus, each hypothetical theory  $T$  turns out to be a global model combining local models obtained in *Step 3* to explain  $G'$  of *Step 1*. The criteria for the evaluation of theories on data corresponding to DJs, in *Step 3*, should differ for different goals. For example, data on beer consumption per day in all districts will be desired for validating (48.13) precisely, if covering all districts by beer market is desired. For this purpose, a DJ for inventing  $T$  and the data for evaluating  $T$  should include *address of consumer* as a variable. On the other hand, this variable will not be reflected neither to invention nor evaluation of  $T$ , if the goal is to check the total consumption of beer all over a country. In a different case, the user may evaluate the novelty of  $T$ , in spite of its low precision and small coverage in the market, if he seeks to discover a clue for creating new market of beer.

Figure 48.6 shows a modification of the last example, in that the model once obtained in *Step 3* is here as insufficient (70% of confidence) with substituting  $\alpha$  with 2 and  $\beta$  with 0.3. This is superficially a failure but is a case of requirement externalization, in that the result of data analysis made the user notice his/her own latent desire to predict beer consumption by higher confidence.

As a result, participants will consider to add data about the correspondence between liquor store and its location given by the variable *place*, so that expert(s) owning such data creates a new function  $place(store)$  for explaining his knowledge about additional data (the latter half of *Step 3*). This can be regarded as an invention made because the insufficient confidence as a result of the previous analysis is attributed to the insufficient consideration of attributes, which may explain beer consumption, such as the liquor store selling beer. By adding the attributes, that is,  $(date, store)$  or with  $(date, place(store))$  in *Step 3*, participants obtain a model as more confident proposal, as in the last step of Fig. 48.6. Here, the store selling beer and its place are connected due to adding DJ. This case shows the nonmonotonic nature of abduction in the procedure.

Figure 48.6 is still a simplistic example. More generally we may iterate abductive reasoning further. For example, additional data about extraordinary natural phenomena including disasters, may be used as in Fig. 48.7 for learning such a causality as *if typhoon is active in China, the air temperature in Tokyo increases*. This causality will be obtained by finding a corresponding pattern from data inventing predicate *typhoon* as

$$\begin{aligned}
 &air\_temperature\_change(date, Tokyo) \\
 &\leftarrow typhoon(date - 1, China) \tag{48.15}
 \end{aligned}$$

Furthermore, by inventing predicate *close* with collecting facts for other places than Tokyo and China, the goal



may be revised to (48.16), meaning one should be concerned about the influence of natural phenomena such as typhoon for predicting the beer consumption per consumer

$$\begin{aligned} &beer\_perconsRev(date, store) \\ &\leftarrow typhoon(date - 1, place2), \\ &close(place(store), place2). \end{aligned} \quad (48.16)$$

## 48.6 Conclusions

In the Innovators Marketplace on Data Jackets (IMDJ), small pieces of information called Data Jackets containing abstracts of confidential data, that is, data that cannot be disclosed, are submitted. Then the process of IMDJ goes as a gamified workshop where participants create and propose ideas to combine and/or analyze data corresponding to DJs. The ideas are evaluated, on the matching of requirements and the utility of the expected knowledge or action scenario to be obtained from the proposed analysis.

This chapter proposed a formalization of the process of IMDJ as of abductive reasoning, and derived a proposal of refined process of the workshop to iterate cycles of steps for communication, reasoning, and sharing DJ. By this formalization, the link of IMDJ to a logical framework in model-based reasoning and to a feasible procedure of innovation came to be clearer by introducing the co-evolution of global, glocal, and local models. These models correspond respectively to the requirements of data users, the proposals of inventors of analysis plans, and the knowledge of experts of data domains.

Such a logical formalization from the viewpoint of model-based reasoning is of high use value, in two

The real intension of user may have been to validate an idea like (48.16), rather than the uttered one as in (48.1). On the way of communication in IMDJ, such a real intention tend to be externalized. As in these examples, participants concretize local and glocal models from goals given as abstract global models and various datasets, by revising and combining local models in the abductive reasoning to derive the goal.

senses, that is, (a) the presented three steps are expected to save IMDJ organizer in explaining the procedure without ambiguity, (b) we may consider in the future to automate a part of the whole process with developing computational methods to cope with the complexity of nonmonotonic reasoning [48.55] such as abduction.

In the future, our direction will be to realize logical IMDJ, mainly in the sense of (a) above, toward humans' logical and creative data-based decision making. An essential problem addressed to the future is how we can support a beneficial real act of participants of IMDJ. That is, they tend to invent new, rather than just adhering to given, requirements as pointed out in the last section, motivated by the communication in the workshop. The frequency of this act cannot be explained by just adding Step 0 before the whole procedure presented in this chapter, nor should be suppressed just for following this procedure because this act results in inventing data reuse plans with novelty and utility.

**Acknowledgments.** This study has been supported by JST CREST, and discussions with Kozo Keikaku Engineering Inc. and other major collaborators have been reflected to this chapter.

## References

- |  |   |
|--|---|
| <p>48.1 J.A. Schumpeter: <i>Theorie der wirtschaftlichen Entwicklung</i> (Duncker &amp; Humblot, Leipzig 1912)</p> <p>48.2 E.M. Rogers: <i>Diffusion of Innovations</i>, 5th edn. (Free Press, New York 2003)</p> <p>48.3 E. von Hippel: <i>Democratizing Innovation</i> (The MIT Press, Cambridge 2006)</p> <p>48.4 P.F. Drucker: <i>The discipline of innovation</i>, Harv. Bus. Rev. <b>63</b>(3), 67–73 (1985)</p> <p>48.5 P.F. Drucker: <i>Innovation and Entrepreneurship</i> (Harper Collins, New York 2006)</p> <p>48.6 Y. Ohsawa, P. McBurney: <i>Chance Discovery</i> (Springer, Heidelberg 2003)</p> <p>48.7 Y. Ohsawa, M. Usui: Creative marketing as application of chance discovery. In: <i>Chance Discoveries</i></p> | <p><i>in Real World Decision Making</i>, ed. by Y. Ohsawa, S. Tsumoto (Springer, Heidelberg 2005) pp. 253–272</p> <p>48.8 Y. Ohsawa, H. Fukuda: Chance discovery by stimulated group of people – An application to understanding rare consumption of food, <i>Conting. Crisis Manag.</i> <b>10</b>(3), 129–138 (2002)</p> <p>48.9 M. Donaldson: <i>Human Minds: An Exploration</i> (Allen Lane, Penguin, New York 1992)</p> <p>48.10 B. Dervin: An overview of sense-making research: Concepts, methods and results. Paper presented at the annual meeting of the Int'l Com. Association. (Dallas, 1983)</p> <p>48.11 B. Dervin: From the mind's eye of the user: The Sense-Making qualitative-quantitative methodol-</p> |
|--|---|

- ogy. In: *Qualitative Research in Information Management*, ed. by J.D. Glazier, R.R. Powell (Libraries Unlimited, Santa Barbara 1992) pp. 61–84
- 48.12 T. Senator: Evidence Extraction and Link Discovery Program, at DARPA Tech 2002, <http://w2.eff.org/Privacy/TIA/feeld.php> (2002)
- 48.13 Y. Ohsawa, Y. Nishihara: *Innovators' Marketplace: Using Games to Activate and Train Innovators*, Understanding Innovation (Springer, Heidelberg 2012)
- 48.14 K. Kira, L.A. Rendell: A Practical approach to feature selection, Proc 9th. Int. Workshop Mach. Learn. (Morgan Kaufmann, San Francisco 1992) pp. 249–256
- 48.15 H. Zhou, T. Hastie: Regularization and variable selection via the elastic net, J. Real Stat. Soc. B. **67**(2), 301–320 (2005)
- 48.16 M. Joshi, V. Kumar, R. Agrawal: Predicting rare classes: Can boosting make any weak learner strong?, 8th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining (KDD'02) (2002)
- 48.17 G.M. Weiss, H. Hirsh: Learning to predict rare events in event sequences, Proc. 4th Int. Conf. Knowl. Discov. Data Mining (KDD-98) (AAAI Press, Menlo Park 1998) pp. 359–363
- 48.18 E. Suzuki, T. Watanabe, H. Yokoi, K. Takabayashi: Detecting interesting exceptions from medical test data with visual summarization, Proc. 3rd IEEE Int. Conf. Data Min. (2003) pp. 315–322
- 48.19 K. Yamanishi, J. Takeuchi, G. William, P. Milne: Online unsupervised outlier detection using finite mixture with discounting learning algorithms, Data Min. Knowl. Discov. **8**(3), 275–300 (2004)
- 48.20 Y. Ohsawa, N. Matsumura, N. Okazaki, A. Saiura, H. Fujie: Mining scenarios for hepatitis B and C. In: *Multidisciplinary Approaches to Theory in Medicine*, ed. by R. Paton (Elsevier, Amsterdam 2005)
- 48.21 N.N. Taleb: *The Black Swan: The Impact of the Highly Improbable* (Random House, New York 2007)
- 48.22 L. Magnani: Chance discovery and the disembodiment of mind. In: *Knowledge-Based Intelligent Information and Engineering Systems*, Springer LNAI, Vol. 3681, ed. by R. Khosla, R.J. Howlett, L.C. Jain (Springer, Heidelberg 2005) pp. 547–553
- 48.23 L. Magnani: Abduction and chance discovery in science, Int. J. Knowl.-Based Intell. Eng. Syst. **11**(5), 273–279 (2007)
- 48.24 O. Eris: *Effective Inquiry for Innovative Engineering Design* (Kluwer Academic Publishers, Boston 2004)
- 48.25 A. Abe: Abduction and analogy in chance discovery. In: *Chance Discovery*, ed. by Y. Ohsawa, P. McBurney (Springer, Heidelberg 2003) pp. 2231–2247
- 48.26 J. Nakamura, Y. Ohsawa: Shift of mind: Introducing a concept creation model, Inf. Sci. **179**(11), 1639–1646 (2009)
- 48.27 H. Terai, K. Miwa: A chance favours a prepared mind: Chance discovery from cognitive psychology. Studies in computational intelligence. In: *Advances in Chance Discovery*, Vol. 423, ed. by Y. Ohsawa, A. Abe (Springer, Berlin 2013) pp. 33–48
- 48.28 D. Bergner, O. Eris: Reframing the data mining process. In: *Data Mining for Design and Marketing*, ed. by Y. Ohsawa, K. Yada (CRC Press, Boca Raton 2008) pp. 19–34
- 48.29 Y. Ohsawa, N.E. Benson, M. Yachida: KeyGraph: Automatic indexing by co-occurrence graph based on building construction metaphor, Proc. Adv. Digit. Libr. Conf. (IEEE ADL'98) (1998) pp. 12–18
- 48.30 R. Fruchter, Y. Ohsawa, N. Matsumura: Knowledge reuse through chance discovery from an enterprise design-build enterprise data store, New Math. Nat. Comput. **1**(3), 393–406 (2005)
- 48.31 X. Llorà, D.E. Goldberg, Y. Ohsawa, N. Matsumura, Y. Washida, H. Tamura, M. Yoshikawa, M. Welge, L. Auvil, D. Searsmith, K. Ohnishi, C.-J. Chao: Innovation and creativity support via chance discovery, genetic algorithms and data mining, New Math. Nat. Comput. **2**(1), 85–100 (2006)
- 48.32 C.F. Hong, H.F. Yang, M.H. Lin, G.S. Lin: Creative design by bipartite keygraph based interactive evolutionary computation, Lect. Notes Comput. Sci. **4253**, 46–56 (2006)
- 48.33 E.M. Goldratt: *Essays on the Theory of Constraints* (North River Press, Great Barrington 1987)
- 48.34 J.M. Carrol: *Making Use: Scenario-Based Design of Human-Computer Interactions* (The MIT Press, Cambridge 2000)
- 48.35 N. Kushiro, Y. Ohsawa: A scenario acquisition method with multi-dimensional hearing and hierarchical accommodation process, New Math. Nat. Comput. **2**(1), 101–113 (2006)
- 48.36 M. Polanyi: *The Tacit Dimension* (Routledge, London 1966)
- 48.37 I. Nonaka: A dynamic theory of organizational knowledge creation, Organ. Sci. **5**(1), 14–37 (1994)
- 48.38 E. von Hippel: "Sticky Information" and the locus of problem solving: Implications for innovation, Manag. Sci. **40**(4), 429–439 (1994)
- 48.39 S. Ogawa: Does sticky information affect the locus of innovation? Evidence from the Japanese convenience-store industry, Res. Policy **26**(7/8), 777–790 (1998)
- 48.40 Y. Ohsawa, M. Akimoto: Unstick tsugoes for innovative interaction of market stakeholders, Int. J. Knowl. Syst. Sci. **4**(1), 32–49 (2013)
- 48.41 H. Plattner, C. Meinel, L. Leifer (Eds.): *Design Thinking – Understand – Improve – Apply* (Springer, Heidelberg 2011)
- 48.42 S. Mohammed, E. Ringeis: Cognitive diversity and consensus in group decision making, Org. Behav. Hum. Decis. Process. **85**(2), 310–335 (2001)
- 48.43 E. Gottesdiener: *Requirements by Collaboration: Workshops for Defining Needs* (Addison-Wesley Professional, Boston 2002)
- 48.44 Y. Ohsawa: Data crystallization: Chance discovery extended for dealing with unobservable events, New Math. Nat. Comput. **1**(3), 373–392 (2005)
- 48.45 L. Hambleton: *Treasure Chest of Six Sigma Growth Methods, Tools, and Best Practices* (Pearson Education, Upper Saddle River 2007)
- 48.46 Y. Ohsawa, H. Kido, T. Hayashi, C. Liu: Data jack-ets for synthesizing values in the market of data, Procedia Comput. Sci. **22**, 709–716 (2013)

- 48.47 Y. Ohsawa, C. Liu, T. Hayashi, H. Kido: Innovators marketplace on data jackets for externalizing the value of data via stakeholders' requirement communication, AAAI 2014 Spring Symp. Big data becomes personal: Knowledge into Meaning (2014) pp. 45–50
- 48.48 G.I. Piatetsky-Shapiro, KDnuggets: *Datasets for Data Mining and Data Science*, <http://www.kdnuggets.com/datasets/index.html>
- 48.49 Microsoft Azure: Windows Azure Marketplace, <https://datamarket.azure.com/>
- 48.50 The Open Knowledge Foundation and CKAN Association: The Open Source Data Portal Software, <http://ckan.org/>
- 48.51 Japan's Ministry of Internal Affairs and Communications, Administrative Management Bureau (AMB): <http://www.data.go.jp>
- 48.52 D. Gentner: Structure-mapping: A theoretical framework for analogy, *Cogn. Sci.* **7**, 155–170 (1983)
- 48.53 M.E. Taylor, P. Stone: Transfer learning for reinforcement learning domains: A survey, *J. Mach. Learn. Res.* **10**, 1633–1685 (2009)
- 48.54 Laboratory of Chance Discovery, The University of Tokyo: Entry sheet for Data Jackets, <https://sites.google.com/site/datajackets/>
- 48.55 Y. Ohsawa, M. Ishizuka: Networked bubble propagation: A polynomial-time hypothetical reasoning method for computing near-optimal solutions, *Artif. Intell.* **91**, 131–154 (1997)

# 49. Models in Pedagogy and Education

Flavia Santoianni

Pedagogy is a discipline concerned with theories and practices of education. Its epistemological model is complex. It may be considered as qualified by two structural directions: *pluralism* and *dialecticity*.

The pluralism of pedagogy is represented by its possible theoretical routes, by the different levels of sharing of disciplinarity and by a multiplicity of aspects. It involves empirical and experimental research, historical and philosophical dimensions, and epistemological and metatheoretical lines. The theoretical plurality of pedagogy concerns subjects, ages and places of education, languages and research methods, and actual directions and interpretative issues. The multidisciplinary plurality of pedagogy distinguishes it in pedagogical sciences, educational sciences, and educational developmental sciences. The disciplinary multiplicity of pedagogy is expressed by the diversity of pedagogical sciences that belong to general pedagogy. Even if pedagogical sciences are multiple, social pedagogy, history of pedagogy and special needs education are disciplines specifically related to the field of pedagogy.

The dialecticity of pedagogy expresses its controversial nature divided between science and philosophy. The scientific approach to pedagogy

<b>49.1 Pluralism</b> .....	1034
49.1.1 Theoretical Plurality .....	1034
49.1.2 Multidisciplinary Plurality .....	1035
49.1.3 Disciplinary Multiplicity .....	1036
<b>49.2 Dialecticity</b> .....	1039
49.2.1 Science and Philosophy .....	1039
49.2.2 Theory and Practice .....	1040
<b>49.3 Applied Models</b> .....	1042
49.3.1 Traditional Models .....	1042
49.3.2 Actual Models .....	1044
49.3.3 Experimental Models .....	1046
<b>49.4 Conclusions</b> .....	1048
<b>References</b> .....	1048

evolves from systematicity to complexity. It develops, namely, in parallel with the construction and the reconstruction of the very idea of science. The systematization of educational sciences strengthens the philosophical role of pedagogy. The so-called *identity crisis* of pedagogy will bring it to rediscover the sense of its own reflexive intentionality. The relationship between theory and practice makes pedagogy a science of education, in particular a theory of educational development processes.

Pedagogy is a discipline concerned with theories and practices of education. The etymological meaning derives from the Greek language and means to guide ( $\alpha\gamma\epsilon\iota\nu$ ) the child ( $\pi\alpha\iota\varsigma$ ), that is, education. Pedagogy today relates to all the ages of humans and to a plurality of relational contexts.

From an epistemological viewpoint, the discipline is characterized by a complex model. As introduced by Cambi [49.1], this model shows two structural directions: pluralism and dialecticity.

## 49.1 Pluralism

The pluralism of pedagogy is represented by its possible theoretical routes, by the different levels of sharing of disciplinarity and by a multiplicity of aspects.

From a pluralistic viewpoint, it is possible to distinguish three coexisting aspects:

1. Empirical and experimental research, related to the model of the hard sciences.
2. Historical and philosophical dimensions, anchored to specific values and perspectives.
3. Epistemological and metatheoretical lines, which express the critical and regulative role of the discipline.

### 49.1.1 Theoretical Plurality

Pedagogical theoretical routes concern a plurality of subjects, the ages of education, the places of education, the languages, the research methods, and several interpretative directions.

#### Subjects, Ages, and Places of Education

The subjects of education have individual biological, psychological, and relational differences [49.2], which modulate their own ways of learning in a wide range of possibilities including cognitive discomfort and specific learning disorders [49.3, 4]. There are studies on gender [49.5] and social differences [49.6] regarding economic, environmental, and cultural deprivation. An emerging field concerns ethnic, linguistic, and cultural differences, which triggered intercultural studies [49.7, 8].

Learning is seen as an on-going process throughout the course of an individual lifespan [49.9]. For this reason, pedagogy concerns multiple ages: basic education for children and youths until university, with the processes of literacy and schooling; continuing education for adults, and educational proposals for seniors. Education is about the cycles of life – in a pedagogical and andragogical perspective – to implement lifelong learning, continuing education, adult education, and theories of adulthood [49.10]. The methodology of these approaches is based on the analysis of experiential contents and on the encouragement of self-management of learning processes. The teacher as a facilitator becomes an organizer of resources [49.11]. Adult education is continuous because it includes different stages of individual life. Adult education encourages change, through transformation and reorganization impulses, for a constant renewal of the person.

The places of education interact with individual and social development [49.12]. Family is the context of pri-

mary socialization, where learning is both explicit and implicit. School is the place of formal education and of self/guided reflection, even if the teaching and learning relationship can be influenced by implicit learning also here [49.13]. Moreover, institutions and associations that manage cultural activities – such as museums, libraries, and theaters – are considered places of nonformal education [49.14].

Also, technologies may be interpreted as a context of education [49.15]: for example, the instruments of social communication, as traditional media (cinema, television, radio, telephone) and electronic media (computer and social networks).

#### Languages and Research Methods

Pedagogy uses different languages. Explicative, analytical, and descriptive language – preferred by scientific pedagogy – coexists with narrative, and interpretative language by which personal stories are analyzed. The philosophical language of axiological nature defines the epistemological scope and the content value of the discipline, while everyday life and common sense language is usually utilized to solve concrete problems. Finally, there are the nonverbal languages [49.16], as the proxemics of communication, and the disciplinary languages [49.17], specifically used by the disciplinary didactics.

Research in education is multifaceted: it may be theoretical research, basic, formal, and structural research, which deals with the relationship with other sciences; at the same time it may be experimental research, which uses observation and measurement to improve teaching practices. It may be historical research, which deals with the evolution of educational institutions, and comparative research, which compares different territorial and national models of pedagogical organization.

Methods of educational research are several: nomothetic, idiographic, experimental, grounded theory, action research, historical, comparative, narrative, and clinical methods. They are quantitative and qualitative [49.18], they connect the scientific and philosophical areas of the discipline, and are often considered interpretative paradigms more than scientific methods.

#### Directions and Interpretative Issues

Pedagogy discusses either recurring and structural issues or emerging issues. Among the recurring and structural issues is the question of the relationship between education, instruction, training, and educational development: what are the mutual influences? A related issue is the question of the relationship between means and ends: which are predominant, from time to time, in

different historical situations? Another theme is related to the critique of educational institutions: what is their role and their more effective work?

Also, there are issues that affect pedagogical antinomies. These can be formal – science and philosophy, theory and practice, knowledge and skills, technology and art. Or these can be theoretical/practical – authority and freedom, culture and professional training, rupture and continuity in education. Pedagogical antinomies can also be practical/educative, such as those concerning the relationships between teacher and students. Other issues relate to the set of educational skills: which one should be the better training? Moreover, the links between pedagogy and politics: what relationships subsist with ideology?

Emerging issues relate to forefront issues, such as how to define the concept of educational development, how to prepare the young on the complexity of the current situations, and how to educate them to respect differences, especially in multicultural societies. Also, how to discuss the concept of value in postmodern societies, and how to dig deeper on the topic of subjectivity and its global interpretation through the concept of the persona, to focus on its role in education and school.

This entails recognizing the relationship between body and mind, and the cognitive, affective and organismic dimensions that are intertwined according to a comprehensive and holistic vision. Another aspect is to implement guidance procedures in learning pathways, within a process of care and taking charge of the learner.

In this sense, the role of the school is significant: how can and should it be autonomous and selective, how should it be influenced by technology? Which positions should be taken towards students in situations of great difficulty towards learning: for instance, the meaning of rejection of school. How about the intention of educating for community participation and critical thinking? Moreover, the openness and inclusion towards students in trouble, with specific learning disabilities or with physical disabilities, as well as towards students from other cultures.

Culture is another topic of interest: to recognize the role played by culture in the school, to affirm a pluralistic and dialectic culture, to educate the young towards a crosscultural mind, to recognize the role played by humanistic knowledge as by scientific and technological ones.

This discourse includes a reflection on communication technologies. The on-going debate on information and communication technology today asks not just what are the chances of access to knowledge through networks of communication, but rather how pedagogy can

render effective the opportunities for learning and sharing in a relational network from an educational point of view. Communication technologies may alter the space-time dimension, so pedagogy needs to reflect on the meaning of virtual experiences. Moreover, these experiences should be joined to traditional routes. New methods of distance education are studied and the criteria by which e-learning can be effective for learning is expanded. Media education explores the use of media and even their social, psychological, and communicative power and their possible individual and collective influences.

Particular attention is given to the concept of freedom, seen both as subjective realization, personal emancipation and as a competent and active citizenship in a supportive community. Finally, attention to the environment is given either as research on learning environments design, in all its diversity, or as ecological education.

### 49.1.2 Multidisciplinary Plurality

Of how many domains may pedagogy consist? The knowledges that compose it, or which are around it, can be distinguished in pedagogical sciences, educational sciences, and developmental sciences [49.19].

#### Pedagogical Sciences

Pedagogy is divided into pedagogical sciences, which all have general pedagogy underlying. These sciences can be numerous and are constantly developing, increasing according to the specialization of knowledge; in particular: social pedagogy, history of pedagogy, special needs education, didactics and experimental pedagogy. Experimental pedagogy, such as didactics, is seen as a theoretical, methodological, and practical framework in a close relationship with general pedagogy [49.20].

#### Educational Sciences

Educational sciences are disciplines that study the general and local conditions of education and the educational situation; they accomplish a general reflection on education [49.21]. Educational sciences include the historical, epistemological, psychological, sociological, and didactical fields of investigation. Therefore, they include disciplines such as history of education, educational epistemology, psychology of education, sociology of education, anthropology of education, and teaching.

In this interpretative framework, philosophy of education can be considered at the border of educational sciences, because it plays a regulatory and mediating role of a metareflective nature in relation to sciences of

education. Hence, it supports general education from an epistemological point of view.

### Educational Developmental Sciences

Educational developmental sciences represent a *trait d'union* with other disciplines that, in the humanities or in the natural sciences, study the development of the human species.

What relationships are between pedagogy and other areas of knowledge? What are the relationships that pedagogy develops with other disciplines [49.22] and which fields and levels of these relationships may be concerned [49.23]?

The pluralistic nature of pedagogy leads it to borrow research objects and methods of investigation from other disciplines; for this reason, sometimes there are chances of confusing its disciplinary identity with other ones. Relations between the different types of knowledge can be multidisciplinary, interdisciplinary, and transdisciplinary [49.24]; these are connections of increasing interdependence.

The relationship between pedagogy and science of education should be multidisciplinary. In this case, different disciplines, with different epistemological stances, with various research methodologies and sectorial operational tools, can share research hypothesis of mutual interest and then collaborate together without losing their disciplinary specificity. Pedagogy is a model of science in itself [49.25].

### 49.1.3 Disciplinary Multiplicity

The multiplicity of aspects that compose the educational pluralism can be represented by the diversity of pedagogical sciences that belong to general pedagogy. Pedagogical sciences are multiple; some ministerial subdivisions refer to the social pedagogy, history of pedagogy and special needs education as disciplines specifically related to the field of pedagogy.

#### Social Pedagogy

Social pedagogy applies the educational theories in operational form to various social contexts and reflects theoretically on these issues [49.26]. Its role is to weld the educational activities in the territories, within their local dynamics, in a transformative and emancipatory way for all the partners involved and the communities to which they belong.

The task of the social pedagogy is to build an integrated educational system between the school and the nonformal agencies such as the family, the church, the working world, the associations, the local authorities, and the informal systems, such as multimedia networks.

Social pedagogy uses a systemic-relational model of education that allows the placing of local interpretative patterns within more global visions. Through this model it manages to connect theory to practice and vice versa, by reformulating from time to time, in a participatory way, practices and theories in relation to the educational needs of the communities in which they are highlighted.

The disciplinary knowledge cannot be separated from the contexts but rather should be able to understand their specificity, the existential experiences that develop in them, and their social emergencies through participatory action research [49.27].

#### History of Pedagogy

History of pedagogy also relates to the history of educational institutions. After the 1970s of the twentieth century, the history of pedagogy gradually interacts with the history of education, since this includes more detailed attention towards the educational systems [49.28].

History of pedagogy studies how the concept of education has developed in the course of the history of humanity [49.29]. The diachronic dimension of the concept of education has developed in the transition from the Greek *paideia*, the free human development through culture, with aims of universal validity, to the *bildung* of the nineteenth century, unitary and not fragmented education, linked to the composition of several domains, from science to art, through which subjects shape their own image, *bild*.

In the Greek *paideia*, education takes place according to perspectives of universalization and interior harmonization through culture. Although there is a dualism of educational models, because intellectual work and manual labor are separated, *paideia* represents the ideal of the human, the global expression of the individual in its full manifestation. The *paideia* embodies the encyclopaedic tension of the classical world, the openness to all knowledge, the idea of humanities as areas of study. Education consists both in the personal relationship between the teacher and the student, both in the competitive tension which affects the corporeal dimension, the physical appearance, shaping itself as practical education within the *polis*, a pedagogical community in a global sense.

In the Greek enlightenment, culture becomes more critical, technical, scientific, and democratic. *Paideia* becomes more attentive to human problems and to the techniques of speech, to the use of words. Education is moral, rhetorical, and historical, increasingly tense to the principle of *kalokagathos*, the beautiful and the good, the cultivation of the specific aspects of humanity through the study.

In Socrates' thought, education is considered as *episteme* and not just as *ethos* or as *praxis*. Education therefore becomes *paideia*, the universalization of the subject through rational discussion, education of mankind through culture and civilization – maieutical, dialogical, and dialectical. An education in which the *humanitas* is the result of a focused education to the *know thyself*. The range is from a pragmatic to a theoretical dimension of educating.

In Platonic *paideia*, education of the individual under the control of reason takes place through the contemplation of ideas (the man is imprisoned in the *cave* of the body and of the *doxa*, of the opinion). The aim of the *paideia* is the recognition of the spirituality of the soul and of its contemplative identity; virtue is identified with knowledge. Political *paideia* is born, which is activated according to the different social classes and their different ways of education (technical learning for manufacturers, training courage for warriors, dialectical education for rulers).

*Paideia* embodies a blend of *musaica* education (gymnastics, poetry, and music) with literary education, thus representing an overall process. However, in this time arises the dichotomy between a rational and philosophical model of education, regarded as superior, and a lower model – technical, professional, and productive – of education. This dichotomy will be destined to last in Western society.

In Aristotelian *paideia*, men realize themselves according to their own shape, defined by the intellectual activity (*nous*). The aim of education is to achieve the virtue of wisdom through the educational mastery and the control of the body. Man is a social animal that develops in a concrete and real way.

In Roman civilization, education becomes literary education but even moral education, civic virtue, respect for tradition, for *patria potestas* and *res publica*. The educational model is based on the values of heroism, dignity, and courage; also piety, and sacrifice. Education acquires the rhetorical character of *humanitas*, the study of the liberal arts, *Humanæ litteræ*, as an introduction to the virtues of the orator, an ideal figure from an ethical standpoint.

The Latin *humanitas* switches to Middle Ages *paideia* Christi. After the conquest of Greece and the contact with the Hellenism, the Roman culture is transformed and acquires the Hellenistic character of education as self-care, self-control, inner balance, literary culture, use of the word, and consciousness of tradition. Ties with the Republican custom become loose: education is about man as universal expression of humankind and not just as a citizen.

By the Hellenistic influence, pedagogy gradually frees itself from the *ethos*, until the *romanitas* model

(which is based on the recognition of the state, on the rights, and on an hegemonic and universal culture) will be integrated with the educational model of the Christian culture, which advocates political and social values, different and *revolutionary* than the classic ones such as equality, solidarity, and humility.

*Enkyklios paideia* coexists with the educational vision of religion in a blend that lasts until the Middle Ages, when Christian *paideia* and its educational program, related to the Christian message – to its values of interiority, sublimation, transcendence – will overshadow the classical culture.

Middle Ages *paideia* turns into the humanistic and renaissance model of education, which is rooted in a man-centered secular worldview, whose author is *Homo faber* (either as an individual, and social subject). The key concepts are those of freedom, progress, emancipation, and rationalization.

Humanistic *paideia* returns to classical *paideia*: it becomes free human education in touch with culture and social life. Humanistic *paideia* resumes its aesthetic and scientific aspects from classical *paideia*, as its expression in the arts, its development in the techniques, its statement of thought as the study of the living and of the existing in a multiplicity of fields (from *created* to *constructed*), and its capacity to change reality.

The educational model is political; it revolves around a secular and civil education, worldly in its character, in controversy against subordination to theology, to encyclopaedism and formalism. It is, however, an aristocratic education, not directed to the middle and popular classes as, instead, the movement of the Reformation that encourages an autonomous approach to culture and its dissemination through a personal reading of the holy texts.

Pedagogy of the counter reformation promotes educational curricula inspired by a rhetorical and grammatical vision of education, characterized by detailed precepts and by strict rules. For instance, the *ratio studiorum*, the planning of educational activities governed by strict regulations, in line with the ethical and religious aims of the order of the Jesuits, whose cultural and educational model converges to the political and social models expressed by religious and civil authority.

In the late renaissance, education becomes more and more tense towards an encyclopaedic vision of the knowledge and becomes, in the late sixteenth and early seventeenth century, utopian pedagogy. Pedagogy aims to correlate a harmonic model of man, proper to humanistic pedagogy, with the design of ideal societies, in which the individual is a civil and social subject. At the same time, the seventeenth century is the century of the *new science*, in which the scientific method is born



and opens up the way for a rigorous foundation of pedagogy through a problematization of the methods and an experimental approach to practice.

In the eighteenth century, the concepts of education and socialization prevail over fragmentation of educational models and education is defined in a social and scientific sense. The bourgeois vision is affected by the political and cultural reformism of the secularized society: education acts as social homologation and a strengthening of civic consciousness for a promotion of the rational, free from prejudices and beliefs.

The modernity remarks the gradual emancipation of pedagogical knowledge from metaphysics – its progressive scientific investment and its consequent relativization – as confirmed by the links with politics and ideology.

The sociopolitical paradigm in pedagogy can be said to be characterized by the social philosophy of education, with ethical and political objectives, by the oscillation between the role of education on policy and vice versa [49.30], and by an historical, critical, and hermeneutical accuracy. The theory of education uses the social commitment as a model of the character of design of pedagogy, for a renewal of society; so it gradually moves away from any philosophy of education that reflects the dominant culture.

In the nineteenth century, *paideia* comes back as German *bildung* and its full educational expression. This century shows the link between education, society, ideology, and politics early but also between art and pedagogy; from an epistemological point of view, pedagogy becomes a rigorous, experimental, *positive*, and autonomous domain at the end of the century. The positivist scientific paradigm indeed brings pedagogy to social and institutional rationalization tasks.

In *bildung*, the human and cultural education interconnects individuality, science, and art to a model of man that elaborates its own internal shape, its image, *bild*. *Bildung* thus recovers the meaning of *paideia* as unification of knowledge against the fragmentation of culture and the influence of technology.

During the twentieth century, the deconstruction of the subject and the specialization of knowledge led to a rethinking of the ideal concept of *paideia*. The many directions taken by pedagogy share the idea of a new model of *bild* – synergistic, developmental, and dysmorphic.

Pedagogy is now *disenchanted*: it shares the issues related to the postmodern societies and attempts to address them. At the same time it does not give up the project of construction of identities and of re-

thinking of subjectivity. The involved categories are: deconstruction, interpretation, planning, responsibility, communication, and solidarity.

Through these categories, scientific and philosophical pedagogy find their meeting point in the idea of the subject as a person. Pedagogy becomes the science of the personal education, in the society of diffused knowledge.

The social dimension becomes collective sharing, solidarity in social practices, motor of local and global development that supports the individual, continuous and never predictable processes of internalization of knowledge. The postmodern neo-*bildung* is therefore a *bildung* without *bild* [49.31, p. 51].

### Special Needs Education

Special needs education is a research science that addresses the difference resulting from disability and deficit and the differences of gender and culture. These issues are also addressed by the pedagogy of gender and intercultural education. The purpose of special needs education is the integration, to provide appropriate responses to differences through special care that do not occur in separate contexts, but in educational shared places [49.32]. Special needs education is therefore integrated with general pedagogy, which deals with the concept of cognitive discomfort.

To go beyond separation towards integration does not mean neglecting special attention and professional figures able to meet special needs. In this sense, integration and special responses shouldn't be opposed. The specific needs of each individual, seen in her/his singularity, should be respected without providing answers to the needs of individuals grouped into categories. Special problems can be faced without control models that require special places and categories, but rather with open models that fit in with a variety of situations.

For this reason it is important to improve the vocational training in this area [49.33]. In particular, today it is required of teachers to take specific training to recognize, manage, and integrate students with specific learning disabilities in school settings.

A situation of disability is composed of multiple elements that interact with each other. The damage cannot be extrapolated from the historical, cultural and environmental context in which it is located. It is just the interaction with the contextual elements that can improve the quality of life and reduce the handicap relative to disability. Special needs education therefore develops special expertise in relation to the perspective of inclusion [49.34].

## 49.2 Dialecticity

The dialecticity of pedagogy expresses its controversial nature divided between science and philosophy – that is, between explanation and understanding, between theory and practice, between the various disciplines that comprise it, each of which offers its own educational theories.

Pedagogy is a discipline in progress, under construction and re-construction.

A discipline that is constantly renewed – like the phoenix, a fantastic mythological animal that is reborn from its own ashes, while retaining its unchanged nature, identitarian and dynamic at the same time [49.35].

Pedagogy always reconstructs itself in different ways, just because it is in search for more and more new interpretative hypotheses about the education of the human species. Every time, pedagogy is reflected in its pluralistic identity, in its holistic, rather than general, unity [49.36], and in its own disciplinary nature – dysmorphic, complex, and postmodern.

While renewing, pedagogy preserves in time the same architecture, consisting of two main epistemological axes.

### 49.2.1 Science and Philosophy

The relationship with the object of its own research, which makes it a science, a discipline with a defined epistemological status, is at the border between the humanistic and the scientific areas. In fact, in pedagogy, scientific and philosophical aspects that belong to the same entanglement may coexist.

#### Pedagogy as a Science

Along the course of the twentieth century, the reflections about the scientificity of pedagogy shift the epistemological axis from an idea of pedagogy as a *discipline* to an idea of pedagogy as a *science*, and farther, to an idea of pedagogy as a *science of education*, in the sense of a field of knowing with its specific objects and methods. The scientific approach to pedagogy evolves from systematicity to complexity. It develops, namely, in parallel with the construction and the reconstruction of the very idea of science.

The demand of pedagogy to be seen as a science, defining its disciplinary framework, is an idea gained in the course of its history. Its source seems to have been the emergence of the scientific paradigm in the eighteenth century, from Locke's empiricism and its reflections on the importance of experience in the processes of knowing.

In the nineteenth century, the scientific paradigm in pedagogy is enriched by the positivist considerations

on the interplay between theory and observation and by the verifiability of scientific statements in relation to the empirical observation of reality. It can be said that the debate on the scientific foundation of pedagogy started in 1806 with the volume of *Herbart* [49.37] *Allgemeine Pädagogik aus dem Zweck der Erziehung abgeleitet*, which highlights the necessity to consider pedagogy as a science, while allowing the antinomic nature of the discipline – both descriptive, and normative.

The demand for a scientific approach that runs through the course of the nineteenth century and is characterized by the positivist approach – axiomatic, nomothetic, and systematizing, as the scientific method of Comte – turns instead, at the beginning of the twentieth century, towards a more conscious reflection on the peculiarities of pedagogical discourse. In 1911, Durkheim compiles the entry *Pédagogie* for the *Nouveau Dictionnaire de Pédagogie et d'instruction primaire* and explains how pedagogy may be considered a science that reflects on education, a practical theory.

The term *science of education*, used at singular, occurs periodically in nineteenth century pedagogy – by Bain in England and by Siciliani, Ardigò in Italy. It is heard in the pedagogy of the second half of the twentieth century – in Germany, there are Brezinka, Lochner, Rohrs – but also in *Dewey*, in *The sources of a science of education* [49.22]. In Dewey's pragmatism, pedagogy may have a scientific and rigorous method, but it is not necessarily systematic. The pedagogical specificity is expressed in the rediscovery of a sense of intentionality, that can be recognized in all the other disciplines, seen as possible *sources* of a science of education. The *trait d'union* between them is represented by specific hypotheses of project intentionally expressed from time to time in educational direction.

Pedagogy as a science itself poses many questions. First of all, since by its nature it enters into a relationship with a variety of knowledge: what may its specific object of investigation (singular or plural, theoretical or applied) be? Secondly: What type of science should it be? The object of investigation of pedagogy is based around the education of men and women in their historical, cultural, and social contextualization. However, subjects, objects, and methods of pedagogy are always to be considered plural: they are indeed individual and social processes, contextually oriented, of multidisciplinary, plural and variable, nature. It is possible to establish relations of reciprocal interplay between them.

The search for an epistemological status for general pedagogy is unique because each discipline builds its own epistemology, its own way to be a science. During the twentieth century, pedagogy has tried to find an

epistemology that expresses its *proprium* and may justify it *iuxta propria principia*.

In the second half of the last century, the crisis of the concept of science has returned a problematic and complex image of it, focused on uncertainty, instability, disorder, diversity, evolution, and situativity [49.38]. The problematization of science and the criticism of its infallibility proposed by Popper, Kuhn, Lakatos, Feyerabend has overcome the traditional principles of intelligibility of classical science, the idea that science should be general, ahistorical, totalizing, regulated by laws, organizational invariants and constants; causal, linear, deterministic, acontextual, observer-independent, objective, and logical. The principles underlying the paradigm of complexity [49.39] rather represent a rethinking of science and, in particular, they propose the idea that science can be idiosyncratic and unique, as well as general, contingent, historical, irreversible, holistic and subject to feedbacks, to inter-relations and re-orientations.

Pedagogy is therefore a complex science and not a systemic one. Science “as an *attitude* and not like a *system*” as *De Bartolomeis* [49.40] wrote in *The pedagogy as a science* of 1953. Since the 1950s of the twentieth century, pedagogy catches both its fragmentation in multiple sciences and its unitariness between science and philosophy [49.41]. What links pedagogy to other sciences is a functional and pragmatic unity, not a systematic or a methodological one [49.42]. In fact, the scientific data of other disciplines involve pedagogy only if applied to specific working hypotheses, which will serve the educational aims.

Between the 1970s and 1980s, sciences of education are born [49.43]. The *Traité des sciences pédagogiques* [49.43], edited by *Debesse* and *Mialaret*, is now considered the official text of formalization of educational sciences. The *Traité* distinguishes between sciences of education and pedagogical sciences – the former more theoretical, the latter more methodological – and it affirms their mutual interdependence. Pedagogy is stated as a regulatory science across multiple sciences of education. To perform this role pedagogy highlights its philosophical dimension.

### Pedagogy as a Philosophy

The systematization of educational sciences strengthens the philosophical role of pedagogy. Since the 1980s of the twentieth century, the so-called *identity crisis* of pedagogy will bring it to rediscover the sense of its own reflexive intentionality.

Through the philosophical dimension pedagogy expresses its critical role.

The research object of the discipline is not only related to the processes of education, but also to the

metareflexive relations with other disciplines. Pedagogy then is shaped as a network of knowledge related to collaborative relationships, while retaining each of them a separate disciplinary autonomy.

At its core, philosophy of education is now recognized as an integrative aspect of the pedagogical expression in a variety of sciences of education. The philosophical approach, however, is no more analytical and systematic, but hermeneutic.

Philosophy of education plays a mediating and reflexive role both between the different domains and both between the various aspects that compose the complexity of pedagogy like its antinomy, historicity, and the openness to utopian models [49.44].

The metatheoretical inquiry uses second level reflections – the theory of theory – to analyze the dynamic relationship with the practice. Any theory of education in fact can not be an a priori nor dogmatic as it has an indivisible relationship with the practice that contributes to the definition of its historical, political, and social role.

The metatheoretical model explains the links between theory and practice through historically determined categories. In addition, the hermeneutic approach supports the pedagogical purposes expressed through the search for the meaning of its own philosophical dimension.

The philosophical aspects of pedagogy evolve from the macroparadigm of the modern to the postmodern.

## 49.2.2 Theory and Practice

The relationship between theory and practice makes pedagogy a science of education, in particular a theory of educational development processes.

Pedagogy is a theoretical and practical discipline, built on the ideal triangles theory-practice-theory and practice-theory-practice. Pedagogy indeed has a theoretical and projectual dimension to deal with the educational practice.

In practical dimension, theories pass an empirical verification and are confirmed, processed or falsified. This allows to review theories again and consequently to change practice.

Epistemological characteristics concerning the relationship between theory and practice in pedagogy are: dynamism, research, and retroaction.

### Dynamism, Research, and Retroaction

*Dynamism*, as it is an ongoing discipline, it is in continuous progress. The theory of education utilizes the theoretical construct of *antipedagogia* [49.45] as an indicator of the need for a continual renewal. Pedagogy, in this sense, becomes pedagogy of research.

*Research's* epistemological statute is flexible and based on inquiry, as it is a critical discipline. Although the various theories of education are many and different, in educational research there are some recurring concepts:

- The focus on the personal and social development of students and the deepening of their biological and psychological peculiarities.
- The promotion of the interests of students towards knowledge and disciplines of study; towards the environment, to sustain an ecological awareness; towards others to promote respect, communication, collaboration, solidarity, and emancipation.
- The use of the game. In its cognitive dimension, to empower the exploratory, constructive, communicative, and creative function that characterizes it. In its emotional dimension, because of the role of its symbolic and therapeutic function.
- The attention to individual and social differences of physical, mental, cultural, gender, and age nature to respect, promote, and enhance the diversity, as appropriate. It should also be necessary to educate for the diversity for aperture, encounter, and dialog.
- The respect of the individual autonomy as an evolutionary project, as an expression of personal identity, as a fulfillment of freedom as a potential for growth, as an opportunity for critical thinking, for the setup of the capacity for choice and for the development of active learning.
- The study of the individual singularity in its cognitive processing aspects, with particular reference to the plurality of facets of the cognitive prism, in its explicit and implicit levels and in the ways in which they interact: in their primary development, in their structural modularity and in their evolutionary compatibility.
- The development of the sensory and aesthetic dimension, to teach to enjoy the various forms of art but also to produce artistically in a creative way.
- The acknowledgment of the role played by the embodiment in the acquisition of knowledge, in particular the ways in which the field of embodied cognition interacts with the teaching and learning contexts.
- The recognition of the role played by the affective sphere and the strengthening of its balanced development, in relation to cognitive development. This research is intertwined with the psychological approach.

- The encouragement of social and ethical education, which fosters respect and responsibility, the concepts of coexistence, citizenship and solidarity, within the current multiethnic and multicultural societies.

*Retroaction*, as it is a recurrent discipline, which tends to reuse its own paradigms. Although theories of education are continuously overcome by later theories, the previous persist, update, redefine themselves and may even go back in use.

The different interpretative lines that compose the architecture of pedagogy define the relationship between instruction, learning and educability.

### Instruction, Learning, and Educability

The statement of instruction includes knowledge, skills, and methodological tools through which educational goals can be conveyed via explicit curricula such as school and institutional curricula or via implicit curricula such as nonformal and experiential ones.

Learning is an unavoidable process that can be activated independently of explicit purposes. It has a biopsychological and pedagogical value that can be driven by educational values but it does not necessarily depend on them.

Educability is the study of the education of the mind in the epigenesis; the study of the possibilities and constraints of education. According to the principles of educability, each individual activates diverse and dynamic learning that interact modifying themselves along the personal history of learning [49.46].

### Education, Training, and Educational Development

The concept of education is never neutral. Education has indeed an intentional, purposeful, and normative nature, which involves questions of philosophical foundation, as value and meaning.

The category of training and educational development is not only vocational training applied to specific contexts. Educational development underlines the subjectivity and its shaping in individually different ways and in several times and places. It may be considered interdependent in relation to education, and it is more operational.

The intertwining between education, training, and educational development may be outlined by applied models of education.

## 49.3 Applied Models

### 49.3.1 Traditional Models

Models of education influence teaching models [49.47]. Those that are currently used have been developed over the course of the twentieth century. Although some of them are dated, their positive characteristics make them still valid today. However, these models may involve negative aspects. For this reason, their understanding is basilar for teaching. In fact, only the informed use of models of education can effectively support their management.

Traditional models of education reflect different types of teaching. Among these models behaviorist teaching and cognitive teaching are included. Metareflective teaching may be entered in this tradition, because it shares with the previous models some didactical aspects.

Traditional models are linked by the following characteristics:

- The relationship between teacher and students is asymmetric.
- The teacher transmits information and the students receive it.
- The effectiveness of teaching is evaluated by the amount of information transmitted.

Traditional models, however, differ from one another in this way:

- Behaviorist teaching is born at the beginning of the twentieth century through behaviorist studies and it grew until the middle of the last century. Learning consists in associating stimuli and responses. Notions are captured and repeated. The role of the teacher and of the learning environment is emphasized.
- Cognitivist teaching is developed since the Second World War, at the middle of the twentieth century, with the birth of cognitive science. Learning means individual information processing and is intended with the purpose of carrying out cognitive tasks. Information is acquired through specific strategies and is rearranged in a personal way. The role of the student in the learning process is highlighted.
- Metareflective teaching has been identified between the 1970s and 1980s in the twentieth century. Learning is about reflecting on processes of knowledge and on acquisition and management of concepts and strategies. Students learn to recognize strategies they use to learn. Metareflective teaching can be considered a point of junction with current models of education. On the one hand it does not discuss

the asymmetric nature of the teaching and learning relationship, and it considers teaching a process of transmission. On the other hand, it does not evaluate the amount of information transmitted as an expression of the effectiveness of teaching. Metareflective teaching instead evaluates the quality of the organization of learning, cognitivist teaching does this partially, as well. This last aspect, with the accompanying idea that learning is not only individual but also collective, makes metareflective teaching close to postcognitivism and then linkable to present models of education.

#### Behaviorist Model

In the behaviorist model, learning is a response to a stimulus, an association between a stimulus and a conditioned habit. Learning is thus conditioning [49.48, 49].

At the beginning of the twentieth century, studies on behavior (*behaviorism*) challenged all knowledge of mind based on intuition and introspection. The mind was considered an unknowable black box: the cognitive behavior of a student is the product of her/his manifest and directly observable actions.

Teaching stimulates students and determines the learning process. The educational environment defines conditions and opportunities for learning. Any behavior can be acquired from the environment, except for the innate reflexes and the primary emotions. The motivation to study develops itself in relation to the external stimuli. The teacher is responsible for student motivation. The model is in fact teacher centered.

The teacher transmits information in a notionistic way. Students acquire information passively without critical reworking. The teacher controls the class by keywords and conventional gestures, such as clapping hands or opening the register. Students change their cognitive behavior under the teacher's control. The result is the disciplined behavioral response and the modeling behavior, through the recognition of the merits and the punishment of errors. Self-control and auto shaping are taught.

Students imitate the teacher through apprenticeships. There is not a significant difference between *know how* and *know that*. Each student must achieve a standard of mastery performance common to all. In this model individual diversity and personal learning paths are undervalued.

Emotional and cognitive processing is linked in learning and learning, which occurs through the same processes. It is believed that emotional states, although anxious, can play in favor of learning. The emotional

tension can stimulate cognitive enhancement. The manipulation of the emotional anxious states is useful for classroom management.

Learning is a sequential process, in which the start and end points have the same value. The path of instruction is divided into short learning units, as a series of steps. Teaching is programmed through explicit objectives predefined from the outset.

The evaluation of the student occurs through the analysis of what the student shows of knowing. Learning is strengthened through exercise and repetition. To avoid a consolidation of misunderstandings or mistakes, students are often interrogated; for example, after the explanation of a topic. It is thought that it is not necessary to leave to students the individual time that could be needed for a personal revision of contents. The time to learn must be reduced and optimized. The average of the results obtained by the class are considered to calculate the amount of time required to perform a task.

The evaluation of the teacher is heterodirect, that is, it depends on external factors, such as colleagues or superiors, who may suggest criteria for the revision of teaching.

### Cognitive Model

The cognitive model considers learning information processing of the mind [49.50]. The classic behaviorist formula becomes stimulus – organism – response.

The mind is compared to a computer and is simulated through it. The input stimulates the information processing mechanisms of the mind that, at the end of the processing, produces the output, a behavioral response produced by learning. Learning, however, is significant and not transmissive: to store information it is needed to manage units of meaning.

The *information processing* implies that the cognitive system works with different hierarchical functions, respecting specific activation sequences. The subject interprets the information and translates it into its own codes selecting the perceptual stimulation. The information selected, translated and encoded, are shaped as mental representations, i. e., linguistic symbols and numbers.

The processed data are stored in the short and in the long-term memory, for the retrieval of information involved in response planning. The cognitive system manages archives of memories and is intended to be as a computer, a system with a limited capacity that may contain a predefined number of information.

The cognitive system processes information sequentially and hierarchically. Learning is therefore a linear process from percepts to concepts, from simple to complex. The ascending vision of knowledge is

widespread in the West together with the idea of the mind governed by reason. Indeed, in the computational cognitive model the mind is considered as rational, guided by rules, abstract and complex.

Postcognitivist models – for which the mind is also influenced by the emotional, bodily and organismic components – have discussed this model; learning is not subjective but distributed and situated.

In the computational cognitive model, the mind is instead considered as a container, an operating system with limited capacity as the hard disk of a computer. Even if attention is given to cognitive processing, the teacher reflects on how much information can be taught in a lesson and therefore how much knowledge can be learned by each student and how this aim may be achieved to let information be understood and stored at best.

Sequences of instruction are organized from simple to complex. The learning contents are simplified and compacted in significant teaching units to facilitate understanding. Only understanding produces a significant long-term storage in memory. Among the objectives of the teaching is in fact the stabilization during time of the information in the long-term memory. Understanding means to interpret the meanings and to rework them, in an autonomous and individual way, through the use of learning strategies.

Another objective of the teaching is the organization of linguistic concepts. The processes of understanding, processing and storage in memory take place through language. The concept of mental representation implies the management of linguistic symbols and of symbolic translation of languages (*scripts, frames*).

The relationship with students tends to be predefined. For this reason, since the Second World War the educational offer has been often generalized as a standard one. However, around 1980s in the Anglo-Saxon contest, lines of research that encourage individually differentiated ways of teaching have been developed within this model [49.51, 52]. If in the previous model the evaluation of a student was limited by the *quantity* of her/his intelligence, for example through the ranges of IQ, now this model rather considers *how* an individual can be smart. The cognitive system is seen as differentiated inside and distinct in a plurality of aspects both at the perceptual, and the processing, level.

The individualization of teaching involves the rethinking of teaching strategies. Teaching should be individualized and personalized, that is, renewed according to a vision of the student as a unique, singular and variable, person in a dynamic relationship with the teacher and other students.

### Metareflective Model

This model is at the border between the traditional models – with which it can share the idea that teaching is asymmetric and transmissive – and the postcognitive models that have developed over the last twenty years of the past century, with which it can share a re-thinking of the mind as distributed and collaborative. The metareflective model represents a possible link between the traditional and current educational models.

If in traditional models the *amount* of the learned contents and the *quality* of the strategies used to process them are evaluated, this model is based instead on the evaluation of the awareness of each student about *how to manage* cognitive and emotional levels of learning through individual strategies [49.53–55]. Each student should in fact structure individual modalities of monitoring and control in order to regulate the flow of learning within its own cognitive system. Reflection on learning influences its self-regulating function of autonomous and self-conscious management. This reflection can be shared with other students in a collaborative way because each student may learn to assess her/his own quality of learning.

The teacher is an example of managerial organization of concepts because she/he achieves to regulate her/his own learning. The purpose of education is to stimulate students' self-reflective function. This function – that is, the awareness about the ways in which each cognitive system uses its resources – involves a second level work about learning that is defined precisely as metareflective, metacognitive, and meta-emotional as appropriate.

The role of the metareflective teacher is to make students aware of the ways in which they learn. The student that reflects on her/his own learning strategies can manage them and assess autonomously her/his preparation.

Among the processes for monitoring and control of cognitive functions are:

- The E.O.L. (ease of learning), that is, an estimation of the ease of learning of the information presented by the educational offer.
- The J.O.L. (judgment of learning), that is, the personal judgment of a student about the probability to recall a learned information.
- The F.O.K. (feeling of knowing), that is, the feeling of knowing a learned information that, however, the student can not remember at all (the phenomenon *on the tip of the tongue*).
- The P.T.R. (prediction of total recall), that is, the prediction of the final total amount of memory recall after a single learning session.

In this model the ways through which tacit knowledge – ideas, theories, and concepts of implicit nature that each individual develops dealing with reality – become aware, referable, and verbalized are also studied. Emerging theories about the surrounding world and the minds of other individuals are validated, reviewed or processed.

Each student has an explicit and implicit learning potential. Implicit learning, mainly present in the structuring of thought in the early stages of cognitive development, can absorb from the environment potentially disordered and disorganized information. One of the objectives of the metareflective model is therefore to make explicit the implicit acquisitions that may represent a potential source of discomfort for the functioning of the cognitive system. In this model implicit learning has a predominantly negative value and it is assessed its educational risk.

Learning opportunities also concern explicit and implicit contexts, defined as formal, nonformal, and informal educational contexts. It is namely possible to learn in a nonvoluntary indirect way by environmental experiences.

Even the teacher uses the metareflective method to review and possibly change her/his teaching strategies.

### 49.3.2 Actual Models

Current models of education have developed in the last decades of the twentieth century without a chronological linearity defining their onset. The recognizable types of teaching within these models have indeed emerged in a parallel and mutually interacting way; for this reason they are hardly distinguishable one from each other. However, it is possible to identify the teaching models of contextualism, culturalism, and constructivism, in the line of development of the patterns of thought of Vigotsky, Bruner, and Piaget.

These three models belong to the broader framework of interpretation called postcognitivism.

Compared to the traditional models of education, in the postcognitivist models:

- Cognitive interaction is considered not only in its abstract processing but also in its emotional, bodily, and organismic dimension.
- The study of the mind, therefore, is no longer in vitro but becomes in vivo.
- Cognitive and emotional aspects of the learning are linked.
- The mind is not separated from the body: knowledge thus includes perceptual and behavioral areas that are often expressed through implicit levels.

- The contexts of learning are intended as specific situations located in space and time.
- Learning is not a *task oriented* process but an embodied, situated, and distributed one.
- The educational relationship becomes almost symmetric, because knowledge is no longer transmitted but shared and co-constructed, negotiated within the learning community.
- The teacher acts as a facilitator of learning and a mediator in the processes of knowledge.

Knowledge is distributed, situated, and embodied [49.56–58]. Distributed knowledge has motivated contextualist models of teaching; situated knowledge has justified culturalist patterns, and both these interpretations have concurred to define constructivist and socio-constructivist didactics.

Didactics is no longer individually oriented (toward the teacher or the student) but is based on the concept of the *relationship* between the teacher and the student and among students.

#### Contextualist Model

In the contextualist model, knowledge is a distributed process, and it is no more the subjective patrimony of each individual. The mind is seen beyond the boundaries of individuality and is distributed among several individuals who perform a work of sharing, socializing, and negotiating learning [49.59, 60].

There is a shift from learning processes to knowledge structures. The student leaves the world of singular learning and enters that of shared learning, in which the knowledge itself is a collective heritage to which everyone can contribute. The educational relationship becomes almost symmetrical, with no-guided processes that are activated in different contexts – formal but also semiformal, nonformal and informal.

The class group becomes a learning community [49.61, 62]. The learning community is a real and/or virtual meeting place in which the educational exchange takes place. Knowledge becomes increasingly linked to the different contexts of belonging and is rooted in them.

Learning is a collective process. Knowledge is shared in the intersubjectivity: every task, that is the knowledge object to study, becomes a problem to be dealt with in the learning community, through dynamic exchange and discussion. Among the aims of the teaching there is not only the education of the individual student but also the education of the learning community. Therefore, the ability of students to learn to express their own views and to know how to relate to each other in a mediating way, continuously revising their own opinions, is enhanced.

Particular emphasis is given to the language as a tool of social communication that allows the educational exchange of ideas, theories and concepts in the processes of social co-construction and in the communitarian negotiation of meanings.

#### Culturalist Model

Even in the culturalist model, knowledge is distributed among multiple parties. The distribution of knowledge, however, includes the cognitive artifacts, i. e., the possible expressions that make up the various cultural frameworks. Attention is given to the different cultures – and to their evolution – in which the mind develops [49.63–65]. Cultures constitute learning environments where knowledge is acquired and is replayed also through the peripheral devices, i. e., the technology tools. The cognitive artifacts and the peripheral devices are contingent and relative to situations defined in space and time; therefore they are located in specific cognitive contexts, belonging to each culture. The mind is situated and embedded in the environment.

Situated knowledge loses its character of generality that it had previously: it is no longer a static and predefined patrimony of information, the *general culture*. The various areas of knowledge become dynamic, constantly changing, different in form and in content, and representative of specific ways to make culture. They are considered *domains* of knowledge, contextualized and interactive areas in which cognitive relations are activated. Knowledge is therefore related to times, places, cultures, relationships, and cognitive domains.

The teacher is proposed as a model of identification for her/his students. Learning means to enter into a cognitive relationship with others, within a learning community in which knowledge is shared, mediated and negotiated. The learning community represent the belonging group, with its own way of making culture, of producing knowledge and of teaching it.

Learning in a learning community means taking the cultural responsibility of personal cognitive heritage. At the same time, it can mean the possibility to detach from it in an autonomous way, even if always in relation to educational figures which have been configured as a model and as an example. What occurs is a process of identification that may lead to the recognition of the individual *cognitive identity*, i. e., the knowledge of the personal cognitive heritage. The acquisition of cognitive identity may involve both integrative and detachment functions.

The student can consider the experience and the personal knowledge as a resource of which she/he becomes responsible: she/he thus acquires the cognitive responsibility for her/his own choices and for the meaning



that these choices may acquire within the learning community. Through processes of initiation, the role of the student becomes that of a member of the community; an apprentice who learns to internalize the educational acts of the belonging adult system through processes of scaffolding. The apprenticeship is done under the guidance of an experienced adult who is also a mediator and a facilitator in the learning community. The student is educated in relation to the experiential knowledge of this educational figure that nevertheless plays a role of reference: the student may indeed choose to follow the example of her/his own adult model. Otherwise, the student may choose to detach herself/himself from the given model and to propose alternative models, after having however had a relationship of comparison with the adult model.

Knowledge thus is developed according to the community sharing of the belief systems, the discursive practices, namely the use of languages for the negotiation of meanings, then through the linguistic exchange but also through the sharing of cognitive artifacts, and of peripheral devices, so generating the productions of different cultures.

#### Constructivist Model

The constructivist model emphasizes the role of the subject in the process of construction of knowledge, even when students learn together. This model promotes the critical and interpretive development of autonomous learning. The experiential relationship with the environment is a process of guided discovery. The teacher collaborates in the development of individual knowledge structures.

In the course of the ontogenetic development, each individual elaborates implicit theories – defined as naive and of common sense – relative to the surrounding environment or to the minds of others [49.66, 67]. The teacher can make these implicit theories explicit through education. The intervention of teaching occurs on demand, when the teacher considers it necessary to clarify a concept or when specifically requested by a student. The teacher keeps a more asymmetric position toward the student than in other current models of education; in fact, the teacher's role is that of a guide. The educational guide takes place in semiformal contexts.

Learning is an adaptive process in which knowledge is constructed to be acquired and produced. In learning environments, educational interactions occur for adaptation and participation, as in sociocultural constructivism.

The process of change that accompanies the construction of knowledge should be autonomous and,

at the same time, guided by the teacher only when necessary. In this way their *visit* in the world of knowledge may enable students to interpret reality in a personal way. According to neo-Piagetian constructivism [49.68], the subjective theories about reality have to be shared within the community of learning to be validated, transformed, or abandoned.

The potential of the individual to know in a learning environment is then developed according to the individual knowledge structures and to their sharing in situations of collective co-construction.

#### 49.3.3 Experimental Models

Experimental models analyze the influence of neuroscience and biological sciences in education. The fields of study addressing the relationship between mind and brain, particularly bioeducational sciences, express the research perspectives of experimental models [49.69].

Experimental models share these views:

- Embodied cognition
- Holistic understanding of the human system as a complex phenomenon
- Relationship between the mental, organismic, and environmental dimension
- Study of the evolutionary processes in ontogenesis and in phylogenesis
- Interpretation of learning as an adaptive process
- Attention to the development of individual knowledge structures
- Designing implicit learning environments for educability
- Nonreductionist, interactionist, and integrative research approaches.

#### Enriched Model

The enriched model is a classical model; perhaps it has always been used in the teaching and learning relationship. The basic idea is that for an environment to be effective, it has to stimulate students as much as possible: therefore it has to be designed in a rich and stimulating way; it should use a multimodal methodology [49.70].

Learning is a process that is based on brain plasticity and on the neural network's modifiability. The teacher works to get from students openness to change, and to educational stimuli. The teacher tries to fully express the potential of each student; a poor and deprived learning environment may not activate students' readiness to learn. This model is widely used in the early school levels.

The enriched model sometimes uncritically reflects the ascensional and linear vision of knowledge, according to which concepts are developed by percepts, and abstract ideas are developed from sensations.

Furthermore, this model can be influenced by a consideration of the development as a continuous consequent progress. This idea involves an interpretation of the development itself as a process that may only occur according to certain conditions that may be present in a given learning environment.

Development in general should not be considered as a process that takes place starting on the basis of its potential genetic and epigenetic bases; it is instead a dynamic and interactive process, discontinuous and variable, which conditions cannot be predefined.

### Organismic Model

In the organismic model, mind is embodied [49.71], and not separated from body and brain as in a part of the traditional Western thought. The interaction between the affective/emotional and bodily/organismic level is studied in the processes of knowledge management. In this model, therefore, bodily and emotional functions are linked to cognitive functions in influencing the explicit and implicit processing of knowledge.

The organismic model studies implicit learning on the perceptual, motor, but also processing side [49.72]. Implicit learning can be automated, automatized, and contribute to the structuring of learning as an implicit form of knowledge. In some of them, it is expected that the implicit may interact with the explicit through specific dynamics of relationship.

Implicit learning is studied in its evolutionary and adaptive significance, as a possible form of knowledge that is based on the conservation and on the selective repetition of the strategies that prove to be the most effective in the ontogenetic and phylogenetic development. This model also explores the role played by implicit learning in the prototypical knowledge processing and as a default support for explicit learning, e.g., in the Elementary Logic Theory [49.73].

The organismic model attributes importance to the proxemics of communication in educational practices, to the influence of physiological states on learning, and to the diversity of learning times that cannot be predefined in relation to cognitive tasks.

### Adaptive Model

The adaptive model highlights the individual peculiarity of the learner and the unpredictability of situations in which learning occurs. In a general sense, the cognitive diversity of individuals interacts with the domain specificity of the various knowledge fields and with the relativity of each context. Consequently, knowledge appears as constantly modifiable and continuously changing [49.74].

The design of the educational offer becomes dynamic, ever predefined. Programming a learning environment foresees both its positive and negative development: evolutionary forces and resistances to change, according to the perception of the efficacy of learning by students.

Educational planning becomes adaptive. To be educational, interactions that eventually occur within it have to be reciprocal: if a student accepts the changes required by the environment as a condition of adaptation, at the same time the environment has to embrace the heterodirect change and to accept to modify itself consequently.

In the adaptive model, the teacher evaluates the possibilities and the constraints of the educational action, that is, he/she evaluates the educability of learners. Each student has a personal cognitive potential that has developed over time through continuous adaptation processes.

The explicit or implicit personal history of learners may be both open and close to change. The levels of modifiability of a cognitive system depend on the modularity of the learning within the system itself, in different ways for each individual.

Cognitive modifiability is related to the explicit and implicit choices that each student takes on in its path of experience. The individual strategic procedures of processing are intertwined with the external influences. The result may be the adaptive efficacy, but also the cognitive discomfort.

The levels of cognitive modifiability therefore also depend on the compatibility and on the adaptive interaction between the ways of learning of a student and the ways of teaching of a teacher.

Educability then concerns how aspects of the cognitive experience can be integrated with each other in the complex mechanism of any adaptive system, along the personal history of learning.

## 49.4 Conclusions

The *mechanism* of the pedagogical discourse is expressed in the complexity of the relationship between ideology, science, and utopia [49.1]. This last aspect, in particular, may represent the *trait d'union* between the scientific and the philosophical component. The utopian nature of pedagogy tends toward the model of an ideal educating society [49.75]. Utopia underlies the design of the pedagogical domain.

The vector of utopia reflects the deep complexity of the issue and its rejection of universality, as-

sertiveness, and uniqueness. This does not preclude, however, the research for models and methods as part of a scientific status of interpretive nature, sensitive to cultures, in dialogue with societies, open to the ideas of project and possibility; therefore, to utopia.

Pedagogy, in conclusion, is a very dynamic and open field of research whose developmental directions are horizontal, multidisciplinary pervasive and continuously in progress.

## References

- 49.1 F. Cambi: *Il Congegno del Discorso Pedagogico* (Clueb, Bologna 1986), in Italian
- 49.2 D.H. Jonassen, B.L. Grabowski: *Handbook of Individual Differences Learning and Instruction* (Routledge, New York, London 2011)
- 49.3 H.L. Swanson, K.R. Harris, S. Graham (Eds.): *Handbook of Learning Disabilities* (Guilford, New York 2013)
- 49.4 B.Y.L. Wong, D.L. Butler (Eds.): *Learning About Learning Disabilities* (Elsevier, San Diego 2012)
- 49.5 P. Tinkler, C. Jackson: The past in the present: Historicising contemporary debates about gender and education, *Gender Educ.* **26**(1), 70–86 (2014)
- 49.6 P. Stephens: *Social Pedagogy: Heart and Head* (Europäischer Hochschulverlag, Bremen 2013)
- 49.7 G.S. Levine, A. Phipps, C. Blyth (Eds.): *Critical and Intercultural Theory and Language Pedagogy* (Cengage Learning, Florence 2010)
- 49.8 J.A. Banks: *Cultural Diversity and Education: Foundations, Curriculum, and Teaching* (Pearson Allyn Bacon, Boston 2001)
- 49.9 J. Clarke, A. Hanson, R. Harrison, F. Reeve (Eds.): *Supporting Lifelong Learning: Perspectives on Learning* (RoutledgeFalmer, London 2002)
- 49.10 P. Jarvis: *Twentieth Century Thinkers in Adult and Continuing Education* (Taylor Francis, Oxford 2012)
- 49.11 M.S. Knowles, E.F.I.I.I. Holton, R.A. Swanson: *The Adult Learner* (Elsevier, London 2011)
- 49.12 Z. Bekerman, N.C. Burbules, D. Silberman-Keller (Eds.): *Learning in Places: The Informal Education Reader* (Peter Lang, New York 2006)
- 49.13 F. Santoianni: *Modelli di Studio. Apprendere con la Teoria Delle Logiche Elementari* (Erickson, Trento 2014), in Italian
- 49.14 A. Rogers: *Non-Formal Education: Flexible Schooling or Participatory Education?* (Kluwer, New York 2005)
- 49.15 V.C.X. Wang, L. Farmer, J. Parker, P.M. Golubski: *Pedagogical and Andragogical Teaching and Learning with Information Communication Technologies* (IGI Global, Hershey 2012)
- 49.16 P. Petrie: *Communication Skills for Working with Children and Young People* (Jessica Kingsley, London 2011)
- 49.17 A.A. Ciccone, R.A.R. Gurung, N.L. Chick, A. Haynie (Eds.): *Exploring Signature Pedagogies: Approaches to Teaching Disciplinary Habits of Mind* (Stylus, Sterling 2008)
- 49.18 S.N. Hesse-Biber: *Mixed Methods Research: Merging Theory with Practice* (Guilford, New York 2010)
- 49.19 M. Gennari, A. Kaiser: *Prolegomeni alla Pedagogia Generale* (Bompiani, Milano 2000), in Italian
- 49.20 G.L. De Landsheere: History of educational research. In: *International Encyclopedia of Education*, Vol. 3, ed. by T. Husen, T.N. Postlethwaite (Pergamon, Oxford 1985)
- 49.21 G. Mialaret: *Les Sciences de L'éducation* (Presses Universitaires de France, Paris 1976), in French
- 49.22 J. Dewey: *The Sources of a Science of Education* (Horace Liveright, New York 1929)
- 49.23 H. Daniels: *Vygotsky and Pedagogy* (RoutledgeFalmer, London 2001)
- 49.24 L.R. Meeth: Interdisciplinary studies: Integration of knowledge and experience, *Change* **10**, 6–9 (1978)
- 49.25 R. Nola, G. Irzik: *Philosophy, Science, Education and Culture* (Springer, Dordrecht 2005)
- 49.26 J. Storø: *Practical Social Pedagogy* (The Policy Press, Bristol 2013)
- 49.27 P. Orefice: Participatory research methods in the education of adult: Theoretical and methodological aspects. In: *Towards the End of Teaching? Innovation in European Adult Learning*, ed. by M. Dale (NIACE, Leicester 2000)
- 49.28 F. Cambi: *Storia Della Pedagogia* (Laterza, Roma-Bari 1995), in Italian
- 49.29 H. Leser: *Das pädagogische Problem in der Geistesgeschichte der Neuzeit* (Oldenburg, München 1925), in German
- 49.30 J. Dewey: *Democracy and Education: An Introduction to the Philosophy of Education* (Macmillan, New York 1916)
- 49.31 F. Cambi: *Abitare il Disincanto. Una Pedagogia per il Postmoderno* (UTET Università, Turin 2006), in Italian
- 49.32 E.C. Winter: Preparing new teachers for inclusive schools and classrooms, *Support Learn.* **21**(2), 85–

- 91 (2006)
- 49.33 G. Golder, B. Norwich, P. Bayliss: Preparing teachers to teach pupils with special educational needs in more inclusive schools: Evaluating a PGCE development, *Br. J. Special Educ.* **32**(2), 92–99 (2005)
- 49.34 M. Nind, J. Rix, K. Sheehy, K. Simmons (Eds.): *Curriculum and Pedagogy in Inclusive Education. Values into Practice* (Routledge, New York 2013)
- 49.35 F. Santoianni: *La Fenice Pedagogica. Linee di Ricerca Epistemologica* (Liguori, Napoli 2007), in Italian
- 49.36 R.L. Zigler: The holistic paradigm in educational theory, *Educ. Theory* **28**(4), 318–326 (1978)
- 49.37 J.F. Herbart: Allgemeine Pädagogik, aus dem Zweck der Erziehung abgeleitet (1806). In: *Pädagogische Grundschriften, Pädagogische Schriften*, Vol. 2, ed. by W. Asmus (Klett, Stuttgart 1982)
- 49.38 E. Morin: *Introduction à la Pensée Complexe* (ESF, Thiron 1990), in French
- 49.39 E. Morin: *Science Avec Conscience* (Fayard, Paris 1982), in French
- 49.40 F. De Bartolomeis: *La pedagogia come scienza* (La Nuova Italia, Firenze 1953)
- 49.41 W. Brezinka: Empirical sciences and other educational theories: Differences and possibilities for agreement. In: *Critical Rationalism and Educational Discourse*, ed. by G. Zecha (Rodopi, Amsterdam 1999) pp. 153–169
- 49.42 A. Visalberghi: *Problemi Della Ricerca Pedagogica* (La Nuova Italia, Firenze 1965), in Italian
- 49.43 M. Debesse, G. Mialaret: *Traité des Sciences Pédagogiques* (Presses Universitaires de France, Paris 1969), in French
- 49.44 F. Heyting, D. Lenzen, J. White (Eds.): *Methods in Philosophy of Education* (Routledge, London 2001)
- 49.45 F. De Bartolomeis: *La Ricerca Come Antipedagogia* (Feltrinelli, Milano 1969), in Italian
- 49.46 F. Santoianni: *Educabilità Cognitiva* (Carocci, Roma 2006), in Italian
- 49.47 F. Santoianni: *Modelli e Strumenti di Insegnamento* (Carocci, Roma 2010), in Italian
- 49.48 J.H. Block, L.W. Anderson: Mastery learning. In: *Handbook on Teaching Educational Psychology*, ed. by D. Treffinger, J. Davis, R. Ripple (Academic, New York 1977)
- 49.49 D.A. Lieberman: *Learning: Behavior and Cognition* (Wadsworth, Belmont 1990)
- 49.50 H. Gardner: *The Mind's New Science: A History of the Cognitive Revolution* (Basic, New York 1985)
- 49.51 H. Gardner: *Formae Mentis. Frames of Mind: The Theory of Multiple Intelligences* (Basic, New York 1983)
- 49.52 R.J. Sternberg, L.F. Zhang: *Perspectives on Thinking, Learning, and Cognitive Styles* (LEA, Mahwah 2001)
- 49.53 J. Metcalfe, A. Shimamura (Eds.): *Metacognition: Knowing About Knowing* (MIT Press, Cambridge 1994)
- 49.54 L.M. Reder (Ed.): *Implicit Learning and Metacognition* (LEA., Mahwah 1996)
- 49.55 D.J. Hacker, J. Dunlosky, A. Graesser (Eds.): *Metacognition in Educational Theory and Practice* (LEA, Mahwah 1998)
- 49.56 D. Kirshner, J.A. Whitson (Eds.): *Situated Cognition: Social, Semiotic, and Psychological Perspectives* (Erlbaum, Hillsdale 1997)
- 49.57 B.C. Smith: Situatedness/Embeddedness. In: *The MIT Encyclopedia of the Cognitive Sciences*, ed. by R.A. Wilson, F. Keil (MIT Press, Cambridge 1999)
- 49.58 C. Bereiter: *Education and Mind in the Knowledge Age* (LEA, Mahwah 2002)
- 49.59 D. Magnusson (Ed.): *The Lifespan Development of Individuals. Behavioral, Neurobiological, and Psychosocial Perspectives* (Cambridge Univ. Press, Cambridge 1996)
- 49.60 P.B. Baltes, U.M. Staudinger (Eds.): *Interactive Minds. Life-Span Perspectives on the Social Foundation of Cognition* (Cambridge Univ Press, Cambridge 1996)
- 49.61 J. Retallick, B. Cocklin, K. Coombe: *Learning Communities in Education: Issues, Strategies and Contexts* (Routledge, Londra 1999)
- 49.62 B. Rogoff, C.G. Turkanis, L. Bartlett: *Learning Together: Children and Adults in a School Community* (Oxford Univ. Press, Oxford 2003)
- 49.63 D.R. Olson, N. Torrance (Eds.): *The Handbook of Education and Human Development* (Blackwell, Oxford 1996)
- 49.64 M. Cole: *Cultural Psychology: A Once and Future Discipline* (Cambridge Univ. Press, Cambridge 1996)
- 49.65 K. Egan: *The Educated Mind. How Cognitive Tools Shape Our Understanding* (Univ. Chicago Press, Chicago 1997)
- 49.66 H.M. Wellman: *The Child's Theory of Mind* (MIT Press, Cambridge 1990)
- 49.67 P. Carruthers, P. Smith (Eds.): *Theories of Theory of Mind* (Cambridge Univ. Press, Cambridge 1995)
- 49.68 S. Carey, R. Gelman: *The Epigenesis of Mind* (Erlbaum, Hillsdale 1991)
- 49.69 E. Frauenfelder, F. Santoianni (Eds.): *Mind, Learning and Knowledge in Educational Contexts* (Cambridge Scholars, Cambridge 2003)
- 49.70 D.H. Jonassen, S.M. Land: *Theoretical Foundations of Learning Environments* (LEA, Mahwah 2000)
- 49.71 M.L. Anderson: Embodied cognition: A field guide, *Artif. Intell.* **149**, 91–130 (2003)
- 49.72 M.A. Stadler, P.A. Frensch (Eds.): *Handbook of Implicit Learning* (Sage, Londra 1998)
- 49.73 F. Santoianni: Educational models of knowledge prototypes development, *Mind Soc.* **10**, 103–129 (2011)
- 49.74 F. Santoianni, C. Sabatano (Eds.): *Brain Development in Learning Environments. Embodied and Perceptual Advancements* (Cambridge Scholars, Cambridge 2007)
- 49.75 C. Metelli Di Lallo: *Analisi del Discorso Pedagogico* (Marsilio, Padova 1967), in Italian

## 50. Model-Based Reasoning in Crime Prevention

Charlotte Gerritsen, Tibor Bosse

Model-based reasoning approaches can be used to formalize and analyze (informal) theories from the field of criminology, to help gain more insight in criminological phenomena that were not clear based on just the informal theory. The analysis of the *displacement of crime* is an important research interest in criminological research. In this chapter, an agent-based simulation model of crime displacement is presented, which can be used not only to *simulate* the spatiotemporal dynamics of crime, but also to *analyze* and *control* those dynamics. Methods are used that are aimed at developing intelligent systems that monitor human-related processes and provide appropriate support. More specifically, an explicit domain model of crime displacement has been developed, and model-based reasoning techniques are applied to the domain model, in order to analyze which environmental circumstances result in which crime rates, and to determine which support

50.1	<b>Ambient Intelligence</b> .....	1053
50.2	<b>Methodology</b> .....	1054
50.3	<b>Domain Model</b> .....	1055
50.3.1	Crime Displacement.....	1055
50.3.2	Formalization.....	1056
50.4	<b>Analysis Model</b> .....	1058
50.5	<b>Support Model</b> .....	1060
50.6	<b>Results</b> .....	1060
50.7	<b>Discussion</b> .....	1062
	<b>References</b> .....	1062

measures are most appropriate. The model can be used as an analytical tool for researchers and policy makers to perform thought experiments, that is, to shed more light on the process under investigation, and possibly improve existing policies (e.g., for surveillance).

Criminology is the study of crime, criminals, and the punishment of criminals [50.1]. To study criminal behavior a number of standard research methods exist, for example, victim surveys, offender surveys, social experiments, and the analysis of police data. Based on these methods, multiple theories have been developed that provide insight into delinquent behavior. However, these theories are usually informal, meaning that they are written in natural language or described graphically and, thus, in principle ambiguous. Approaches based on computational modeling can prove to be very useful to address this void. Criminological theories can be translated into a formal, unambiguous, machine-readable notation so that they can be used for model-based reasoning and simulation, for example, [50.2–4] and help gain more insights in criminological phenomena that were not clear based on just the informal theory.

Within the field of criminology, one of the main research interests is the analysis of the *displacement of crime* [50.5–7]. Typically, certain locations in a city seem to attract many criminal activities, but only for

a short period. These locations where many crimes occur are called *hot spots* [50.7]. Questions that are important in understanding the displacement of crime are:

- When do hot spots of high crime rates emerge?
- Where do they emerge?
- And, perhaps most importantly, how can they be prevented.

In recent years, computational modeling and simulation have proved to be a useful instrument to answer such questions. Most of these approaches take the routine activity theory as a point of departure, that is, the assumption that crime is likely to occur when a motivated offender finds a suitable target, while capable guardians are lacking [50.5]. By creating an *artificial world* in a computational environment, populating it with computational entities representing the *agents* addressed by the theory (i.e., offenders, targets, and guardians), and establishing mathematical rules that govern the agents' behavior, the spatiotemporal implications of

the theory can be studied with the help of the computer.

Nevertheless, when investigating the literature on computational modeling of displacement of crime, a wide variety of different computational modeling approaches can be distinguished. Among the approaches that are applied, one can find agent-based modeling [50.8–11], population-based modeling [50.9], cellular automata [50.12, 13], different spatial analysis techniques [50.14], and evolutionary computing techniques [50.11]. The underlying principle behind *agent-based modeling* approaches is the *agent* metaphor, that is, the idea to compose a model of autonomous pieces of software that make their own decisions, based on information they observe in their direct environment. For instance, the agent-based approach presented in [50.8] simulates the spatiotemporal dynamics of crime as a result of individual decisions of offenders, targets, and guardians (e.g., to move around, or to perform assaults or arrests). In contrast, *population-based modeling* approaches, for example, [50.9] do not distinguish individual agents, but instead describe the dynamics of crime in terms of mathematical formulae (mostly differential equations) over variables that represent the densities of certain subgroups in the population. Furthermore, *cellular automata* (CA) are discrete models that consist of grids of cells that are in particular states (e.g., on or off), of which the dynamics are determined by rules that take the states of adjacent cells into account. Liu et al. [50.13] have used CAs to simulate individual crime events, in order to generate plausible crime patterns. In their approach, the main elements are offenders, targets, and crime places, of which different attributes can be manipulated, such as motivation of offenders, capability of guardians, and accessibility of places. Additionally, a number of *spatial analysis techniques* are used in [50.2]; these techniques include, among others, geographical information systems (GIS) and analytical methods. Finally, *evolutionary computing* is a subarea of artificial intelligence (AI) that attempts to find optimal solutions to mathematical problems by exploiting a computational variant of biological evolution. More specifically, by representing candidate solutions to an optimization problem in terms of individuals in a population, and having the population evolve using operations such as recombination and mutation (where the better performing individuals have a higher probability to reproduce), good solutions can be found to problems in a variety of domains, including criminology. For example, in [50.11] some results are presented that were achieved with GAPatrol, an evolutionary multi agent-based simulation tool devised to assist police

managers in the design of effective police patrol route strategies.

Although these approaches all share the aim of investigating crime displacement, the perspectives taken differ. For example, some authors try to develop simulation models of crime displacement in existing cities, which can be directly related to real-world data, for example, [50.13], whereas others deliberately abstract from empirical information, for example, [50.9]. The idea behind the latter perspective is that the simulation environment is used as an analytical tool, mainly used by researchers and policy makers, for thought experiments, to shed more light on the process under investigation, and perhaps improve existing policies (e.g., for surveillance) [50.15]. Also, some authors take an intermediate point of view, for example, [50.4, 8]. They initially build their simulation model to study the phenomenon per se, but define its basic concepts in such a way that it can be directly connected to empirical information, if this becomes available.

This intermediate perspective is also taken in the current chapter. Its main goal is to develop an agent-based simulation model of crime displacement, which can be used not only to *simulate* the spatiotemporal dynamics of crime, but also to *analyze* and *control* those dynamics. This second aim distinguishes it from most existing approaches, which are mainly descriptive (instead of prescriptive).

To achieve this goal, we make use of techniques from AI, and in particular from ambient intelligence (AmI). Ambient intelligence [50.16–18] represents a vision of the future where humans will be surrounded by pervasive and unobtrusive electronic environments, which are sensitive and responsive to their needs. In order to develop such intelligent environments, Bosse et al. [50.19] introduced a methodology to endow intelligent systems with the possibility to reason explicitly about the mental and physical states of humans. In this chapter, this methodology is reused in order to develop an intelligent system that reasons about crime displacement.

More specifically, this chapter will first describe the development of an explicit *domain model* of crime displacement (which describes displacement in terms of states of the world over time, and transitions between these states). On top of that, model-based reasoning techniques [50.20] will be applied to the domain model, in order to analyze which environmental circumstances result in which crime rates, and to determine which support measures are most appropriate [50.21]. Hence, both an *analysis model* and a *support model* will be developed.

In the remainder of this chapter, first some background information about the area of AmI will be provided, followed by an introduction to the basic methodology for the development of intelligent human-aware model-based systems that will be used. Based

on this methodology, the next sections will introduce, respectively, the domain model, analysis model, and support model for crime displacement. After that some preliminary simulation results will be presented, followed by a discussion.

## 50.1 Ambient Intelligence

AmI [50.16–18] represents a vision of the future where human beings will be surrounded by pervasive and unobtrusive electronic environments, which are sensitive and responsive to their needs. Such an environment has a certain degree of awareness of the presence and states of living creatures in it, and supports their activities. It analyzes their behavior, and may anticipate on it. AmI integrates concepts from ubiquitous computing and AI with the vision that technology will become invisible, embedded in our natural surroundings, present whenever we need it, attuned to the humans' senses, and adaptive to them. In an AmI environment, people are surrounded by networks of embedded intelligent devices that can sense their state, anticipate, and when relevant adapt to their needs. Therefore, the environment should be able to determine which actions have to be undertaken in order to keep this state optimal.

For this purpose, acquisition of sensor information about humans and their functioning is an important factor. However, without adequate additional *knowledge* for analysis of this information, the scope of such applications is limited. As argued by *Bosse et al.* [50.19], AmI applications can show a more human-like understanding and base personal care on this understanding when they are equipped with knowledge about the relevant physiological, psychological, and/or social aspects of human functioning. For example, this may concern elderly people, patients depending on regular medicine usage, surveillance, penitentiary care, psychotherapeutical/self-help communities, but also, for example, humans in highly demanding tasks such as warfare officers, air traffic controllers, crisis and disaster managers, and humans in space missions; for example, [50.22].

Within human-directed scientific areas, such as cognitive science, psychology, neuroscience, and biomedical sciences, models have been and are being developed for a variety of aspects of human functioning. If such models of human processes are represented in a formal and computational format, and incorporated in the human environment in devices that monitor the physical and mental state of the human, then such devices are able to perform a more in-depth analysis of the human's functioning. This can result in an environment

that may more effectively affect the state of humans by undertaking actions in a knowledgeable manner that improve their wellbeing and performance. For example, the workspaces of naval officers may include systems that, among others, track their eye movements and characteristics of incoming stimuli (e.g., airplanes on a radar screen), and use this information in a computational model that is able to estimate where their attention is focused at. When it turns out that an officer neglects parts of a radar screen, such a system can either indicate this to the person, or arrange on the background that another person or computer system takes care of this neglected part. Note that for a radar screen it would also be possible to make static design changes, for example, those that improve situation awareness, for example, picture of the environment [50.23]. However, as different circumstances might need a different design, the advantage of a dynamic system is that the environment can be adapted taking both the circumstances and the real-time behavior of the human into account.

In applications of this type, an ambience is created that has a better understanding of humans, based on computationally formalized knowledge from the human-directed disciplines. The use of knowledge from these disciplines in AmI applications is beneficial, because it allows taking care in a more sophisticated manner of humans in their daily living in medical, psychological, and social respects. In more detail, content from the domain of human-directed sciences, among others, can be taken from areas such as medical physiology, health sciences, neuroscience, cognitive psychology, clinical psychology, psychopathology, sociology, criminology, and exercise and sport sciences.

Although it does not directly fit in the description of AmI, the system envisioned by the current paper has a number of similarities with the types of systems sketched above. That is, it will also take information about humans and their dynamics as input (namely the spatial distribution of individuals over the city, and information about crime rates), it will also be equipped with (formalized) knowledge from human-directed disciplines (in this case criminological knowledge about crime displacement), and it will also generate sup-

port measures as output (i. e., advice to reduce crime). Thus, in order to develop the intelligent system for reasoning about crime displacement, it makes sense to

reuse approaches from the AmI area. In particular, the methodology from [50.19] is used, which is introduced below.

## 50.2 Methodology

In this section, the adopted approach to develop intelligent human-aware systems is presented in detail [50.19]. Here, human-aware is defined as being able to analyze and estimate what is going on in the human's mind (a form of mindreading) and in his or her body (a form of bodyreading). Input for these processes are observed information about the human's state over time, and dynamic models for the human's physical and mental processes. For the mental side, such a dynamic model is sometimes called a theory of mind [50.24] and may cover, for example, emotion, attention, intention, and belief. Similarly for the human's physical processes, such a model relates, for example, to skin conditions, heart rates, and levels of blood sugar, insulin, adrenalin, testosterone, serotonin, and specific medicines taken. Note that different types of models are needed: physiological, neurological, cognitive, emotional, social, as well as models of the physical and artificial environment (In this chapter, the main focus is on social/environmental states and models, that is, locations of persons, and decisions to move to other location. Nevertheless, the model is sufficiently generic to be extended with the other types of states as well).

A framework can be used as a template for the specific class of AmI applications as described. The structure of such an ambient software and hardware design can be described in an agent-based manner at a conceptual design level and can be given generic facilities built in to represent knowledge, models, and analysis methods about humans, for example (Fig. 50.1):

- Human state and history models
- Environment state and history models

- Profiles and characteristics models of humans
- Ontologies and knowledge from biomedical, neurological, psychological, and/or social disciplines
- Dynamic process models about human functioning
- Dynamic environment process models
- Methods for analysis on the basis of such models.

Examples of useful analysis methods are voice and skin analysis with respect to emotional states, gesture analysis, and heart rate analysis. The template can include slots where the application-specific content can be filled to get an executable design for a working system. The analysis method used in this chapter mainly addresses displacement of crime, that is, it calculates how a certain distribution of persons over space would lead to movement of criminal activities.

A general approach for embedding knowledge about the interaction between the environment and the human(s) in AmI applications is to integrate dynamic models of this interaction (i. e., a model of the *domain*) into the application. This integration takes place by embedding domain models in certain ways within agent models of the intelligent application. By incorporating domain models within an agent model, the intelligent agent gets an understanding of the processes of its surrounding environment, which is a solid basis for knowledgeable intelligent behavior. Three different ways to integrate domain models within agent models can be distinguished. A most simple way is to use a domain model that specifically models human behavior in the following manner:

*Domain model directly used as agent model.* In this case a domain model that describes human processes and behavior is used directly as an agent model, in order to simulate human behavior. Note that here the domain model and agent model refer to the same agent.

Such an agent model can be used in interaction with other agent models, in particular with *ambient agent models* to obtain a test environment for simulations. For this last type of (artificial) agents, domain models can be integrated within their agent models in two different ways, in order to obtain one or more (sub)models; see Fig. 50.2. Here the solid arrows indicate information exchange between processes (data flow) and the dot-

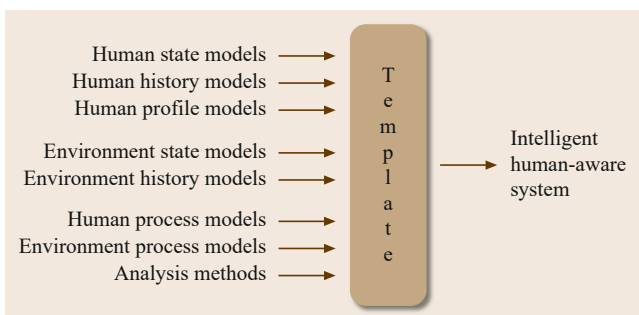
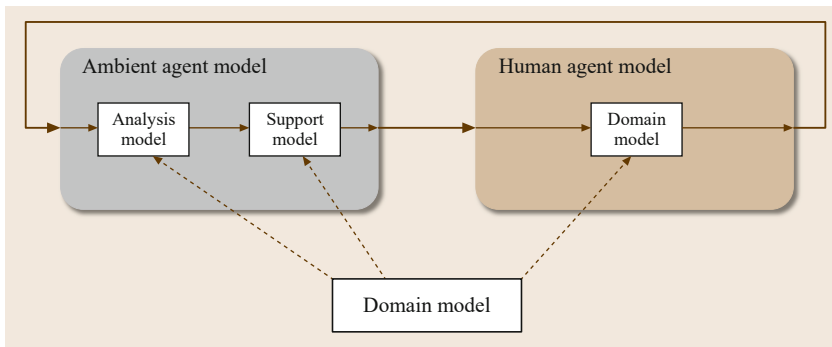


Fig. 50.1 Framework to develop intelligent human-aware systems (after [50.19])





**Fig. 50.2** Overview of the multiagent system architecture (after [50.19])

ted arrows the integration process of the domain models within the agent models.

As shown in Fig. 50.2, the following submodels can be obtained based on a domain model:

- **Analysis model:** To perform analysis of the human's states and processes by reasoning based on observations (possibly using specific sensors) and the domain model.
- **Support model:** To generate support for the human by reasoning based on the domain model.

Note that here the domain model that is integrated refers to one or more human agents, whereas the agent model in which it is integrated refers to an artificial

agent (the intelligent system). In the following sections, this methodology will be applied to the domain of crime displacement. First a domain model is presented which represents the spatiotemporal dynamics of crime. Next, an analysis model is presented, which is able to reason about the domain model in order to predict crime rates for particular situations. And finally, a support model is presented, which is able to suggest to the user the most appropriate measures to reduce crime rates. For example, in case the analysis model predicts that the crime rates at the railway station will increase with 20% in the next year, and that these rates can be kept stable by increasing the amount of police by 5%, then it may propose to invest in 5% more police forces.

## 50.3 Domain Model

This section presents the domain model for crime displacement. The important concepts used are introduced in Sect. 50.3.1, and their formalization is described in Sect. 50.3.2.

### 50.3.1 Crime Displacement

As explained in the introduction, most large cities in the world contain a number of *hot spots*, that is, locations where the majority of the crimes occur [50.7, 25]. Such locations may vary from railway stations to shopping malls. These hot spots usually have several things in common, among which the presence of many passers-by (which makes the location attractive for criminals) and the lack of adequate surveillance. However, after a while the situation often changes: the criminal activities shift to another location. This may be caused by improved surveillance systems (such as cameras) at that location, by an increased number of police officers, or because the police changed their policy.

Another important factor in explaining crime displacement is the *reputation* of specific locations in

a city [50.6]. This reputation may be a cause of crime displacement, as well as an effect. For example, a location that is known for its high crime rates usually attracts police officers [50.25], whereas most citizens will be more likely to avoid it [50.26]. As a result, the amount of criminal activity at such a location will decrease, which affects its reputation again.

To summarize, in order to model the process of crime displacement, several aspects are important. First, one should have information about the *total number* of agents in the different groups involved, that is, the number of *criminals*, number of *guardians*, and number of *passers-by*. Next, it is assumed that the world (or city) that is addressed can be represented in terms of a number of different *locations*. It is important to know how many agents of each type are present at each location: the *density* of criminals, guardians, and passers-by. Furthermore, to describe the movement of the different agents from one location to another, information about the *reputation* (or *attractiveness*) of the locations is needed. This attractiveness is different for each type of agent. For example, passers-by like lo-

cations where it is safe, for example, locations where some guardians are present and no criminals. On the other hand, guardians are attracted by places where a lot of criminals are present, and criminals like locations where there are many passers-by and no guardians. Finally, to be able to represent the idea of hot spots, the *number of assaults* per location is modeled. The idea is that more assaults take place at locations where there are many criminals and passers-by, and few guardians, cf. the routine activity theory by [50.5].

The interaction between the concepts introduced above is visualized in Fig. 50.3. This figure depicts the influences between the different groups at one location. Here, the circles denote the concepts that were mentioned above in italics, and the arrows indicate influences between concepts (influences on attractiveness have been drawn using dotted arrows to enhance readability). (Note that Fig. 50.3 does not depict the influence of some *basic attractiveness* of a location for certain groups (i.e., an attractiveness that is independent of the distribution of agents at the location). For the sake of readability, this notion has been left out of the picture, but it often plays a role in reality. For instance, locations like a railway station will be visited more often by passers-by than other locations, simply because people need to go there to reach their desired destination. Therefore, the notion of basic attractiveness will also be considered in this chapter).

### 50.3.2 Formalization

In order to build the domain model for crime displacement, the concepts that were introduced above (in italics) are formalized in terms of mathematical vari-

ables. The variable names that are used are summarised in Table 50.1.

Next, a number of mathematical equations are introduced to represent the causal relations between these variables. Most of these ideas are taken over from [50.9, 27, 28]. First, the calculation of the number of agents at a location is done by determining the movement of agents that takes place based on the attractiveness of the location. For example, for criminals, the following formula is used

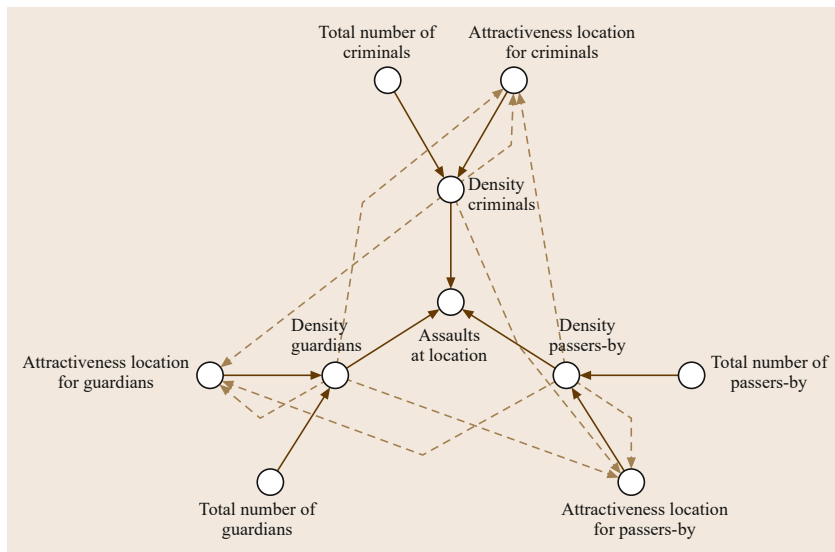
$$c(L, t + \Delta t) = c(L, t) + \eta(\beta(L, c, t)c - c(L, t))\Delta t.$$

This expresses that the density  $c(L, t + \Delta t)$  of criminals at location  $L$  on time  $t + \Delta t$  is equal to the density of criminals at the location at time  $t$  plus a constant  $\eta$  (expressing the rate at which criminals move per time unit) times the movement of criminals from  $t$  to  $t + \Delta t$  from and to location  $L$ , multiplied by  $\Delta t$ . Here, the movement of criminals is calculated by multiplying the relative attractiveness  $\beta(L, c, t)$  of the location (compared to the other locations) for criminals with the total number  $c$  of criminals (which is constant). From this, the density of criminals at the location at  $t$  is subtracted, resulting in the change of the number of criminals for this location. For passers-by, a similar formula is used

$$p(L, t + \Delta t) = p(L, t) + \eta(\beta(L, p, t)p - p(L, t))\Delta t.$$

However, as opposed to [50.9], the movement of the guardians is not (necessarily) modelled using this formula. Instead, to represent guardian movement, different strategies can be filled in.

Next, the attractiveness of a location can be expressed based on some form of reputation of the loca-



**Fig. 50.3** Interaction between criminals, guardians, and passers-by

**Table 50.1** Variables used in the domain model

Name	Explanation
$c$	Total number of criminals
$g$	Total number of guardians
$p$	Total number of passers-by
$c(L, t)$	Density of criminals at location $L$ at time $t$ .
$g(L, t)$	Density of guardians at location $L$ at time $t$ .
$p(L, t)$	Density of passers-by at location $L$ at time $t$ .
$\beta(L, a, t)$	Attractiveness of location $L$ at time $t$ for type $a$ agents: $c$ (criminals), $p$ (passers-by), or $g$ (guardians)
$ba(L, a, t)$	Basic attractiveness of location $L$ at time $t$ for type $a$ agents: $c$ (criminals), $p$ (passers-by), or $g$ (guardians)
Assault_rate( $L, t$ )	Number of assaults taking place at location $L$ per time unit.

tion for the respective type of agents. Several variants of a reputation concept can be used. The only constraint is that it is assumed to be normalized such that the total over the locations equals 1. An example of a simple reputation concept is based on the densities of agents, as expressed below.

$$\beta(L, c, t) = \frac{p(L, t)}{p} \text{ for criminals ,}$$

$$\beta(L, p, t) = \frac{g(L, t)}{g} \text{ for passers-by .}$$

This expresses that criminals are more attracted to locations with higher densities of passers-by, whereas passers-by are attracted more to locations with higher densities of guardians. This definition of reputation is used in [50.9]. Although this definition is simple, which makes the model well suited for mathematical analysis, it is not very realistic. To solve this problem, in this chapter, the following linear combinations of densities are used.

$$\beta(L, c, t) = \beta_{c1} \left( 1 - \frac{g(L, t)}{g} \right) + \beta_{c2} \frac{p(L, t)}{p} + \beta_{c3} ba(L, c, t) ,$$

$$\beta(L, p, t) = \beta_{p1} \left( 1 - \frac{c(L, t)}{c} \right) + \beta_{p2} \frac{g(L, t)}{g} + \beta_{p3} ba(L, p, t) .$$

(Note that these attractiveness formulae are not normalized yet. To ensure that the values stay between 0

and 1, each attractiveness value is divided by the sum of the values over all locations. Moreover, the influence by agents from the same group is not considered.)

This expresses that criminals are repelled by guardians, but attracted by passers-by. Similarly, passers-by are repelled by criminals, but may be attracted by guardians. In addition, for each type of agent some basic attractiveness can be defined. The weight factors ( $\beta_{xy}$ , which may also be 0) indicate the relative importance of each aspect. Again, for the guardians no formula is specified, since this depends on the guardian movement strategy that is selected.

Finally, to measure the assaults that take place per time unit, also different variants of formulae can be used [50.9]. In this chapter, the following is used

$$\text{assault\_rate}(L, t) = \max(c(L, t)p(L, t) - \gamma g(L, t), 0) .$$

Here, the assault rate at a location at time  $t$  is calculated as the product of the densities of criminals and passers-by, minus the product of the guardian density and a constant  $\gamma$ , which represents the capacity of guardians to avoid an assault. The motivation behind this is that the maximum amount of assaults that can take place at a location is  $c(L, t)p(L, t)$ , but that this number can be reduced by the effectiveness of the guardians (which corresponds exactly to the routine activity theory). In principle, this assault rate can become less than 0 (the guardians can have a higher capacity to stop assaults than the criminals have to commit them); therefore the maximum can be taken of 0 and the outcome described above. Based on this assault rate, the total (cumulative) amount of assaults that take place at a location is calculated as

$$\text{total\_assaults}(L, t + \Delta t) = \text{total\_assaults}(L, t) + \text{assault\_rate}(L, t)\Delta t .$$

Although the domain model is presented here in a purely mathematical notation, its actual implementation has been done in the agent-based modeling environment LEADSTO [50.29]. This environment is well suited for the current purposes, since it integrates both qualitative, logical aspects and quantitative, numerical aspects. The basic building blocks of LEADSTO are executable rules of the format  $\alpha \rightarrow \beta$ , which indicates that state property  $\alpha$  leads to state property  $\beta$ . Here,  $\alpha$  and  $\beta$  can be (conjunctions of) logical and numerical predicates.

## 50.4 Analysis Model

This section extends the domain model introduced in the previous section to an analysis model. The analysis model (and the support model, see next section) is created by taking the domain model as a basis, and applying model-based reasoning to it. In particular, two types of reasoning are applied taken from [50.20]: forward and backward reasoning. In short, these types of reasoning make use of the following kinds of (simplified) rules (where  $X$  and  $Y$  are variables in a model, for example, as in Fig. 50.3):

- If you believe  $X$  and believe that  $Y$  depends on  $X$ , then you also believe  $Y$ .

$$\text{belief}(X) \wedge \text{belief}(\text{depends\_on}(Y, X)) \rightarrow \text{belief}(Y) .$$

- If you desire  $Y$  and believe that  $Y$  depends on  $X$ , then you also desire  $X$ .

$$\text{desire}(Y) \wedge \text{belief}(\text{depends\_on}(Y, X)) \rightarrow \text{desire}(X) .$$

To illustrate the idea, assume that we focus on an existing city, of which the average number of criminals, guardians, and passers-by at the different locations is known (to a certain extent). Thus, specific numbers can be assigned to the variables  $\text{density\_criminals}$ ,  $\text{density\_guardians}$ , and  $\text{density\_passers\_by}$  in Fig. 50.3 (which correspond to  $c(L, t)$ ,  $g(L, t)$ , and  $p(L, t)$  in Table 50.1). Then, via forward reasoning (the first rule shown above), the model can predict how the number of assaults will change over time.

One step further, instead of taking the actual densities of guardians at the different locations, the analysis model can also be used to investigate how the crime rates would change in case the densities of guardians were different. To this end, the analysis model is extended with the possibility to specify particular crime

prevention *strategies*. The idea is that, in addition to the rules that govern the behavior of criminals and passers-by, the behavior of the guardians can be specified by selecting one out of multiple strategies.

In current practice, the crime prevention policies that are applied by law enforcement agencies are – mostly – reactive [50.25, 30]. That is, these agencies often only increase the level of guardianship at locations where crimes have been committed in the past. As a consequence, this often means that such a decision is made too late, because the damage has already been done. Instead, we hypothesize that a more anticipatory strategy (e.g., a strategy to invest in more guardians at locations where one predicts that a hot spot *will emerge*) may be more efficient.

To be able to investigate this, the analysis model is equipped with multiple strategies for movement of guardians (varying from reactive to anticipatory, and combinations of the two). The selected strategies are based on [50.27, 28], in which they were already tested against some initial scenarios. In total, the analysis model contains 10 different strategies (see also Table 50.2):

- The first strategy is a *baseline* strategy. In this case guardians do not move at all. Their density at the different locations remains stable over time.
- The second strategy (called *reactive 1*) states that the amount of guardians that move to a new location is proportional to the density of criminals at that location.
- The third strategy (*reactive 2*) states that the amount of guardians that move to a new location is proportional to the percentage of the assaults that have recently taken place at that location.
- The fourth strategy (*reactive 3*) states that the amount of guardians that move to a new location is proportional to the percentage of all assaults that have taken place so far at that location.

**Table 50.2** Guardian movement strategies considered by the analysis model

Strategy	Formalization of $\sigma(L, t)$
Baseline	0
Reactive 1	$(c(L, t)/c)g - g(L, t)$
Reactive 2	$\text{aar}(L, t)g - g(L, t)$
Reactive 3	$\text{taar}(L, t)g - g(L, t)$
Reactive 4	$(p(L, t)/p)g - g(L, t)$
Anticipate 1	$(c(L, t) + \eta_2(\beta(L, c, t)c - c(L, t))\Delta t)/cg - g(L, t)$
Anticipate 2	$p(L, t) + \eta_2(\beta(L, p, t)p - p(L, t))\Delta t/pg - g(L, t)$
Anticipate 3	$((c(L, t) + \eta_2(\beta(L, c, t)c - c(L, t))\Delta t)/c + (p(L, t) + \eta_2(\beta(L, p, t)p - p(L, t))\Delta t)/p)/2g - g(L, t)$
Hybrid 1	$((\text{aar}(L, t)g - g(L, t)) + (p(L, t) + \eta_2(\beta(L, p, t)p - p(L, t))\Delta t)/pg - g(L, t))/2$
Hybrid 2	$((\text{taar}(L, t)g - g(L, t)) + (p(L, t) + \eta_2(\beta(L, p, t)p - p(L, t))\Delta t)/pg - g(L, t))/2$

- The fifth strategy (*reactive 4*) states that the amount of guardians that move to a new location is proportional to the density of passers-by at that location.
- In the sixth strategy (*anticipate 1*), the amount of guardians that move to a new location is proportional to the density of criminals they expect that location to have in the future.
- In the seventh strategy (*anticipate 2*), the amount of guardians that move to a new location is proportional to the density of passers-by they expect that location to have in the future.
- In the eighth strategy (*anticipate 3*), the amount of guardians that move to a new location is proportional to the amount of assaults they expect that will take place at that location in the future. This predicted amount of assaults is approximated by taking the average of the expected densities of criminals and passers-by.
- The ninth strategy (*hybrid 1*) is a combination of *reactive 2* and *anticipate 2*. Here, the amount of guardians that move to a new location is the average of the amounts of guardians determined by those two strategies.
- The tenth strategy (*hybrid 2*) is a combination of *reactive 3* and *anticipate 2*. Here, the amount of guardians that move to a new location is the average of the amounts of guardians determined by those two strategies.

To formalize these strategies, the following formula is used

$$g(L, t + \Delta t) = g(L, t) + \eta\sigma(L, t)\Delta t.$$

This formula is similar to the formulae used for criminals and passers-by, but the amount of guardians that move per time unit is indicated by the factor  $\sigma(L, t)$ , which depends on the chosen strategy. The different definitions of  $\sigma$  are shown in Table 50.2. For example, for the baseline strategy,  $\sigma(L, t) = 0$ , which means that the amount of guardians at time point  $t + \Delta t$  is equal to the amount at  $t$ .

In the strategies *reactive 2* and *3*, the average assault rate  $aar(L, t)$  and the total average assault rate  $taar(L, t)$  are calculated by

$$aar(L, t) = \frac{\text{assault\_rate}(L, t)}{\sum_{X:\text{loc}} \text{assault\_rate}(X, t)},$$

$$taar(L, t) = \frac{\text{total\_assaults}(L, t)}{\sum_{X:\text{loc}} \text{total\_assaults}(X, t)}.$$

As can be seen from Table 50.2, the idea of the anticipation strategies is that the guardians use formulae that are similar to the formulae for movement of criminals and passers-by to predict how they will move in the near future. Obviously, these predictions will not be 100% correct, since they do not consider interaction between the different types of agents, but our assumption is that they may be useful means to develop an efficient strategy.

Furthermore, different values can be taken for the parameter  $\eta_2$  in the anticipation strategies. This parameter represents the speed by which the criminals and/or passers-by move in the predicted scenario (or, in other words, the distance in the future for which the prediction is made). For example, by taking a very high value for  $\eta_2$  in the *anticipate 1* strategy, guardians get the tendency to move to locations that are predicted to have a high density of criminals in the very far future.

As mentioned earlier, the idea of having different strategies is that the analysis model can test which one performs best. A question is however how to define the notion of a *good* strategy. One possibility is to look at effectiveness, for example, by considering the strategy that results in the lowest crime rates (total\_assaults) as the best. However, in reality also the *costs* of crime prevention play an important role. Various mechanisms to improve guardianship exist (e.g., adding and moving security guards, burglar alarms, fencing, lighting), but they all involve costs [50.30]. Thus, instead of only measuring the amount of assaults that result from each strategy, in the calculation of the *best* strategy one should compensate for the costs involved. For this reason, the following formula (which was not included in [50.20]) has been added

$$\begin{aligned} \text{total\_costs}(t + \Delta t) \\ = \text{total\_costs}(t) + \sum_{X:\text{loc}} \sigma(X, t)\varepsilon\Delta t. \end{aligned}$$

This formula counts the total costs that are spent on crime prevention (for all locations involved) during the simulation. Parameter  $\varepsilon$  represents the guardian movement costs per time step.

## 50.5 Support Model

On top of the analysis model presented above, also a support model for crime prevention has been developed. This model takes as input certain information about the future scenario for which the user desires support. Based on this information, it generates advices about which strategies are recommended to prevent crime in this scenario.

More specifically, the model first needs to have some information about the state of the world. In particular, the user needs to specify the geography of the city (i. e., which locations are relevant?), and the initial densities of the different types of agents for each location. In addition to this, the user needs to define a scenario, that is, (s)he needs to indicate the total time span for which the system is to provide support, and to specify for each location how its basic attractiveness will change during this time span. For instance, in case a circus will temporarily come to town, the basic attractiveness of the location of the circus is likely to increase. Finally, the user has to specify the maximum amount of money (s)he desires to spend.

To summarize, the support model takes the following information as input (which needs to be entered by the user of the system):

- Geography of the city (i. e., which locations are relevant?)
- Initial densities of the different types of agents ( $c(L, t)$ ,  $g(L, t)$ ,  $p(L, t)$ ) for each location
- Total time span of the scenario
- Basic attractiveness for the different types of agents ( $ba(L, c, t)$ ,  $ba(L, g, t)$ ,  $ba(L, p, t)$ ) for each location over time
- Maximum budget.

On the basis of these settings, the support model requests the analysis model to perform simulations for all possible strategies, to determine for each of these strategies to which crime rates it would lead, and what its costs would be. After that, the support model selects the *best* strategies, and presents information about those strategies to the user. The strategies that are assessed as best are those strategies of which the costs are lower than the maximum budget. Moreover, concerning the remaining strategies, in case some strategy  $s_1$  turns out both more expensive and less effective than some strategy  $s_2$ , then this strategy  $s_1$  is removed from the selection. Upon request, the model can also provide the user more detailed information about the dynamics of the effect of a particular strategy in the scenario.

## 50.6 Results

A prototype implementation of the model has been developed. To illustrate the behavior of the prototype, below (part of) the dynamics of an example execution are shown in detail.

This example addresses a scenario where there are three locations, and 3900 agents. The population considered consists of 600 (potential) criminals, 300 guardians, and 3000 passers-by. Initially, these agents are distributed equally over the three location (i. e., at each location, there are 200 criminals, 100 guardians and 1000 passers/by). Moreover, all locations start with the same basic attractiveness (= 0.33 on a [0, 1] scale). After 50 time steps the attractiveness of the locations changes: location 1 becomes very attractive (= 0.6), location 2 becomes slightly less attractive (= 0.3), and location 3 becomes much less attractive (= 0.1). The scenario lasts 100 time steps and the maximum budget the user can spend is 100.

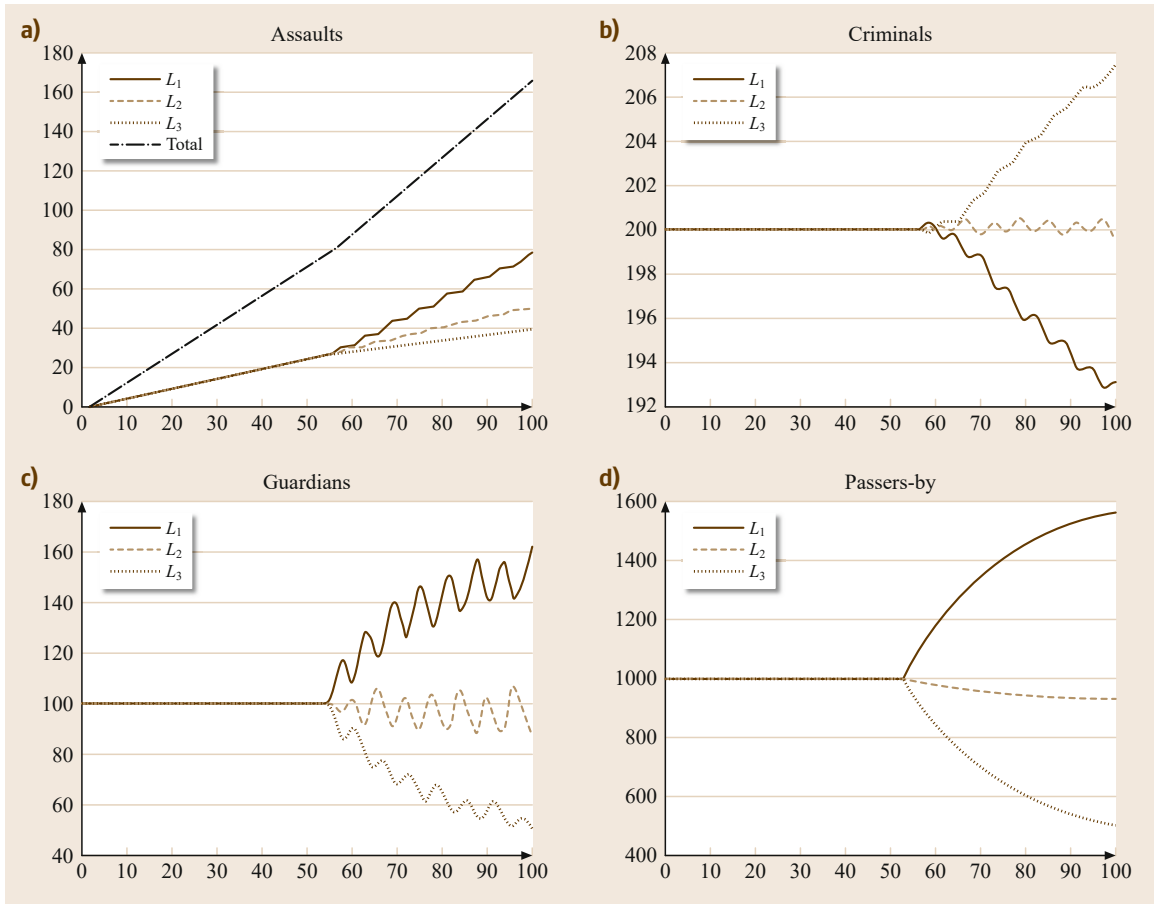
When executing the system based on these settings, for the analysis model would predict the dynamics of the scenario for each of the different strategies, as mentioned above. As an illustration, such a prediction

is visualized for one particular strategy (in this case, the *reactive 2* strategy, Table 50.2) in Fig. 50.4. Figure 50.4a–d shows the assault rate, and the amount of criminals, guardians, and passers-by at the different locations. In all graphs, the red line indicates location  $L_1$ , the green line indicates location  $L_2$ , and the blue line indicates location  $L_3$ . The black line in Fig. 50.4a shows the total amount of assaults, that is, the sum of the assaults at the three locations.

As can be seen in Fig. 50.4, over the first 50 time points, the number of the different types of agents at the locations stays equal. After time point 50, the amounts change. The guardians move away from location 3 to location 1, which is the most attractive location. The criminals move away from location 1 because they want

**Table 50.3** Recommendation by the support model

Recommended strategies	Predicted costs	Predicted % assaults prevented
Anticipate 1	30.38	7.7
Anticipate 2	93.16	70.9
Anticipate 3	45.10	40.6



**Fig. 50.4a-d** Results of an example simulation run by the analysis model

to move away from the guardians. The passers-by move toward location 1 since they want to be at the safest location (i. e., the location with the highest amount of guardians and the lowest amount of criminals). In this case, the strategy used by the guardians seems to work well, because the total number of assaults (i. e., the black line in Fig. 50.4a) grows not much faster than it did during the first 50 time points. When comparing this, for instance, with a baseline strategy in which the guardians are static (which is also tested by the analysis model but not shown in Fig. 50.4), this turns out to be a significant improvement.

All in all, the analysis model tries out all possible strategies and provides the results to the support model. Based on this, the support model selects the most promising strategies (in the context of the user's

preferences), and presents them as a recommendation to the user. Table 50.3 shows what this recommendation looks like for the current scenario.

As can be seen from Table 50.3, the system predicts that the three *anticipate* strategies are *best*, that is, they have costs that are below the budget of the user and are nevertheless effective. Moreover, the system predicts that strategy *anticipate 1* will be cheapest, but that strategy *anticipate 2* will be most effective.

Although this is only a single example scenario, it clearly illustrates that the model is able to generate an appropriate advice on police investment, which may actually be used by policy makers in order to reduce crime rates. For a more detailed comparison between the different strategies in various scenarios, see [50.27, 28].

## 50.7 Discussion

In this chapter, model-based reasoning techniques were applied to the domain of criminology. We presented an approach to analyze crime displacement. The approach was inspired by an existing methodology from AmI [50.16–18], which proposes that intelligent human-aware systems are composed of three separate components, namely a *domain model*, an *analysis model*, and a *support model* [50.19]. In the context of crime displacement, the role of the domain model was to simulate the dynamics of crime displacement, but on top of that, the analysis model proved useful to reason about such simulations for different settings, and the support model was able to generate advice on the basis of the results of this reasoning. The advice consists of a selection of guardian movement strategies that are recommended for a particular scenario, augmented with additional information about the costs and effectiveness

of these strategies. A prototype version of the model has been implemented, and some initial tests have pointed out that the model provides realistic advices.

Despite these encouraging results, one should be careful not to overgeneralize them. Currently, they were achieved in simulations that used several specific parameters and simplifying assumptions. Nevertheless, after further testing, the model may provide useful input for policy makers, in order to elaborate their thoughts about efficient strategies, and possibly improve existing surveillance policies.

**Acknowledgments.** This work previously appeared as: T. Bosse, C. Gerritsen: A model-based reasoning approach to prevent crime, *Stud. Comput. Intell.* **314** 159–177 (2010), *Proc. Int. Conf. Model-Based Reason. Sci. Technol.*, MBR'09, ed. by L. Magnani, W. Carnielli

## References

- 50.1 Merriam-Webster: <http://www.merriam-webster.com/dictionary/criminology>
- 50.2 C. Gerritsen: Caught in the Act: Investigating Crime by Agent-Based Simulation, Ph.D. Thesis. (VU Univ., Amsterdam 2010)
- 50.3 L. Liu, J. Eck (Eds.): *Artificial Crime Analysis Systems: Using Computer Simulations and Geographic Information Systems* (Information Science Reference, Hershey 2008)
- 50.4 N. Malleon, P. Brantingham: Prototype burglary simulations for crime reduction and forecasting, *Crime Patterns Anal.* **2**(1), 47–66 (2008)
- 50.5 L.E. Cohen, M. Felson: Social change and crime rate trends: A routine activity approach, *Am. Sociol. Rev.* **44**, 588–608 (1979)
- 50.6 D.T. Herbert: *The Geography of Urban Crime* (Longman, Harlow 1982)
- 50.7 L.W. Sherman, P.R. Gartin, M.E. Buerger: Hot spots of predatory crime: Routine activities and the criminology of place, *Criminology* **27**, 27–55 (1989)
- 50.8 T. Bosse, C. Gerritsen: Social simulation and analysis of the dynamics of criminal hot spots, *J. Artif. Soc. Soc. Simul.* (2010), doi:10.18564/jasss.1498
- 50.9 T. Bosse, C. Gerritsen, M. Hoogendoorn, S.W. Jafry, J. Treur: Agent-based versus population-based simulation of displacement of crime: A comparative study, *Web Intell. Agent-Syst. Int. J.* **9**, 147–160 (2011)
- 50.10 P.L. Brantingham, U. Glässer, K. Singh, M. Vajihollahi: *Mastermind: Modeling and Simulation of Criminal Activity in Urban Environments, Technical Report SFU-CMPTTR-2005-01* (Simon Fraser Univ., Burnaby 2005)
- 50.11 D. Reis, A. Melo, A.L.V. Coelho, V. Furtado: Towards optimal police patrol routes with genetic algorithms. In: *Intelligence and Security Informatics*, ed. by S. Mehrotra, D.D. Zeng, H. Chen, B. Thuraisingham, F. Wang (Springer, Berlin, Heidelberg 2006) pp. 485–491, LNCS 3975
- 50.12 K. Hayslett-McCall, F. Qui, K.M. Curtin, B. Chastain, J. Schubert, V. Carver: The simulation of the journey to residential burglary. In: *Artificial Crime Analysis Systems: Using Computer Simulations and Geographic Information Systems*, ed. by L. Liu, J. Eck (Information Science Reference, Hershey 2008) pp. 281–300
- 50.13 L. Liu, X. Wang, J. Eck, J. Liang: Simulating crime events and crime patterns in RA/CA model. In: *Geographic Information Systems and Crime Analysis*, ed. by F. Wang (Idea Group, Singapore 2005) pp. 197–213, in RA
- 50.14 E.R. Groff: *The Geography of Juvenile Crime Place Trajectories*, Ph.D. Thesis (University of Maryland, College Park 2005)
- 50.15 H. Elffers, P. van Baal: Spatial Backcloth is not that important in simulation research: An illustration from simulating perceptual deterrence. In: *Artificial Crime Analysis Systems*, ed. by L. Liu, J.E. Eck (IGI Global, Hershey 2008) pp. 19–34
- 50.16 E. Aarts, R. Collier, E. van Loenen, B. de Ruyter (Eds.): *Ambient intelligence: Proc. First Eur. Symp. EUSAI 2003*, Lect. Notes Comput. Sci., Vol. 2875 (Springer, Berlin, Heidelberg 2003)
- 50.17 E. Aarts, R. Harwig, M. Schuurmans: Ambient intelligence. In: *The Invisible Future*, ed. by P. Denning (McGraw Hill, New York 2001) pp. 235–250
- 50.18 G. Riva, F. Vatalaro, F. Davide, M. Alcañiz (Eds.): *Ambient Intelligence* (IOS, Amsterdam 2005)
- 50.19 T. Bosse, M. Hoogendoorn, M.C.A. Klein, R. van Lambalgen, P.P. van Maanen, J. Treur: Incorporat-



- ing human aspects in ambient intelligence and smart environments. In: *Handbook of Research on Ambient Intelligence and Smart Environments: Trends and Perspectives*, ed. by F. Mastrogiovanni, N.Y. Chong (IGI Global, Hershey 2011) pp. 128–164
- 50.20 T. Bosse, F. Both, C. Gerritsen, M. Hoogendoorn, J. Treur: Methods for model-based reasoning within agent-based ambient intelligence applications, *Know. Based Syst. J.* **27**, 190–210 (2012)
- 50.21 T. Bosse, R. Duell, M. Hoogendoorn, M.C.A. Klein, R. van Lambalgen, A. van der Mee, R. Oorburg, A. Sharpanskykh, J. Treur, M. de Vos: An adaptive personal assistant for support in demanding tasks, *Lect. Notes Comput. Sci.* **5638**, 3–12 (2009), *Proc. 4th Int. Conf. Augmented Cognition and 13th Int. Conf. Human-Computer Interaction, HCI'09* (Springer, Berlin, Heidelberg)
- 50.22 D.J. Green: Realtime compliance management using a wireless realtime pillbottle – A report on the pilot study of SIMPILL, *Proc. Int. Conf. eHealth, Telemedicine and Health, Med-e-Tel'05, Luxembourg* (2005)
- 50.23 C.D. Wickens: Situation awareness and workload in aviation, *Curr. Directions Psychol. Sci.* **11**(4), 128–133 (2002)
- 50.24 S. Baron-Cohen: *Mindblindness* (MIT Press, Cambridge 1995)
- 50.25 J.E. Eck, S. Chainey, J.G. Cameron, M. Leitner, R.E. Wilson: *Mapping Crime: Understanding Hot Spots* (U.S. Department of Justice, Washington 2005), <http://www.ojp.usdoj.gov/nij/pubs-sum/209393.htm>
- 50.26 W. Skogan: Fear of crime and neighborhood change. In: *Communities and Crime*, Crime and Justice, Vol. 8, ed. by A.J. Reiss Jr., M. Tonry (Univ. Chicago Press, Chicago 1986) pp. 203–229
- 50.27 T. Bosse, C. Gerritsen: An agent-based framework to support crime prevention, *Proc. 9th Int. Conf. Auton. Agents Multi-Agent Syst.*, ed. by W. van der Hoek, G.A. Kaminka, Y. Lespérance, M. Luck, S. Sen (ACM, Toronto 2010) pp. 525–532
- 50.28 T. Bosse, C. Gerritsen: Comparing crime prevention strategies by agent-based simulation, *Proc. 9th IEEE/WIC/ACM Int. Conf. Intell. Agent Technol., IAT'09 (IEEE CS, 2009)* pp. 491–496
- 50.29 T. Bosse, C.M. Jonker, L.J. van der Meij Treur: A language and environment for analysis of dynamics by simulation, *Int. J. Artif. Intell. Tools* **16**(3), 435–464 (2007)
- 50.30 S. Brand, R. Price: *The Economic and Social Costs of Crime, Research Study 217* (Home Office, London 2000)

# 51. Modeling in the Macroeconomics of Financial Markets

Giovanna Magnani

Since the stock price bubble of 1920 and the following 1929–1933 Great Depression, financial crises have become increasingly frequent and globalized. When in the late 2007 the Global Financial Crisis began to show the flawed characteristics of the US capitalist system while spreading throughout all other economies of the world, the ideas of the post–Keynesian School of Economics – a school of economic thought having its origins in *The General Theory* – and in particular, those of Hyman Minsky, became prominent. Minsky's conception of “crisis-prone markets” has become fundamental not only to interpret the 2007 credit crunch – as well as a sort of “ignored prediction” – but also to elucidate the features of the post-modern capitalistic system and its evolution. This chapter begins with a review of Minsky's thought on the *inherently unstable* nature of capitalism. It then examines Irving Fisher's debt deflation model and its application to interpret financial crises and recessions. A reflection on the issues of *finance-led capitalism* in the neo-liberal era completes the first part of the chapter where it is argued that the Minskyian model, if integrated with the social structure of accumulation theory, is very relevant for interpreting the causes and the evolution of the 2007 crisis. The second part of the chapter progresses with the investigation around the constructs of *risk and uncertainty*, and their modeling in Economics and Business Studies.

51.1	<b>The Intrinsic Instability of Financial Markets</b> .....	1066
51.1.1	The Interpretation of the General Theory .....	1066
51.1.2	The Nature of the Capitalist System .....	1068
51.1.3	Cash Flows Analysis and Classification of Financial Postures.....	1069
51.2	<b>The Financial Theory of Investment</b> ...	1071
51.2.1	Aggregate Profit Determination.....	1071
51.2.2	The Two-Price Model and the Determination of Investment .	1072
51.3	<b>The Financial Instability Hypothesis Versus the Efficient Markets Hypothesis</b> .....	1074
51.4	<b>Irving Fisher's Debt-Deflation Model</b> .	1074
51.4.1	Debt Deflation as a Cycle Theory .....	1075
51.4.2	How Debt Deflation Model Fits the Great Depression.....	1077
51.4.3	How Debt-Deflation Model Fits Current Economic Conditions.....	1077
51.5	<b>Policy Implications and the Shareholder Maximization Value Model</b> .....	1079
51.5.1	Stability is Destabilizing.....	1079
51.5.2	From the Debt Deflation Model to Policy Proposals.....	1080
51.5.3	Financialization, Neoliberalization, and the 2008 Crisis .....	1081
51.6	<b>Integrating the Minskyian Model with New Marxists and Social Structure of Accumulation (SSA) Theories</b> .....	1085
51.7	<b>Risk and Uncertainty</b> .....	1086
51.7.1	Models of Risk and Uncertainty in Economics and Business Studies .....	1086
51.7.2	Models of Risk.....	1092
51.7.3	Models of Uncertainty.....	1093
	<b>References</b> .....	1098

## 51.1 The Intrinsic Instability of Financial Markets

In the following sections, we review Hyman Minsky's financial instability hypothesis, starting from his interpretation of John Maynard Keynes' *General Theory*. Subsequently, we examine Minsky's financial theory of investment and the investment theory of the business cycle.

The central idea of Minsky's theory is that *stability is destabilizing* meaning that even a stable economic (capitalistic) environment will endogenously move toward unstable conditions. Capitalism is seen as prone to crisis, with the origins for economic collapse rooted inside the growth process itself. More specifically, a capitalist economy with sophisticated financial institutions is inherently flawed because of its cyclical nature, where short- and long-term expectations are unstable because of the presence of uncertainty.

During an economic downturn, banks and other financial institutions act more conservatively in their lending, as such practices produce efficient outputs to firms which are in turn taking on less risk (as they have a sustainable level of indebtedness). But during times of growth, firms start to engage in riskier activities borrowing money to accelerate growth to even higher rates as their optimistic forecasts are validated. Economic units become riskier – starting from a situation where they are simply hedge units to evolve into speculative or *ultra-speculative* ones – as financial institutions become laxer in their lending practices. Euphoria spreads over and low interest rates make leveraging affordable.

An economic system increasingly based on speculative units is going to be *intrinsically* much exposed to financial crises because it becomes more fragile; thus, any small fluctuation in financial markets (in particular an increase in interest rates) can have great rebounds on real variables.

Indeed, the ability to repay debt and to refinance business is a function of expected future profits. These latter are a function of investment and in turn, the ability to finance future investments is strictly correlated to the expectations that those investments will be able both to repay debts and to refinance business.

As the vicious spiral of increased speculative and ultra-speculative lending positions carries on, units which face liquidity problems caused by their over-indebtedness are forced to refinance or to sell their activities at declining prices (debt-deflation). This leads to falling investment, demand, and profits. Whenever over-indebted investors are forced to sell even their less-speculative positions to make good on their loans; at this point, a major sell-off begins and markets create a severe demand for cash: an event that has come to be known as a *Minsky moment*. The crisis of 2007

has been classified by many commentators as a *Minsky moment*. Paul McCulley, a bond fund director at the Pacific Investment Company, has been the first who coined the term following Hyman Minsky's financial instability hypothesis. The term has been subsequently clarified by George Magnus, a senior economic advisor at the global investment bank UBS. According to Magnus, a *Minsky moment* in financial markets is the point at which “credit supply starts to dry up [...] systemic risk emerges and the central bank is obliged to intervene” [51.1]. This stage is first characterized by “a prolonged period of rapid acceleration of debt” in which the most traditional debt is replaced by new debt borrowed to repay the already existing one. Then “the moment’ occurs when lenders became more and more cautious or restrictive” and at this point “the risks of systemic economic contraction and asset depreciation become all too vivid” [51.2].

According to Minsky, two sets of solutions emerge as necessary in order to reverse the income decline caused by a financial shock. First, central banks should act as lenders-of-last-resort (the so-called *big bank*) and second, the governments should run high deficits that can sustain firms' profits, employment, and final demand even if income has declined (the so-called *big government*). Countercyclical government spending, which constitutes a significant share of aggregate demand, can reverse the tendency toward debt-deflation that emerges during the crisis. Summing up, the bigger the government, the greater the stability of the economy.

### 51.1.1 The Interpretation of the General Theory

During the 1980s, when mainstream macroeconomic research started to study the role of finance within the economic system, Hyman P. Minsky's theories became of foremost importance. Minsky argued that the structure of the contemporary capitalism is made of exceedingly complex financial arrangements. He indeed preferred to call himself a *financial Keynesian* rather than a post-Keynesian because of his aim to clarify and extend Keynes' theories including the complex financial relations, markets, and institutions which characterize the contemporary capitalist structure. Indeed, Minsky's *financial instability hypothesis* – as it is generally called in the Academia – derives from his studies of Keynes, based in particular on *The General Theory of Employment, Interest and Money* [51.3].

His *financial Keynesian* perspective contrasts sharply with that in the mainstream macroeconomic de-

bates of the 1960s and 1970s where the role of financial relations was considered to be limited. The financial instability hypothesis arose out of an attempt by Minsky to understand Keynes in light of the extreme financial disturbance of the 1920–1930 decade which led to the Great Depression of 1929–1933. Such an interpretation is alternative to the one which led to the birth of the so-called *neoclassical synthesis* mainly developed by John Hicks and popularized by the mathematical economist Paul Samuelson, in that it stresses Keynes' explanation of complex capitalist economy's behavior made of sophisticated financial institutions. As mentioned already, Minsky believed that such a kind of economic system tends to be *inherently flawed*. This flawed feature is related to its cyclical nature, meaning that it is made of a succession of transitory phases where economic behaviors change and where, according to this definition, the system cannot by its own processes continuously sustain full employment [51.4].

With respect to the financial structure within a capitalist economy, we can identify the presence of (i) *private portfolios*, made of real-capital assets or speculative financial assets – and (ii) *banks*, generally defined as institutions specialized in finance. Monetary and financial institutions determine the way in which funds required for gathering stocks of capital assets and those needed for the production of new ones are obtained. Following Keynes, Minsky agrees with the assumption that the proximate cause of the transitory nature of each cyclical phase is the instability of investment, but the deeper cause is the *instability* related to *portfolios decisions* and *compositions*, and to *financial interrelations*. In order to better frame such an argument, we need to illustrate its foundation on a specific modeling of the construct of *uncertainty*. Uncertainty in this context means that future is not predictable, and therefore changing views about what will happen affect people's choices about portfolios compositions, financial decisions, and relations in general. The presence of uncertainty makes the formation of short- and long-term expectations precarious. Since short-term expectations are for instance – from firms' point of view – the basis for current production decisions whereas long-term ones affect investment decisions, a variation in one or both can change the equilibrium conditions between production and investment, as well as agents' portfolio choices.

While *classical economics* (i. e., the school of economic thought born in the eighteenth century with Adam Smith, David Ricardo, and John Stuart Mill) and the *neoclassical synthesis* (i. e., the consensus view of macroeconomics which emerged in the mid-1950s in the United States) are based upon a *barter paradigm*, the Keynesian theory of investment rests upon a *speculative-financial paradigm* [51.5, p. 21]:

“[...] in the General Theory Keynes adopts a City or Wall Street paradigm: the economy is viewed from the board room of a Wall Street investment bank.”

By adopting the City paradigm, Keynes assumes that the economy is a sophisticated monetary system where *money* has not only simple trade functions, but it also acts as a *financial veil* between “the real asset and the wealth owner.” As pointed out by *Davidson* [51.6], Keynes specifically argued that money has also another important function: since people know that they cannot predict future, society has attempted to create institutions that will provide people with some control over their uncertain economic destiny; the use of money permits individual to have some kind of control over their cash inflows (CIFs) and outflows and so, of their monetary economic future. Keynes' conception of money is clearly stated by his sentence [51.7, p. 169]:

“There is a multitude of real assets in the world which constitutes our capital wealth- buildings, stocks of commodities, goods in the course of manufacture and of transport, and so forth. The nominal owners of these assets, however, have not infrequently borrowed *money* in order to become possessed of them. To a corresponding extent the actual owners of wealth have claims, not on real assets, but on money. A considerable part of this *financing* takes place through the banking system, which interposes its guarantee between its depositors who lend money and its borrowing customers to whom it loans money wherewith to finance the purchase of real assets. The interposition of this veil of money between the real asset and the wealth owner is a specially marked characteristic of the modern world.”

Looking at the economy from a “Wall Street board room”, Minsky sees a world made of commitments to pay cash in the future in exchange for cash today, where the most crucial type of commitment is *business debt*, assuming that it is the very peculiar component of a modern capitalist economy. The ability of firms to repay debt and to refinance businesses in order to make new investments is a function of the expected future profits (gross profits), where gross profits themselves are, in turn, largely determined by investments. This means that the ability to finance future investments is strictly correlated with the expectation that future investment will be higher enough so that future cash flows will be able both to repay debt and to refinance the business. Minsky argues that an economic system with private debts is deeply influenced by agents' (firms and bankers) views about the future course of investment and thus the determination of units' liability structures

and in the end, future production, income, and employment. It is uncertainty – the Keynesian type of – for which “there is no scientific basis on which to form any calculable profitability whatever. We simply do not know” [51.8, p. 214]. In the next sections, we will progressively illustrate Minsky’s financial theory of investment and his investment theory of the business cycle, in light of the assumptions he made upon Keynes’ vision of behavior in an uncertain capitalistic economy. But first we need to portray Minsky’s model of the capitalist system itself.

### 51.1.2 The Nature of the Capitalist System

#### The Financial Nature of Capitalism and Money Supply Endogeneity

Minsky recognizes different distinct forms of capitalist systems, but at the same time he argues that they share similar characteristics. According to his analysis [51.9], our capitalist economy is first of all a monetary production system where money is at the center of economic decisions, and where the aim of any economic activity is to get a monetary gain. Following Keynes, money is understood as a *veil* that *camouflages* ultimate ownership of wealth [51.10] and that any economic theory (i. e., the neoclassical synthesis which considers money only as a *bartering veil*) which ignores this major aspect of money cannot be considered a useful tool to design appropriate policies. This kind of veil is different from the one in *quantity theory of money*. The veil of quantity theory basically says that the price level in an economy depends on how much money is in the economy: when money supply changes, the real economy does not because when money supply changes by a certain amount, everything else does as well. If it doubles, then prices double: money supply is the force that changes the price level

$$MV = PY, \quad (51.1)$$

$$M \rightarrow P, \quad (51.2)$$

where  $M$  is the money supply,  $V$  is money’s velocity of circulation,  $P$  is the general price level, and  $Y$  is the real value of national output (i. e., GDP).

The idea that money supply has to be treated as endogenous is largely accepted by post-Keynesian theories: accordingly, credit volume is generated endogenously inside the private sector because of banks introducing money into the economy directly by financing current productive activities. Credit is brought into existence by banks (without preventive savings acts), through the creation of deposits. The financing of production is therefore linked to banks’ availability and capacity of injecting money into the economic system through investment acts [51.11]. Hence, money is not

only a medium of exchange or a store of value, but also the financial mean through which banks allow credit to units. In such a prospective, banks’ credit potential is not fixed by the *monetary base* (i. e., a measure of the money supply that is the sum of all the money in circulation plus deposits, and commercial banks’ reserves with the Central Bank). Money supply is a function of the interaction between firms, banks, workers, and financial markets.

Within the Minskyan framework, money has an *endogenous* nature: it is a type of *obligation* appearing on the market as investment or current production activities are in the process of being financed. It has to be noted that according to Minsky, money supply endogeneity is compatible with the *liquidity preference* hypothesis [51.12] since he extends his financial instability hypothesis to banks and credit markets, analyzing creditor and debtor risk throughout all the business cycles phases [51.11], thus taking into consideration the liquidity preference functions of each actor. Liquidity preference means that the less liquid (i. e., easily tradable) the investment is, the greater the premium demanded by investors in the face of greater risk.

One of the most innovative arguments in Minsky’s theory is that when demand for money stimulates supply, it not only increases money velocity of circulation, but also causes the introduction of *financial innovations*. The release of less liquid and more onerous financial instruments (called *quasi-money*) is the reason why during boom periods the amount of money in circulation rises. Thus, money supply is not *given* once and for all: the amount changes during the economic cycle (during a boom period money demand will stimulate for an increasing supply). Although the effective quantity of money in circulation is demand driven, it is *not* unbounded. Money supply is positively related to the rate of interest with a given financial structure, but it increases when banks and financial intermediaries *squeeze* inactive money and issue new substitutes for money, reacting to monetary policy or just exploiting profit opportunities during the cycle [51.13].

#### The Role of Banks and of Financial Innovations

Banks are an essential component of the system at all stages of the economic process. In addition, bankers are both a source of dynamism and destabilization and, therefore, need to be managed [51.10]. Banks can be thought as all kind of institutions that contribute, directly or indirectly, to the financing and funding of economic activities. At the same time the aim of the banking activity is that of maximizing expected profits. Banks are considered by *Minsky* truly speculative enterprises since the maturity of their debts is far shorter than

the maturity of their assets, so that there is a refinancing cost upon which they have to speculate [51.5]. By allowing loans, banks are subject to credit risk; therefore, they first of all assess firms' expected profitability. Banks will profit from their lending activity only if firms are able to meet their debt payment commitments. Hence, bankers' expectations on entrepreneurship are pivotal in determining the economic course of events (even if also bankers' expectations are formed in a world made of uncertainty), since they are the prerequisite to obtain funds.

In order to maximize profits, banks attempt to make an increasingly efficient use of their lending potential, and push for an increase in loans. At the same time, similar to commercial banks, financial intermediaries that are profit-seeking agents, also constantly try to extend credit, by financing new positions. Financial intermediaries issue new kinds of financial instruments which serve as *effective* money without increasing required reserves [51.14]. A given amount of bank loans and demand deposits supports a higher volume of finance to the whole economy.

### The Centrality of Profits

In arguing about the financial nature of modern capitalism, and in particular of investment, Minsky recognizes three ways in which the former affects the latter. First, in financing of positions in the existing stock of capital assets; second, in financing investment and production/distribution activities; third, in meeting payment commitments as stated on financial contracts. He argues that the techniques employed to finance positions in capital assets affect asset prices and that prices reflect expected profits that can be earned by using capital assets into production and the payment commitments that have to be agreed to finance ownership. Since a debt involves an exchange of money today for promises to pay money tomorrow, the smaller the amount that has to be promised to obtain current money to finance capital assets, the greater the demand for such capital assets. Supply of capital assets is fixed and therefore an increase in their demand will cause an increase in their price.

It is necessary for *quasi-rents* – that is, the difference between the total revenue from selling output produced with the aid of capital assets and costs associated with that production – to be greater than future payment commitments (at least on a relevant time horizon), so that it is suitable to proceed to the production and acquisition of capital activities. The maintenance over time of a positive gap between cash-inflows and cash-outflows depends on the realization of expectations on future profits, but also on the debt-structure inherited from the past and on the course of financial costs in the current period [51.14]. Since we can say that

profits are inflows used by firms in order to meet debt commitments, and expected profits are the stimulus for making investments and determine the possibility to renew existing debts or to generate new ones – expectations on future profits are the determinants of current investment and financing decisions. The centrality of profits, seen as the engine for growth and financial dynamics (debt structures), is confirmed by the use of *Kalecki's equations of profit* by Minsky (Sect. 51.2.1).

### 51.1.3 Cash Flows Analysis and Classification of Financial Postures

All economic activities (real and financial assets) are acquired by firms through a combination of equity and short/medium long-term debt. Since debt implies a commitment in order to repay principal and interests, periodical cash outflows (COFs) are generated. On the other hand, investments and productive activities are expected to give birth to CIF. Note that current profits are a sign of the generation and of the amount of future profits and thanks to this, firms decide or not for greater indebtedness.

An economic unit – or the economy itself – needs to generate enough CIF or to have enough idle cash balance (ICB), in order to meet its COF [51.15]. If the net cash inflow (NCF) is negative and no ICB is available, an economic unit will be considered illiquid or even insolvent [51.9]. We can identify different sources and uses of funds for each economic activity, also depending on the level of analysis (business unit, sector, and whole economy): possible sources and main NCF are different. Our economy is one in which borrowing and lending on the margins of safety are commonplace. Each financing transaction involves an exchange of money today for money later on; the future cash receipts which will enable the borrower to fulfill the money-tomorrow parts of the contract are conditional upon the performance of the economy over a longer or shorter period. All economic activities involve income transactions, balance-sheet transactions, and portfolio transactions; we can classify sources of CIF into three categories [51.4, 16]. Cash flows from income operations are those deriving from productive activity or from investment: they are wages, salaries, and profits. Cash flows from balance sheet operations are those deriving from financial contracts (interest and principal). Cash flows from portfolio activities are those related to transactions involving real assets or financial ones (assets acquisitions or the issue of liabilities). It is the relative weight of these cash flows that determines the exposure grade of a system to a financial crisis. In particular, financial instability derives from the propagation of the practice through which a unit finances

long-term activities by underwriting new debt (position refinancing). The model in which Minsky investment theory is developed is a closed economy with a very small State (that is to say, public intervention has no relevant dimension), where capital accumulations implies debt, which in turn is constituted first of all by banking finance. There are two types of financing: the short term, which is to finance productive activities and the long-term, needed to finance capital assets (in general all those illiquid postures in capital assets such as plants and equipment, buildings, etc.). Usually these financings are made partly with own funds and partly with external ones; we can assume that banks finance current production and that the present stock of capital assets can be financed also through other financial intermediaries or directly by private savers, with instruments whose liquidity is directly linked to their convertibility in bank money [51.14]. As we mentioned, Minsky defines money as a particular type of obligation, created as production process activities, investments and postures on capital assets are getting financed.

It is important to remark that, according to Minsky, agents' ultimate purpose is not material production itself, but accumulation of money (production of money with money) through speculation, by holding portfolios of activities that are not quickly convertible into money which means, by engaging in *financial postures* (i. e., any capital value from which to expect future profits for instance machines, but also securities). There are three kinds of financial postures for the acquisition and ownership of activities not quickly convertible into money, including investment goods: hedge, speculative and *Ponzi*. Minsky argues that the stability of the economic system is deeply dependent on the mix of these financial postures: the greater the presence of hedge financing, the greater the stability of the structure, while on the other hand, the presence of excessive speculative and *Ponzi financing* will increase the tendency of an economy to become more and more unstable.

A necessary though not sufficient condition for their financial profitability is that the expected gross capital income exceeds the total payment commitments over time. In particular [51.9]

$$\begin{aligned} \text{Gross Capital Income} &= \\ &\text{Total Receipts From Operations} \\ &\quad - \text{Current Labor and Material Costs} \\ &\text{and} \\ \text{Gross Capital Income} &= \\ &\text{Principal and Interest Due on Debts} \\ &\quad + \text{Income Taxes} + \text{Owners Income} . \end{aligned}$$

This means that the total receipts of a business firm can be divided into the payments for current labor and pur-

chased inputs and a residual, gross capital income that is available to pay income taxes, the principal and interests on debts, and to be used by the owners.

Back to our sorting, a hedge financing unit implies that CIF (quasi-rents) from participation in income production exceeds debt payment commitments (principal plus interests) in every period and for each interest rate. Hence, a sharp rise in interest rates cannot reverse the condition where the actual value of capital assets exceeds the book value of debts. The actual value of activities is always non-negative even in the presence of very large changes in interest rates. A hedge unit is expected to be very liquid; there is no expectation that cash flows from operations will be lower than balance-sheet commitments at any time; therefore, there is no expectation that one will have to refinance. However, some idle cash and superfluous assets are kept aside to cover possible disappointments in expectations [51.17]. As time passes debt decreases, while equity and idle cash (retained for precautionary purposes) increase.

A speculative financing unit implies that cash flows from participation in income production, when totaled over the foreseeable future, exceed outstanding debt, but in the near term it is expected that cash flows from operations will not cover the capital component induced by debt (even if they are always sufficient to cover interests); therefore at least in the short term negative cash flows are expected. A refinancing is thus necessary, but only for the capital component of debt, thereby exposing the unit to interest rates fluctuations. A speculative unit may get more easily into troubles than a hedge unit because for a certain set of interest rates the activity actual value is positive, but a sudden increase can convert profits into losses.

A Ponzi finance unit is a speculative financing unit for which the net income portion from short-term cash flows is less than near term interest payments on debt. A Ponzi finance unit implies that cash flows from operations do not even cover cash disbursements due to interest payments. The unit bet on a favorable variation under market conditions and/or on gaining an exceptional profit that allows compensation of initially accumulated losses (e.g., who speculates at bullish trend). Both speculative and Ponzi units can fulfill their payment commitments only by borrowing again (or by disposing assets). The amount of refinancing of Ponzi units is greater than that of the speculative units, since the former must refinance in order to cover both principal and interest. In this case, the refinancing process is needed both to cover not only the capital components of balance sheet cash commitments but also their income components [51.16]. Debt increases over time till the (expected) realization of a final profit (that reverses the sign of the investment expected actual value). It has to be

noted that Ponzi financing can be hard to detect because it can be hidden by (creative) accounting practices.

We can identify two kinds of risk: *economic risk* and *financial risk*. Economic risk is linked to the possibility not to come up to firm's expectations about future profits, because of a sudden unforeseen worsening of commodities' markets conditions. Economic risk can arise because of lower incomes or higher expenditures than expected. Causes may be several, for instance a hike in the price for raw materials, lapsing of deadlines for construction of a new operating facility, disruptions in a production process, emergence of a serious competitor on the market, the loss of key personnel, the change of a political regime, natural disasters, etc. Financial risk (which is an umbrella term for any risk associated with any form of financing) is related to the possibility of unexpected worsening of financial markets conditions. Risk may be taken as *downside risk*, the difference between the actual return and the expected return (when the actual return is lower), or the uncertainty of that return. Risk related to

an investment is often called *investment risk*. Risk related to a company's cash flow is called *business risk*. The science that has evolved around managing market and financial risk under the general title of modern portfolio theory was initiated by Harry Markowitz in 1952 with his article, *Portfolio Selection*. We can therefore argue that hedge financing units are exposed only to economic risk, while speculative and Ponzi financing units are exposed both to economic and financial risks, thus making these two positions unstable and fragile.

As we have mentioned, the stability of an economy depends upon the mixture of hedge, speculative, and Ponzi financing, and in turn, the weight of Ponzi and *near-Ponzi* speculative finance is conducive to instability. During a tranquil period of economic growth, the weight of speculative and Ponzi financing can increase a lot because of euphoric expectations about the future. Thus, for these kinds of positions a rise in interest rates can transform a positive net worth into a negative net worth; and these together can intrinsically produce the conditions for interest rates to swing.

## 51.2 The Financial Theory of Investment

### 51.2.1 Aggregate Profit Determination

"A capitalist economy works well as an investing economy, for investment generates profits" [51.18, p. 104]. Profit expectations make debt financing possible: investment takes place because positive future profits are expected, but these profits will be forthcoming only if future investment takes places. This means that investments are the key variable in order to determine whether or not debt payment commitments will be met.

In order to determinate aggregate income, *Min-sky* [51.10, pp. 515–516] builds on Kalecki's equation of aggregate profit [51.19], which assumes a world where investment takes place, workers spend all their wages in consumption, and capitalists' profits are all saved.

$$C = W_c N_c + W_I N_I, \quad (51.3)$$

with  $C$  = consumption,  $W_c$  = money wage rate in the production of consumption goods,  $W_I$  = money wage rate in the production of investment goods,  $N_c$  = employment in consumption goods,  $N_I$  = employment in investment goods,  $W_c N_c$  = wage bill in the production of consumption goods, and  $W_I N_I$  = wage bill in the production of investment goods.

Assuming that  $P_c Q_c$  is consumption ( $C$ ) summed over all goods, then

$$P_c Q_c = W_c N_c + W_I N_I, \quad \text{so that} \quad (51.4)$$

$$\pi_c = P_c Q_c - W_c N_c = W_I N_I, \quad (51.5)$$

where  $\pi_c$  are profits in consumption goods. Thus, profits in consumption goods equal wages in investment goods.

Profits in investment goods  $\pi_I$  are instead

$$\pi_I = P_I Q_I - W_I N_I = \pi_I. \quad (51.6)$$

Since  $\pi_I + \pi_c = \pi$ , and  $P_I Q_I = I$ , we have

$$\pi = W_I N_I + \pi_I = I. \quad (51.7)$$

Kalecki's result can be expanded to

$$\pi = I + Df \quad (51.8)$$

if government is introduced ( $Df$  = government deficit), if consumption out of profits ( $C$ ) and savings out of wages ( $s$ ) are allowed, then we have

$$\pi = I + Df + C\pi - sW \quad (51.9)$$

and

$$\pi = I + Df + C\pi - sW + BPS \quad (51.10)$$

if the economy is open ( $BPS$  = balance of payments).

Total profits are the sum of capitalist consumption, investment, public deficit, net external surplus (exports minus imports) minus workers' savings. Today's profits depend on today's investment; they equal investment in a world in which we imagine work-



ers consuming all their wage and capitalists saving all profit. The curve that represents financing with internal funds is a function of realized investments, from which gross profits and so internal resources for further investments depend. At a *micro* level, firms' investments depend on *expected* profits (and on the level of interest rates as well), while at a macro level *realized* profits depend on the overall level of business firms' investments.

Since both investment and positions in capital assets must be financed, as a result financing terms influence prices of capital assets, the effective demand for investment and the supply price of investment outputs [51.16]. Once the determinants of investment are understood, a complete theory of financial instability can be constructed.

### 51.2.2 The Two-Price Model and the Determination of Investment

The two-price model is the analytical tool by which Minsky integrates his theory of money and finance into his theory of investment. From a microeconomic perspective, it is necessary to consider the price of *capital assets* as a key variable for investment. Such price embodies firms' changeable profit expectations, making investment an unsteady component of demand. According to Keynesian theory, the investment is governed by the marginal efficiency of capital, which is the marginal rate of profit expected from investment. The calculation of investment opportunity cost is carried out by comparing the marginal efficiency of capital with the interest rate. The limit of investment corresponds to the point at which, for increasing interest rates, marginal efficiency decreases. The Keynesian approach to investment theory was criticized by Kalecki who pointed out that it only allows us to determine the ex-post level of investment, but says nothing about ex-ante investment choices. In fact, entrepreneurs make their calculations of investment opportunity cost, by taking into account the present prices of capital goods and not the future ones. Investment and ownership of capital assets are indeed undertaken in the expectation that they will produce money. Financial assets (which are commitments to pay cash over some period) have current prices, which are the capitalization of the future cash flows as laid down in contracts [51.16].

The first set of prices is *prices of current output* and includes the price of consumption goods and the price of investment goods. With respect to consumption goods, producers have given short-term expectations about costs and demand. They fix prices to recover direct costs (the main component being, in the aggregate

closed economy, money wages) and to earn a residual: deducting costs, they obtain gross profits. With respect to investment goods, we have the supply price of capital assets ( $p_i$ ) that is determined by adding a mark-up to direct costs of production (a price sufficient to induce a supplier to provide new capital assets). If external (borrowed) funds are involved, then the supply price of capital also includes finance costs (interest rate, fees, etc.), such increase is due to the presence of *lender's risk*. The second set of prices is *prices of capital assets*, which in particular affects the determination of the demand price ( $p_k$ ) of investment goods. Demand price of capital assets (securities)  $p_k$  is the maximum price that the investor is willing to pay. Minsky argued that it is strictly dependent on the amount of borrowings of external funds required: the higher is the weight of external finance, the greater the risk of insolvency for the buyer. That is why a *borrower's risk* has to be incorporated in the demand price for capital assets. The amount of an investment depends on the ratio between the demand price of investment goods and the supply price of new capital goods.

The demand curve of a capital good is positively correlated with the long-term expected profits obtained from it and with the money supply, while is negatively correlated with the debtor's risk (more on debtor and creditor risk later on in this section). The supply curve of a capital good is positively correlated with production costs, with the producer's short-term profits expectations and with the creditor's risk. As we can see in Fig. 51.1, investments increase until  $p_k \geq p_i$ , where  $I = I_1$ .

For each investing unit, and for investment in general, a mix of gross retained earnings and external finance is needed [51.16]. The amount of investment

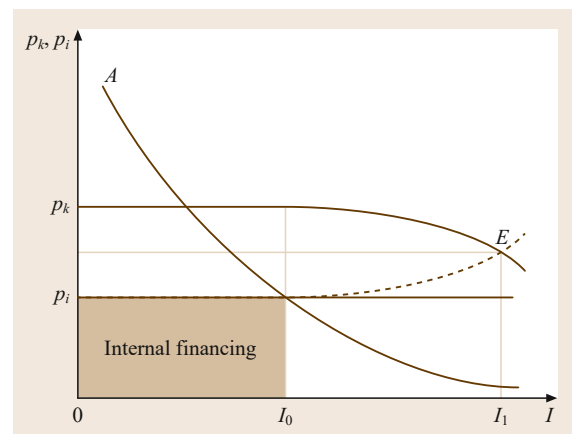
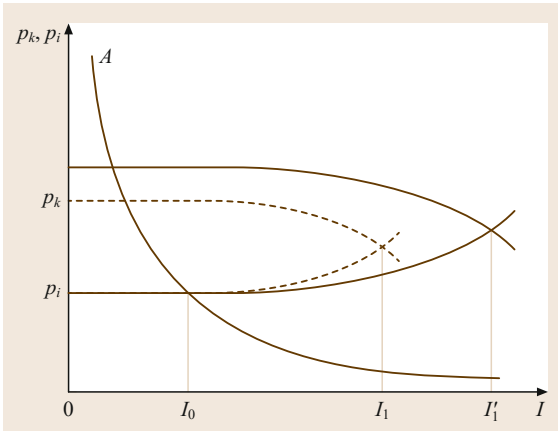


Fig. 51.1 The determination of the (real) investment level in the presence of risks (after [51.20])

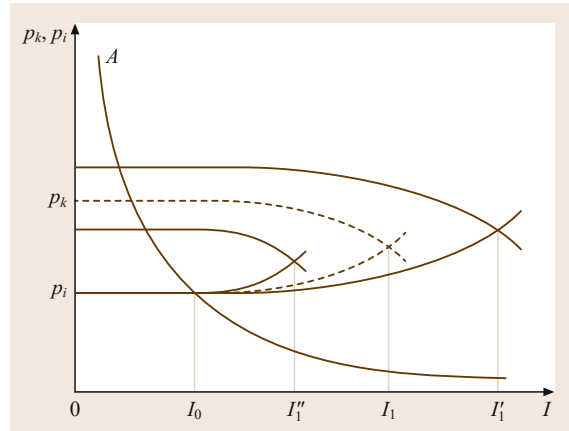


**Fig. 51.2** Investment during the boom phase (after [51.20])

(at a micro level) is determined in the way shown in Fig. 51.1, where  $p_k$  and  $p_i$  represent, respectively, the demand price the investor is ready to pay for a certain type of investment good, and the supply price that is the price at which the producer of the investment good offers the latter. The equilateral hyperbola  $A$  represents all possible levels of investments that can be financed given the amount of *internal* funds, for each level of capital goods prices. Self-financing  $= p_i I_0$ . Given the supply unit price of capital goods  $p_i$  and the amount of internal resources for self-financing, beyond  $I = I_0$ , it is necessary to resort to external financing. It follows that, beyond such level  $I_0$  the demand and supply curves shapes modify, because of the rising of the borrower and lender risk (both increasing in  $I$ ).

The *borrower's risk* is the *subjective* risk linked to the possibility that an increase in the amount of financing made by using external funds reduces the safety margin (due to increased illiquidity), thus reducing the investor's portfolio diversification. Therefore, beyond the internal financing threshold, the demand price  $p_k$ , will reduce as investment grows. The *lender's risk*, which is subjective as well, objectifies in the form of interest expenses and increasing burdens. Hence, in the effective capital asset supply price must also include the (increasing) cost of external financing. Kalecki's principle of increasing risk is therefore intrinsically stated in Minsky's analysis. The principle states that, as expected external funds increase, there is an expected increase in the debt–equity ratio, which affects the perceived risk of engaging in an economic activity with external funds. The profit equation and the principle of increasing risk highlight the importance of financial factors in the determination of investment.

The determinants of investment at a micro level are then dependent on:



**Fig. 51.3** Investment during the recession phase (after [51.20])

- The investor's (buyer of capital goods) profit expectations; hence, on quasi-rents, which in turn influence the demand curve for capital goods position.
- The capital goods producer's production costs and mark up, which influence the capital goods supply curve position.
- The borrower and lender's perceived risks, which act respectively on both the demand and supply curves (for  $I > I_0$ ). The greater the borrower's and lender's risks are, the lower is the investment level (*ceteris paribus*).
- The inherited leverage (debt/equity ratio) level from past operations, which influences the availability of self-financing resources.

We can show how both demand and supply curves behave during the economic cycle. During the growth phase, we can observe (Fig. 51.2) an investment boom:

- Optimistic profit expectations and reduced perceived risks cause an increase in the amount of demand for new capital goods, causing the supply curve to shift downward.
- A multiplication of financing activity, a reduction in interest rates and thus an increase in indebtedness.
- A general increase of investments, the beginning of an economic boom.

During the recession phase (Fig. 51.3):

- Profit expectations are pessimist, and perceived risks increase.
- The amount of financing activities reduces and/or interest rates increase.
- Investment level reduces triggering an economic recession.

### 51.3 The Financial Instability Hypothesis Versus the Efficient Markets Hypothesis

The two fundamental propositions of the financial instability hypothesis are the following [51.16]:

1. Capitalist market mechanisms cannot lead to sustained, stable-price, full-employment equilibrium.
2. Serious business cycles are due to the financial attributes that are fundamental to capitalism.

These two propositions are in sharp contrast with the neoclassical theory which holds that unless exogenous forces disturb economy, the system will remain self-sustaining, with stable-prices and full-employment. Minsky criticizes standard economic theory because of its bad performance in explaining reality. According to him, the *crisis* – in economic theory – has two facets: *logical holes* appeared in conventional theory and the impossibility for conventional theory to explain financial crises. The first failure lies in its complete non adherence to empirical evidence. Classical economic theory, which has many variants such as *theory of efficient markets, classical or neoclassical theory, general equilibrium theory, mainstream economic theory* (including old and new Keynesian theory) lies on the mantra that free markets can cure any economic disequilibria that may arise, while government interference has to be considered as a problem. With respect to the second failure, Minsky criticizes the classical efficient market theory that assumes all agents to be able to *know* their future *intertemporal* budget constraints and act accordingly thus avoiding loan defaults, insolvency, and bankruptcy events. According to such a theory, future can be predicted efficiently thanks to the collection and analysis of reliable information on both

the probability of events that have already occurred and the probability of events that will occur in the future. In such a context perfect information is considered to be available to all decision makers. Rational agents are therefore not involved in insolvency problems thus, according to such theories, financial crises are impossible to happen.

In *Minsky's* theoretical framework, agents make arbitrages between the assets (productive and financial) they own on the basis of their future expected rate of return from income and capital gains [51.16]. Their fundamental speculative *decision* is which assets to keep and which to sell, and how to fund their investments [51.12, 16]. Following Keynes, Minsky explains the mechanism through which human beings use *practical equivalents* as a convention which is retained until it confirms reality, but abandoned when evidence becomes different and in contrast with it. Practical equivalents are the process through which (in a situation of uncertainty) a decision is taken making uncertain propositions equivalent to certain ones. These conventions are used ad hoc for practical purposes. Conventions play a central role in Minsky's analysis. In normal times, there is always a *consensus* that exists and stabilizes the decision-making process [51.21]. Hence, in order to make decisions, agents construct mental models of how they think the economic system works and will work in the future. They know that these are representations of reality that do not replicate the *true* model; economic agents thus know that they can be *systematically* wrong. Errors are possible since the future is uncertain.

### 51.4 Irving Fisher's Debt-Deflation Model

This section deals with Irving Fisher's debt-deflation model, its explanation, and application to interpret financial crises and recessions. After the 2007 crash and financial ruin, Fisher's theories have shown to be fundamental in determining the underlying mechanisms of the crisis.

Fisher's theory of *great depressions* is based on the interaction of an initial situation of over-indebtedness which, in conjunction with a dynamic process of deflation, produces a contracting economy. Fisher explicitly ties loose money to over-indebtedness, triggering speculation and asset bubbles. If over-indebted units hit by an exogenous shock start facing problems with debt

commitments, they begin to liquidate debt through distress selling of their assets at decreasing prices.

Fisher outlines the nine factors which interact with each other to create the process of boom to bust for a Great Depression:

1. Debt liquidation leads to distress selling and to
2. Contraction of deposit currency; this contraction of deposits and of their velocity, caused by distress selling, leads to
3. A fall in the level of prices (a swelling of the dollar)
4. A still greater fall in the net worth of business
5. A fall in profits

6. A reduction in output, trade, and employment. These losses, bankruptcies, and unemployment, lead to
7. Hoarding and
8. A further reduction in the velocity of circulation.

The above eight changes cause (9) a fall in the nominal rate of interest and a rise in the real rates of interest. The way out, according to Fisher, is deflation; the lender-of-last-resort function of central banks and government have to support the financial system through stimulating it with fiscal measures.

### 51.4.1 Debt Deflation as a Cycle Theory

During the peak of the Great Depression of 1929, Irving Fisher wrote *The Debt Deflation Theory of Great Depressions* [51.22], following his previous work *Booms and Depressions* [51.23] with the aim of finding an explanation for those years' dramatic events. As depression worsened, in 1932, Fisher became convinced that the crisis could not be simply interpreted as a downturn in the business cycle, however severe, but that was something radically different which needed a new theoretical explanation. In January 1932, therefore, he began to devise a new theory of great depressions, based on the interaction of two most important factors: (i) an initial situation of over-indebtedness; (ii) a dynamic process of deflation. The main point of the theory is that over-indebtedness acts in conjunction with deflation to produce a contracting economy causing bankruptcies, rising unemployment, and falling profits. Irving Fisher was the first economist to emphasize the potential connections between violent financial crises, which lead to *fire sales* of assets and falling asset prices, with general declines in aggregate demand and the price level.

The author starts with investigating the 1929–1932 debt-deflationary situation by listing (in logical order) the nine factors he recognized to be the cause of such trend, describing each one of them. The first main factor, as we just mentioned, is *over-indebtedness*. According to Fisher, debts are an intrinsic feature of a monetary economy and they are essential both for production and distribution processes; over-indebtedness is that degree of indebtedness that “multiplies *unduly* the chances of becoming insolvent” [51.23, p. 9]. It is a matter that affects both singular individuals and whole communities and it means that debts are too high relatively to other economic factors. As lenders and borrowers have become too incautious and, at a point, over-indebtedness is discovered, *distress selling* is likely to arise. The debtor begins to liquidate some of its assets (both tangible and intangible ones optionally) to run after its debt commitments, and/or debtor's

bank or broker cash-in its collateral; hence, the debtor becomes victim of distress selling either on its own initiative and on initiative of its creditors. The process of distress selling perverts the demand and supply equilibrium because in this particular event sales are not made – as Fisher [51.23] points out – in order to attract the highest possible price. Since distress sellers are being forced to sell, they usually do not receive a price as favorable as if they were able to wait for ideal selling conditions; the effect of a whole community involved in distress selling is a reduction in the general price level.

Linked to distress selling and *stampede liquidation* (as he defines the anxious selling of assets to repay debts), Fisher argues that such a situation has a major effect on the volume of currency (deposit currency) in circulation, shrinking it. The author is referring to debts to commercial banks which are paid by checks out of a deposit account, implying the disappearance of that amount of deposit currency. Thus, the reduction of circulating money is tied to debt-volume, especially debts to commercial banks. Such a disturbance passes to prices, with a consequent reduction in the price level (deflation) in other words, a swelling of the currency. This situation can be simply explained as Fisher did in the *The Purchasing Power of Money* [51.24, p. 25]:

“From the mere fact, therefore, that the money spent for goods must equal the quantities of those goods multiplied by their prices, it follows that the level of prices must rise or fall according to changes in the quantity of money, *unless* there are changes in its velocity of circulation or in the quantities of goods exchanged.”

Fisher uses the *equation of exchange* to assume that an increase in the quantity of circulating money has some tendency to raise the price level, and vice versa. In any given year,  $PT = MV$ : the price level multiplied by the yearly volume of trade is equal to the money in circulation multiplied by the number of times it circulates in a year.

Though, if we consider debts in *real* terms, we can realize how each unit of money that has to be paid by the debtor becomes bigger: he receives less money for its liquidating assets while at the same time owning the same amount of money as before on its debts. The acts of liquidation actually enlarge the real value of debts instead of reducing them. Fisher refers to this effect by introducing the third factor, *the price level* (deflation) and by pointing out that both creditors and debtors do suffer from what in literature is called *Money Illusion*, meaning that a few people are able to recognize the real significance of a unit of money by measuring how many goods that a unit can buy – and therefore, understanding

the real value of money and not the mere nominal one. Because of this *misunderstanding* [51.23, p. 18]:

“The creditor is unaware of receiving more than he is properly entitled to, and the debtor is unaware of paying more than he properly owes. One gains and the other suffers [...]”

Fisher argues that the point of the economic cycle at which stampede liquidation starts and a mass payment by the weaker debtors – which in turns produces the swelling of the currency damaging stronger debtors as well as weaker ones – takes place, a vicious spiral begins and a depression is on its way. Once the mass payment begins, each individual is forced to do so: if he stays out, the mass liquidation will swell its whole debt instead of only part of it. The two most important processes of debt-deflation theory are those Fisher calls *The Debt Disease* (over-indebtedness) and *The Dollar Disease* (a swelling currency), while decreased currency volume is just a link between the two [51.22, p. 341]:

“Disturbances in these two factors – debt and the purchasing power of the monetary unit – will set up serious disturbances in all, or nearly all, other economic variables.”

The fourth main factor is *Net Worth*: the fall of prices reduces the value of business assets while liabilities remain fixed, hence net worth value shrinks. The fifth factor is linked to deflation as well, and it is *Profits*. Because of the decreased value of receipts relative to that of expenditures (which we can assume as almost fixed) profits are reduced, and sometimes turned into losses. A depression might be defined as the contraction of net worth and profits. Variation in profits (and/or in profits' expectations) causes the variation in the general strategy of the firm, in particular with respect to current production and investment decisions. The sixth main factor therefore contains the three variables that are consequently affected by a fall in profits, which are *Production, Trade, and Employment* that come out of worsened. All this factors taken together produce a lack of confidence and pessimism (the seventh factor, *Optimism and Pessimism*) that translates into a general rush toward money: phenomena of hoarding thus multiply. The velocity of circulation of money reduces since people are scared and begin to spend less and slowly; thus deflation worsens again and consumption contracts even further (If, e.g., the currency is halved and it now also moves at half velocity than in the past, it will do only a quarter of its former work. Prices and/or trade must contract in the same degree) The effort by each agent to improve his own position leads to a worsening of the overall situation: “Every man who hoards does it

for his own protection; yet by hoarding he aggravates the very condition that started his fear” [51.23, p. 36]. The ninth and last factor is the *Rate of Interest*: since every debt bears interest, we can expect that a *cycle-tendency* in the former will produce a *cycle-tendency* in the latter. The disturbance consists in the fact that during depression nominal interest rates remains relatively low, while real interest rates are actually rising. We know that from the Fisher equation, we have  $i = r + \pi$ , where  $i$  is the nominal interest rate,  $r$  is the real interest rate, and  $\pi$  the inflation level. Since during a depression currency is swelling (deflation), the debt that for instance was contracted last year worthing 100\$, with a nominal interest rate of 5%, today will worth 106 of last year's \$; thus, if  $\pi = -1$ , the real interest rate is not 5, but 6%.

The Debt-Deflation Theory of Great Depressions is the explanation Fisher gives to describe the apparent prevailing boom-bust pattern of the economic cycle. With his *Cycle Theory*, Fisher defines different types of economic cycles and their possible causes. He dismisses the idea of the business cycle as being a single, self-generating one as a myth. In its place he introduces the notion that there are many interacting cycles within the economy, also interacting with noncyclical forces such as growth and chaotic tendencies. He divides cyclical tendencies into two types, forced cycles and free cycles. Forced cycles are imposed onto the economy by outside forces, such as the yearly season cycle, day–night cycle, and monthly and weekly cycles imposed by religion and custom. The free cycle is self-generating and is commonly thought of when referring to the business cycle.

According to Fisher, each case of over-indebtedness has its own *set of starters* that, thanks to their investment decisions, initially triggers a boom economy. Over-indebtedness causes can be various; according to Fisher, the most common one appears to be the presence of new opportunities and/or big profits perspectives, as compared with ordinary actual profits and interest. Investment opportunities can be represented by new inventions, new industries, and development of new resources, opening of new lands or new markets. Easy money is the biggest cause of over-borrowing [51.22, p. 349]:

“The public psychology of going into debt for gain passes through several more or less distinct phases: (a) the lure of big prospective dividends or gains in income in the remote future; (b) the hope of selling at a profit, and realizing a capital gain in the immediate future; (c) the vogue of reckless promotions, taking advantage of the habituation of the public to great expectations; (d) the development of down-

right fraud, imposing on a public which had grown credulous and gullible.”

The psychological factor is a critical one: agents have the perception that a *new era* has begun, and they start to be less cautious and risk-averse. They overlook profits perspectives and over-indebt themselves, thus giving start to the vicious cycle that sooner will trigger a depression.

### 51.4.2 How Debt Deflation Model Fits the Great Depression

According to Fisher, the debt deflation model well suit what happened in the United States and in Europe between 1929 and 1932. The World War propelled a series of inventions – not only destructive ones – and a number of technological improvements (especially in the manufacturing sector) begun to spread up and the American economy faced an investment boom. This situation encouraged many firms to borrow heavily in the expectation of higher profits, including those in the farming sector, stimulated by the sharply rising demand for food. The process that led to over-indebtedness found its *starters* in those investors incentivized by innovations on the expectation of high future profits.

In corporations, the practice of preferring investment in equities rather than bonds became the rule. The Wall Street crash was the detonator, triggering the downward spiral predicted by the debt deflation theory.

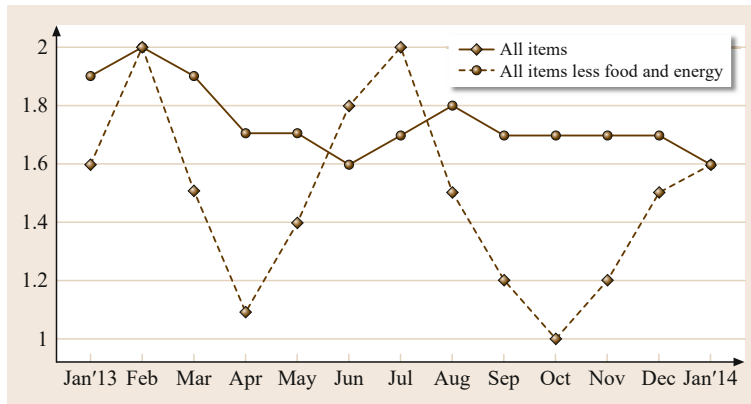
With regard to the first factor, over-indebtedness, Fisher was deeply convinced that on the eve of the stock market crash many firms and households were heavily indebted, as well as the government itself. One of the factors behind such enormous amount of debt was certainly the war.

Between 1929 and 1932, mass debt-liquidation took place, decreasing all American debts by 23% – except public debt, which increased. Deposit banks lost 21% of its volume and 61% of its velocity. The commodity price index decreased, losing 38%; industrial stocks lost 77%. In regard to net worth performances, they are best indicated by the number of firms' bankruptcies, around 29 000 in 1931; net profits of 163 industrial and noncorporations became losses. Production, employment, and trade all kept falling. However, with reference to the United States, it is also necessary to consider other aspects. *Booms and Depression* accurately reconstructed the dramatic events of 1929–1932 that had thrown the world into depression: in particular, the link between bank failures in Austria and Germany (United States was the main lender of both countries), the ensuing crisis of pound (motivated by the huge amount of short-term credit granted to these countries

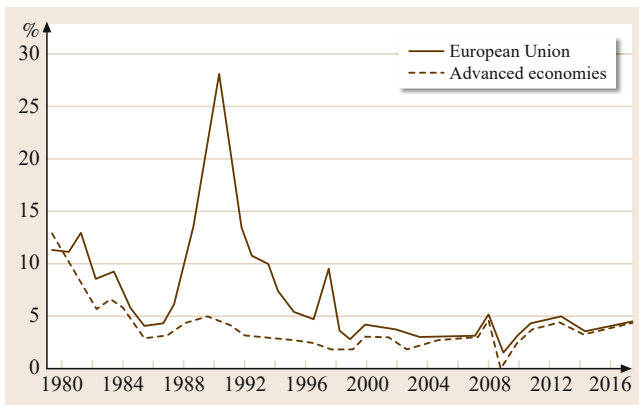
by the British monetary authorities) and the simultaneous speculative attack on American gold reserves. This in particular, as Fisher wrote, demoralized the US banking community, spread hoarding and prompted bank runs.

### 51.4.3 How Debt-Deflation Model Fits Current Economic Conditions

Although inflation has been the norm during the latter half of the twentieth century, there were long periods in the history of the United States during which prices actually fell. In 1836, the money supply contracted by at least 30% pushing prices down. Between 1875 and 1896, prices fell in the United States by 1.7% per year. Between 1930 and 1933, prices fell almost 10% per year. In 2008–2009, the United States experienced the first deflation since the 1950s. In 2009, the Fed has *doubled* the money supply in less than a year and lowered interest rates to almost zero in order to fight deflation. Deflation's threats are several, in particular the possibility for nominal interest rates to reach very low levels thus increasing real interest rates. Moreover, if nominal rates reach the zero level, central bank monetary policies will be ineffective and economy could be stuck into the so-called *liquidity trap*. The liquidity trap is defined as a situation in which the short-term nominal interest rate is zero. In this case, many argue, increasing money in circulation has no effect on either output or prices. The liquidity trap was originally a Keynesian idea and contrasts with the *quantity theory of money*, that expects prices and output to be, roughly speaking, proportional to money supply. According to Keynesian theory, money supply has its effects on prices and output through the nominal interest rate. Increasing money supply reduces the interest rate through a money demand equation, while lower interest rates stimulate output and spending. The short-term nominal interest rate, however, cannot be less than zero, based on a basic arbitrage argument: no one will lend 100 dollars unless he/she gets at least 100 dollars back. This is often referred to as the *zero bound* on the short-term nominal interest rate. Hence, the Keynesian argument goes, once the money supply has been increased to a level where the short-term interest rate is zero, there will be no further effect on either output or prices, no matter by how much the money supply has been increased. The ideas which underlie the liquidity trap were conceived during the Great Depression, when the short-term nominal interest rate were close to zero. At the beginning of 1933, for example, the short-term nominal interest rate in the United States – as measured by three-month Treasuries – was only 0.05%. Sixty decades later another example is that of Japan where



**Fig. 51.4** US 12-month percent change in CPI for All Urban Consumers (CPI-U), not seasonally adjusted, Jan 2013–Jan 2014. Source: Bureau of Labor Statistics. CPI Detailed Report Data for January 2014 Blank Cells = Data not available because it has not been released by the Bureau of Labor Statistics



**Fig. 51.5** Inflation (%) – EU and Advanced Economies. Source: International Monetary Fund, April 2011 World Economic Outlook

the short-term nominal interest rate collapsed to zero in the second half of the 1990s: the Bank of Japan more than doubled the monetary base through traditional and nontraditional measures to increase prices and stimulate demand.

Once the economy moves toward a deflationary situation, it shifts from a relatively short-lived recession to a much more serious and persistent depression. The perspective of future lower prices damages investment by decreasing them, lowers demand, and raises unemployment immediately causing a mass transfer of wealth from debtors to creditors. Usually, such wealth redistribution has no first-order impact on the economy. However, in the face of large shocks, deflation in the prices of assets leads to a decline in the nominal value of assets on banks' balance sheets. For a given value of banks' liabilities, also denominated in nominal terms, this deterioration in banks' assets threatens insolvency. As banks reallocate assets away from loans to safer government securities, some borrowers, particularly small ones, are unable to obtain funds, often at any price. Furthermore, if bank portfolio reallocation is long-lived,

the shortage of credit for these borrowers helps explain the persistence of the downturn. As the disappearance of bank financing forces lower expenditure plans, aggregate demand declines and contributes again to the downward deflationary spiral.

In order to apply the debt deflation theory for an understanding of twentieth and twenty-first century's financial crises, we must first expand it to an *open economy setting*. The international contest characterizing our age is different from 1930s' one when capital movements were small; globalization patterns are the crucial feature of these era where financial deregulation (the so-called Washington consensus) had led to high levels of indebtedness, especially in the South East Asia. Risky-kind of debts, allowed throughout complex financial transactions related to securitized debt obligations as well as to derivatives financial contracts, and in general the boom of financial innovations (leading to decreases in capital goods prices) are typical causes propelling units to further indebted themselves. The financial system results in a complex structure bearing more and more risk: when over-indebtedness is reached, a pro-cyclical boom spiral begins. As a consequence of excessive *financialization* and *deregulation*, the Washington consensus as a matter of fact increased the possibility of a rising endogenous excessive indebtedness, thus causing (as in Fisher's theory) debt deflationary processes [51.25], asset devaluation, and recession.

It could seem that the current financial crisis, since 2007, is evidence against Fisher's theory since it did not imply a significant deflation in consumer prices. The US consumer price index fell 2.1% in the year ending in July 2009 (Fig. 51.4), but it was an isolated event, caused substantially by a sudden and catastrophic drop in commodity prices. Still, the rate of annual inflation measured by the consumer price index did fall from the 3.4% average for the four years between September 2004, until the Lehman crisis in September 2008

down to 0.7% for the two years from October 2008 to October 2010. If people expected the 2004–2008 rate of inflation to continue, then the decline of nearly three percentage points in inflation for last two years has already, as of this date, magnified the real value of debts by over 5%. Moreover, if the low inflation is widely expected to continue for a while, it will lower long-term bond yields, amplifying the real value of debts relative to precrisis expectations by even more. Declining long-term bond yields push up the real value of long-term debts immediately, even more than the immediate decline in prices affects the real value of money today. The effects on balance sheets are of course disruptive to business, to confidence, and thus to the aggregate economy.

From Figs. 51.5–51.7, we can observe the inflation trends in Japan, United States, and the European Union (this latter compared in the same figure with those of Advance Economies) starting from 1980 to 2016 (data contain projections). In all cases, the inflation rate calculated using the Current Consumer Price Index shows its declining performance, reaching the lowest level in 2008 corresponding to the financial crisis.

That is probably enough to give substantial support to Fisher’s debt-deflation theory in the current era. It is particularly notable that in the recent crisis there has been a mortgage refinancing boom. The otherwise large effect of inflation was, in a sense, reduced because many mortgage holders exercised the call option to move to a lower interest-rate mortgage. Anyway relying on call provisions clearly is not the key to solve instability of purchasing power. Because they are one-sided, that is, they protect borrowers against the risk of declining inflation, but do not protect lenders against

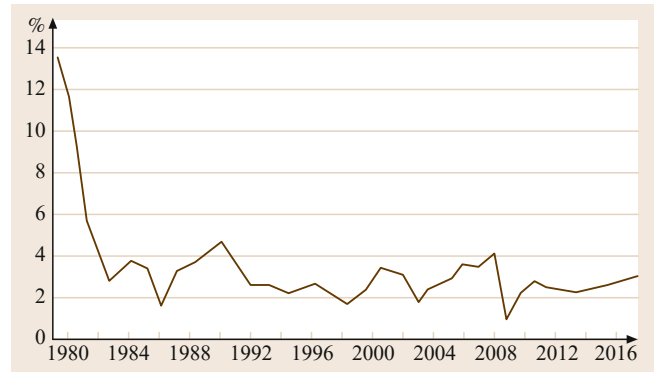


Fig. 51.6 Inflation (%) – US Source: International Monetary Fund, April 2011 World Economic Outlook

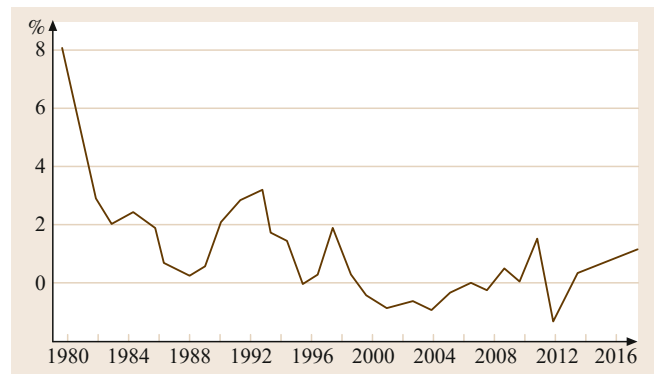


Fig. 51.7 Inflation (%) – Japan. Source: International Monetary Fund, April 2011 World Economic Outlook

the risk of increasing inflation, they have the effect of increasing the cost of borrowing.

## 51.5 Policy Implications and the Shareholder Maximization Value Model

### 51.5.1 Stability is Destabilizing

The central idea in Minsky’s financial instability hypothesis is that a solid economic system can transform itself into a fragile and unstable one due to endogenous changes in cash flows [51.13, p. 12]:

“The financial instability hypothesis states that tranquil growth and prosperity in modern capitalist economies naturally bring about changes in cash-flow interconnections leading from solidity to fragility, and that the normal functioning of the economy may easily convert fragility into open financial crisis.”

Minsky’s financial instability hypothesis incorporates the concept that *stability is destabilizing*: it is during tranquil growth periods that conditions for an economic crisis are laid down. When the economy is basically constituted of hedge financing positions and there are good safety margins about future positive profits, a financial crisis can happen only because of an unexpected drop in incomes. At the contrary, an economic system increasingly based on speculative units is going to be *intrinsically* much exposed to financial crises since even small financial markets fluctuations will have great rebounds on real variables. In particular, an increase in interest rates can cause:



- The impossibility of refinancing for speculative units that must sell their assets
- A drop in prices, income, and expected and realized cash flows of each unit
- An overall increase in the level of indebtedness.

We can start by analyzing the process through which an economic system switches from stable to unstable by imagining a financially robust economy, where almost every agent is in a hedge-financing position. The economy experienced cycles (including recession phases) in the past, but is now in tranquil prosperity. Agents remember past crises: they maintain, therefore, ample margins of safety in their liability structures, and balance sheets are awash with cash and very liquid assets. As long as the economic development goes on without shocks, optimistic forecasts are validated. Banks and firms expect large positive future quasi-rents, and bet on an upward trend in the value of real and financial assets. As growth goes on, optimism propagates and firms have the opportunity to rise their debt–equity ratio associated both with higher short-term financing of fixed capital and long-term financial assets. Banks, that are profit seeking units, have interest in allowing firms’ requests for funds, thus they also start using innovative financial instruments to increase credit availability. The improvement of expectations, the reduction for liquidity preference and of safety margins, and the expansion of quasi-money (financial innovations) cause an increase in capital assets demand and prices. Such a context is the midst of an investment boom which, in turn, drives production and sustains profits (note that these latter are an important factor for lenders’ decision making). Optimism propagates and indebtedness increases again. The presence of low interest rates make financial leveraging affordable, and raises the conditions of fragility of the whole system. Nonetheless until euphoric expectations are realized, there is no way to stop financial fragility growth. With the emergence of a state of *euphoria*, speculation has two main aspects: (1) owners of capital assets speculate by financing with external funds ownership of investment goods and securities; (2) families and firms speculate on financial activities they own and on the way they can finance ownership of such activities [51.12]. In particular, speculative positions of families, firms, and banks are characterized by a high ratio between financial commitments and income. The intrinsic presence of a temporal shift in the financing process becomes crucial when units understand that they overvalued their expectations and that they now have to face a liquidity problem linked to high debt payment commitments. Units are thus forced to refinance. This process sooner or later will make interest rates to grow (banks are

indebted as well), reducing the actual value of investments: it translates into a reduction of safety margins. We shall recall that banks have long-term activities, due to the financing of investments, while they have short-term liabilities. With an increasing number of loans during the boom phase, this temporal structure can determine a condition where cash COF are not covered by CIF in each period. Banks are forced to increase interest rates in order to face increased liabilities. This is the upper turning point of the cycle. In order to avoid bankruptcy, indebted units sell their activities at declining prices (triggering debt-deflation), with a debt revaluation in real terms, and a loss of liquidity. It follows a drop of investments, demand, and profits, that are dragged down by pessimistic expectations.

In the course of an economic crisis, speculative activities that were undertaken during the boom period are the first to collapse; moreover, these will also have rebounds on hedge activities because of the so-called *interconnections of budgets*, that is, our financial system is a stratified one, where one unit’s debts reflect into another’s credits. The number of bankruptcies – households’ insolvency included – increases, and assets price deflation prolongs as far as Ponzi and the most of speculative units exit the market. In this phase, liquidity preference reaches its maximum level, investments activity almost stops; income, profits, and assets value drop, while unemployment raises. The only positive aspect of a financial crisis is that, according to Minsky, it lowers the financial fragility of the system, which is a necessary endogenous condition for the following economic recovery.

To sum up, Minsky argues that our economy is unstable because of capitalist finance. In particular, if a mixture of hedge, speculative and Ponzi positions, and of internal and external financing can rule innocuously for a while, but, at some stage – encouraged by a period of economic boom and widespread euphoria – they will breed endogenous incentives for units to change the mix, shifting toward riskier financial positions and more external financing. It follows an investment boom phase where debt–equity ratios raise and margins of safety erode. The severity of the subsequent financial crisis (and recession scenario) will depend on the relative size of government intervention and on the breadth of the central bank’s *lender-of-last-resort* action.

### 51.5.2 From the Debt Deflation Model to Policy Proposals

*Fisher* argued that one of the causes of the collapse of US economy in 1929 (and subsequent depression that lasted until 1933) was the choice by the Federal Reserve to abandon the stabilization policy that had been

pursued during the twenties by Benjamin Strong, Governor of the Federal Reserve, who died in 1928. *Fisher* was convinced that it is always economically possible to stop or prevent long depressions simply through *reflation* [51.26], in other words, pursuing a monetary expansion to bring prices back up to their pre-29 levels at which debts were contracted. An essential point in Fisher's plan was the belief that pursuing an increase in demand, even if necessary, was inevitably only a first step. To get out of a depression, a substantial rise in prices was needed. A second point is the transmission mechanism which explains how an expansionary monetary policy can induce an increase in output and prices. An increase in the means of payment available to agents produces an increase in aggregate demand, which in turn boosts prices. As already mentioned, during a serious depression, agents tend to hoard their own liquid assets, thus it may be argued that a greater supply of liquidity could have no real effect. But Fisher was aware of this point. In the second half of 1932, the worst period of the Depression, he became a supporter of a plan for *stamped money*. Stamped money was a promise of payment issued by a public body or municipality, usually with the guarantee of a bank, which circulated as a banknote but that, within a given period of time, could be taken out of circulation and converted into legal tender. Its peculiarity was that it was subject to a periodic *tax* (for example two cents per dollar per week) in the form of a stamp that had to be affixed on the back of the note. This made the plan a self-liquidating operation at no cost to the finances of the authority that issued the notes. The notes were obviously characterized by an high velocity of circulation, as every agent had an evident interest in spending the scrip quickly so as to avoid the tax [51.22].

### 51.5.3 Financialization, Neoliberalization, and the 2008 Crisis

The so-called *subprime crisis* has been the peak of a decade of a housing boom in the US economy, stimulated by a number of regulatory changes of the financial system which gradually moved it toward conditions of fragility. The subprime crisis can be thus interpreted as a so-called *Minsky moment* which soon transformed into a so-called *Minsky meltdown*; excessive financialization and deregulation led to over-indebtedness, debt-deflation, asset devaluation, and recession. A *Minsky meltdown* is when financial instability becomes so acute that only an immediate, global, and massive government intervention can avoid a systemic banking failure. A *Minsky meltdown* thus starts when the *Minsky moment* is so intense to trigger a recession.

The world that emerged after the Great Depression was characterized by low private debts, high govern-

ment debt (war finance), and a lean financial system. Such features boosted rapid economic growth, corporations mostly used retained profits to finance expenditures, finance was kept small, regulated, and relatively irrelevant, risky financial practices were outside commercial banking. Over time a process of *neoliberalization* (see next section) changed all those regulations which were relaxed or defeated by financial institution through innovations. Private debt grew, and risky practices emerged. The weight of finance moved away from institutions toward markets – the so-called *originate to distribute* model. Weakened labor through part-time practices and lower remuneration was emerging (often under the label of labor *flexibilization*. *Money manager capitalism* and practices such as securitization were born and spread over.

#### Neoliberalization

In the 1980s, the President of the United States, Ronald Reagan and the Prime Minister of the United Kingdom Margaret Thatcher were pushing the free-market doctrine across the world to make privatization and deregulation policies politically viable out of their countries. Deregulation of the banking system spread from the United States to Europe to many developed countries and to the emerging economies of the Asian *tigers*, Russia, and Thailand.

Neoliberalism started as an intellectual movement in the 1980s and then transformed in the so-called *Washington Consensus* in the 1990s. The ideological roots of neoliberalism can be traced to the liberal theory of self-regulating markets developed during the belle époque and later, thanks to the contributions of Friedman and Hayek. The term *Washington Consensus* is used to describe the policy prescriptions given by the economist John Williamson as a baseline of directions for nations in need of assistance from international economic entities such as the World Bank and the International Monetary Fund. The *Washington Consensus* was originally laid out in 1989, and refers to all the most evident neoliberal policies: free trade, flexible labor, active individualism, anti-statism, and the cutting of institutional welfare policies, in particular after-war Keynesian ones. Neoliberalism is thus a set of political beliefs which firstly include the belief that the only legitimate intervention of the state in economic life is that of safeguarding individual, especially commercial, freedom as well as private property rights.

During the Thatcher and Reagan governments, the IMF was supporting worldwide liberalization of capital movements; in 1997 a global agreement on financial services was made with the supervision of the World Trade Organization [51.27].

Through a process of *neoliberalization* US banks were experiencing deregulation and were increasingly investing in foreign-exchange trading, expanding geographically as well as widening their range of services. With the concept of neoliberalization, we refer to the processes of regulatory restructuring under post-1970s and post-2008 capitalism. As argued by *Peck et al.* [51.28], such a process implies: commoditization, movement of huge amounts of capital and speculative (innovative) financial tools throughout the world in search of profit opportunities, privatization, deregulation, and trade liberalization.

Since the 1970s, a new kind of global economy has started to shape. Restrictions on capital have been progressively reduced, free currency exchange, free world trade, and free capital circulation were intensifying. At the same time, extensive privatization was taking place, while a process of weakening of the negotiation power of labor, and its flexibilization through part-time and lower remuneration was emerging. Such a set of policies, summed up with an aggressive dismantling of the welfare state, has produced a declining share in both real wages and investment, and thus slower growth of effective demand. In particular with respect to the US case, Reagan increased government military expenditures, which resulted the only demand component with a positive tendency and the public deficit was worsened by tax reductions.

In such a sociopolitical context, finance appeared to be increasingly valuable. It is however only during the 1990s that a complete picture of the new economy was clear; labor is progressively subdued to finance, consumers are increasingly over-indebted and class power is gradually restored. Federal Reserve's Chairman Greenspan was the promoter of a number of financial deregulatory measures. More importantly his monetary policy goals have shown to be significantly different from those of the post-WWII era, which were focused on full employment and rising real wages. Greenspan's policies gave prominence to fight inflation rather than unemployment, thus rising wages were brought down since considered inflationary; but as *Palley* [51.29] argues, the same logic was not applied to raising profits (in accordance with the Chicago School theories). Under Reagan the conjunction of restrictive monetary policies and expansive fiscal policies, the differential in the level of interest rates in favor of the United States led to capital inflows and to a revaluation of the US dollar.

### Money Manager Capitalism

*Minsky's* work on structural economic change [51.16] and on the financial instability hypothesis is not only useful for explaining the economic changes of the pe-

riod between 1945 and 1966, but it catches a broader historical analysis to help interpret the subsequent developments of the capitalistic system and the evolution of the financial structure. *Minsky* [51.30] argues that capitalist economies undergo various stages of development. During expansionary periods, financial innovations prosper, in turn relaxing financing constraints, and leading to increased investments. This implies an increasing reliance on outside sources of funding and debt leveraging that, in turn, boosts investment, aggregate demand, and profits. *Minsky* believes that capitalist economies which own such a structure – based on reinforcing expectations and on the possibility of increasing financial innovations – are intrinsically prone to boom and bust cycles. Indeed, when an endogenous shock causes a shift in entrepreneurial prospects, investment will drop bringing down aggregate demand and profit too. A wave of risk aversion and desire of liquidate debts leads to bankruptcies and debt-deflation progresses – unless the Big Government and the Big Bank intervene.

The evolution of the US financial industry in the post-WWII period shows the emergence of various innovative instruments and institutions. Especially during the inflationary period of the 1970s, the differential between securities and mutual funds yields and those of insured deposit institutions was increasing, thus leading to new profit opportunities and the birth of money market funds. *Minsky* [51.31, p. 12] defined the five stages of capitalistic development in the United States,

“(i) commercial capitalism, (ii) industrial capitalism and wild-cat financing, (iii) financial capitalism and state financing, (iv) paternalistic, managerial and welfare state capitalism, and (v) money manager capitalism.”

We focus on the transformation from what *Minsky* calls managerial capitalism – shaped by the experiences of the Great Depression and World War II – to the current system of managed-money capitalism (which he believed had emerged by the early 1980s).

During the period of managerial capitalism (the post-war period) the financial structure was conservative, with low debt levels and contained speculation; firms owned most of their net positions in governmental debt. Macroeconomic conditions were generally stable and the accumulation of wealth was possible, as well as distributed equally. But it is during such conditions that money manager capitalism is born, especially because of the shared confidence in profitable future. During the 1970s inflation, financial and technological innovation led funds to flow from bank deposit accounts to mutual funds and securities. Money manager capital-

ism is defined as one dominated by highly leveraged funds (in particular pension and mutual funds) searching for the maximum return within an environment made of little regulation or supervision of financial institutions and underestimation of risk. Money managers mix nontransparent and intricate instruments that have the quality to quickly spread throughout the world. Market stimulates managers to bear high risk by giving them greater remunerations.

Managed-money capitalism is part of the trend toward an increase in the proportion of financing that takes place through markets rather than intermediaries [51.30, p. 70].

Securitization of home mortgages began in the early 1980s. As Minsky saw it emerging, he understood the risks it would imply, and in 1987 he wrote a Policy Note telling that such an instrument was the result of the increased importance of money manager capitalism (the market) and the decline of banks (commercial banks).

#### Financialization: A Finance-Led Economy

From the mid-1980s, the growth of financial sector indebtedness has been huge: from one-fifth of GDP to 120% of GDP [51.32], especially because of the growing securitization process and the issue of increasingly volatile products into portfolios to fund positions in securities. As we mentioned in the previous paragraph, the relative stability of the post-war period led to the increase of financial innovations, easy credit availability, higher competition, and leverage ratios. Thanks to such easy credit, asset prices started rising, prompting other financial innovations and raising leverage ratios again. This in turn caused loans to expand, especially those for home buying, thus raising real estate values, expanding loans even further and increasing leverage ratios in an attempt to cover the raised value of real estate. This structure can be considered a Minskyan's Ponzi position.

The wave of financialization after the 1980s has been huge. With the term financialization [51.29], we refer to the process through which financial markets, financial institutions, and financial actors (élites) acquire importance and are able to seriously influence economic policies as well as economic outcomes. The relationship between financialization and neoliberalism is strong: neoliberalism and its counterpart in globalization are heavily sustained by an extraordinary expansion and promotion of financial activity. Those who support financial deregulation improvement have argued that financialization provides superior risk-management [51.33, p. 6]:

“for example, securitization was supposed to slice risk into different parts (by means of different se-

curities) and allocate it to those who were best equipped to hold it”

in a sense thus increasing the stability of the entire system. At the opposite, it could be argued that since the beginning of the financialization era, a series of major crises (such as the debt crisis in 1982, the Savings & Loans crisis in the USA in the 1980s, the Peso crisis in 1994, the 1998 Asian Crisis, the Dot.com bubble in 2000, and lastly the current global financial crisis, started in 2007) have showed this assumption to be unsustainable.

During the post-1980s era, financial markets greatly expanded and put increased pressure on nonfinancial corporations (NFCs) to generate increasing earnings and distribute them to financial markets. Firms that failed to meet financial markets' expectations faced fall in stock prices and threats of hostile takeovers. Such environment produced what Crotty calls the *neoliberal paradox* [51.34]. Unable to increase their profits due to adverse conditions in the product markets, firms were forced to pay an increasing share of their internal funds to financial markets.

Financialization gave birth to two new models of growth. The Anglosaxon growth model, based on consumption (consumption-led growth model), and the German, Japanese, and Chinese model, based on export (export-led growth model). The two models rely on each other. The Anglosaxon countries developed a credit-financed consumption as the most dynamic component of demand growth and a current account deficit and capital inflows, while in the other group of countries consumption remained low, but demand growth was fuelled by very high exports. The US, the UK, Ireland, and Greece ran large current account deficits, whereas Germany, Japan, and China had large surpluses. International imbalances have played a major role in the debate on the causes of the 2008/2009 crisis. As argued by *Stockhammer* [51.33], financial globalization and liberalization led to exchange rate crises. Exchange rates changes are determined by capital flows an amount of capital inflows can reverse because of a sudden shock and thus rapid withdrawal. This can lead to exchange rate crises which are furthermore made possible because of the widespread practice of the currency-carry trade in international markets. By implication, assets and liabilities will then be denominated in different currencies; thus abrupt exchange rate realignments may have disastrous effects on firms' or banks' balance sheets. Exchange rate crises have been a common feature of emerging economies, such as for Mexico in 1994, and in the case of South East Asian crisis of 1997–1998 and Argentina in 2001. Exchange rates crises all have led to long-lasting recessions.

Second, capital flows liberalization has allowed some countries to run current account deficits (much more than during the Bretton Woods agreement period). Because the balance of payment is composed of the current account and the capital account, a deficit in the current account can be balanced by net capital inflows allowing countries to run current account deficits and balancing them by attracting capital inflows.

Financialization has profound effects on income distribution. The dominance of capital against labor has been one of the results of the neoliberal break-up of labor stability. For instance, in the US, Europe, and Japan, wages are stagnating since the 1980s. Wages stagnation has been spurred by increasing income inequality, the rise of the so-called *rentiers income* and the growth in the financial sector especially in the form of bonuses.

With respect to the economic environment we have just described, the mainstream corporate governance doctrine spreading from the Anglo Saxon countries to Europe and to the Emerging Markets has been the *shareholders value maximization theory*. This conception argues that since the firm belongs to stockholders, managers should run it in order to maximize shareholders value. Such a process leads to increases in dividend disbursements and share buybacks. At the same time management remunerations are increasingly linked to profit and stock market value, which in turn increases managers' incentives to keep stock prices on a high level. It has been argued [51.35] that one of the main causes of US growing inequity is the stock-based compensation of the executives. The key management change has been from a *retain and invest* strategy to a *downsize and distribute* one (on organizational restructuring see [51.36]). Shareholder maximization value practices were born in the United States in the 1980s, then spread to UK and to continental Europe as well. In the 1990s, both Germany and Japan leading executives were calling for the adoption of more American-style corporate practices, in order to compete in increasingly globalized capital markets. Convergence hypothesis theories highlight the role of global capital flows in eliminating inefficient forms of governance, since the early 2000s OECD and World Bank were busting the adoption of common standards.

Neoclassical shareholder theory defines shareholders as the only *residual claimants* of a firm, they bear

the risk of business success/failure and the return/loss they get is what has left after other stakeholders have been paid. Furthermore, they are the only stakeholders which have an incentive in investing productive resources [51.37]. The problem with the concept just described is that it does not consider workers as residual claimants as well. *Freeman and Evan* [51.38] and *Blair* [51.39] argue that workers do own like shareholders the status of residual claimants because they invest in firm-specific human capital with the expectation of having returns from such investment during their future career. They do bear risk as well, because they would suffer in case of lack of inter-firm labor mobility. This vision is much more related to the *stakeholder theory* [51.40]: managers should take decisions so as to take account of all stakeholders of a firm (not only financial claimants, but also employees, customer, communities, etc.). *Evan and Freeman* [51.41, pp. 102–103], asserted that “management has a duty of safeguarding the welfare of the abstract entity that is the corporation” and of balancing the conflicting claims of multiple stakeholders to achieve this goal. They further argued [51.41, pp. 102–103]:

“A stakeholder theory of the firm must redefine the purpose of the firm. [...] The very purpose of the firm is, in our view, to serve as a vehicle for coordinating stakeholder interests.”

The shareholder value orientation leads to increases in dividend disbursements and share buybacks, since management remunerations are increasingly linked to profit and stock market value which increases managers' incentives to keep stock prices on a high level.

With respect to the role of the state in the economy, from the 1980s onward, it was a priority of neo-liberals to downsize its presence in the economy. *Stockhammer* [51.33] shows that the state share, as measured by the size of state employment and transfers, has not been reduced in most OECD countries (except for Ireland, United Kingdom, and Netherlands); at least there has not been a growth in the share of state intervention, but the level has remained as its 1970s level. Nonetheless privatizations and deregulations have in practice reduced the influence of the state into the economy.

## 51.6 Integrating the Minskyian Model with New Marxists and Social Structure of Accumulation (SSA) Theories

Since the summer of 2007 to the spring of 2008 the Global Financial Crisis has been spreading throughout the world. Furthermore, since late 2009 fears of sovereign debt crises in the euro zone including Greece, Ireland, Spain, and Portugal arose. In the European Union, especially in countries where sovereign debts have increased sharply due to bank bailouts, bond yield spreads widened and risk insurance on credit default swaps between these countries and other EU members, most importantly Germany, grew.

I hereby argue that Minsky's financial instability hypothesis needs to be combined with the sociopolitical scenario (i. e., a *neoliberalized* and *financialized* economy) described in the earlier sections. I indeed agree with the structural Keynesian synthesis [51.29, 42] of the new Marxists view [51.43] and of the social structure of accumulation SSA theory [51.44]. In interpreting the crisis, a common aspect stressed by these theories is the acknowledgment on the critical role played by the neoliberal model and its adoption by many advanced economies in the 1980s. These approaches indeed believe that the crisis has its roots in the real economy; this could seem at first in contrast with the pure Minskyan explanation of crises as endogenous financial system processes of excessive optimism and indebtedness. Nonetheless it will help us to generate a more extensive comprehension of the subject of our work.

Both New Marxists and SSA theorists adopt an *under-consumptionist* position according to which the economy cannot reach full employment because of the lack of demand caused by an excessive wage squeeze. The new Marxists use a monopoly capital mode of analysis to explain the crisis, arguing that it is the result of the three decades of wage stagnation and widening income inequality caused by the contradictions inherent in the neoliberal model of growth. In this perspective Karl Marx was the first to argue that capitalism produces recurring and increasingly dangerous crises because of its inherent feature to direct large amounts of wealth toward a few pockets of wealthy elites, leaving a growing number of people without the possibility of affording to buy goods and services produced by firms. It follows the emergence of an *industrial reserve army* of poor and unemployed, and a contracting economy. The new Marxist (e.g., [51.45]) theory – as well as that of Marx, Minsky, and Fisher – stresses the role of accumulation of private debt as a means of sustaining the cycle. However, the new Marxist framework does not incorporate the mechanisms of Minsky's financial in-

stability hypothesis, even if the role of debt as a means of sustaining the cycle is a common feature.

The SSA theory focuses on the problem of neoliberal wage-squeeze and exploitation, interpreting stagnation as a purely real phenomenon caused by a lack of aggregate demand caused by a worsened income distribution. Such an approach does not by the way incorporate the financial instability hypothesis as well.

On the other hand the third approach, the structural Keynesianism, developed by *Palley* [51.29, 42], offers a synthesis of the two points of view. First, it shares the Marxist perspective that there is a real economy problem regarding a wage squeeze and an unequal income distribution, which ultimately gives rise to a Keynesian lack of aggregate demand. Second, structural Keynesianism recognizes that finance plays a critical role in fuelling asset price bubbles and over-indebtedness which sustains the lack of demand caused by the wage squeeze. Then it adds two other remarks:

- By assuming that financial innovation and deregulation are the sources to fuel demand, it is possible to incorporate Minsky's financial instability hypothesis into the structural Keynesian approach.
- It takes into consideration the US model of global economic engagement in causing the crisis, that spreads throughout the world by means of three channels: leakage of spending on imports, off-shoring of jobs, and off-shoring of new investment.

As mentioned, the processes identified in Minsky's financial instability hypothesis play a critical role in the actual crisis, but we must consider a larger economic scenario involving the neoliberal growth model implemented around the 1980s. Thus integrating the structural Keynesian approach into the Minskyan one, financial markets have been the means through which the neoliberal model of growth could regenerate and sustain demand escaping its *stagnationist* tendency. This explains why the crisis took the form of a financial crisis; financial markets have been *the place where to sustain and generate demand* in order to counteract the wage squeeze. Economic policy deteriorated the workers position and neoliberalism used financial innovation to sustain increasing lacks of demand.

Minsky's financial instability hypothesis explains how the neoliberal *growth* model avoided stagnation for so long and could reproduce through over-indebtedness and leverage (produced in financial markets) instead of wage growth, leading to the crash identified by Minsky.

With respect to structural Keynesianism, new Marxists and SSA conclusions about the *cure* for the crisis, they show differences only in terms of degree of optimism. They all endorse the belief that monetary and fiscal policies are not sufficient to restore full employment: they identify the need to reverse neoliberalism and restore the link between wages and productivity

growth. Structural Keynesians believe it is possible to manage appropriate institutions that, combined with traditional Keynesian policies, can produce full employment and shared prosperity. New Marxists and SSA theorists are more pessimistic and they see an institutional design with a larger public sector and more nationalization, especially regarding the financial sector.

## 51.7 Risk and Uncertainty

In the previous section, we have been dealing with a few macroeconomics theories of investment and financial markets. In the theoretical frameworks described, the concepts of risk and uncertainty have from time to time emerged. We have for instance mentioned the issue of lender's and borrower's risks in the Minskyian model for investments, as well as the role of risk and uncertainty in investment decisions (e.g., the determination of the (real) investment level in the presence of risks). As mentioned, according to Keynes, investors' speculative decisions about what assets to retain and what to sell is based on a mechanism through which human beings use *practical equivalents* as a convention which is retained until it confirms reality, but abandoned when evidence becomes different and in contrast with it.

This example about agents modelling for investment in the presence of uncertainty, and the use of practical equivalents as a mechanism to make decisions, is one of the various models advanced in the economic theory – as well as in business studies – to introduce uncertainty in economics agents decision-making processes.

### 51.7.1 Models of Risk and Uncertainty in Economics and Business Studies

The topic of uncertainty is prominent in the twentieth century's economic discourse, starting with the marginalist revolution of Carl Menger, the Austrian School, and then thanks to the seminal contributions by John Maynard Keynes and Frank Hyneman Knight. Standard neo-classical economics assumes that actors have complete knowledge of means-end relationships and they maximize their utility on the basis of a given set of knowledge, technology, and preference ordering. The notion of complete knowledge allows for the functioning of markets according to the neoclassical model and for the development of Pareto optimal equilibria. Much of twentieth-century economics, especially General Equilibrium Analysis, deals with the mathematical formulation of the functioning of the economic system under conditions of *perfect knowledge*, and *perfect rationality*. The notion of *incomplete knowledge* can

be traced back to the *marginalist revolution*, with Carl Menger, and the Austrian School of Economics.

The Austrian School looks at the human limits in cognitive capacity as the crucial source of uncertainty in the production process [51.46]. Furthermore, it sees uncertainty in terms of limited knowledge of future outcomes, and uses such an assumption to assert that rational state-planned production activities are not conceivable. Only recently, attention has been devoted to the analysis of imperfect markets, as a result of incomplete information [51.47].

One of the most important contributions to the conceptualization of uncertainty in Economics is the seminal work by Knight titled *Risk, Uncertainty, and Profit* [51.48]. The author argues that agents are not able to produce optimal forecasts of all future states; therefore, in a dynamic economy, they cannot make decisions leading to equilibria outcomes. According to Knight, disequilibria are caused by uncertainty because omniscience is not possible. Uncertainty undermines perfect information and therefore the ability to predict the *distribution probability of future instances* (i. e., it is difficult for the decision maker to predict the probability whereby future outcomes will arise). In the Knightian acceptance of uncertainty, not only it is difficult to estimate the frequency distribution of future possible outcomes, but often it is even hard to classify events themselves. Neither they can be grouped on the basis of features of similarity, nor it is helpful to make reference to the past (e.g., a series of historical data) because some events are, just, not knowable. In all these cases, decision makers face a situation pervaded by *uncertainty*. Instead, in those cases when (i) either the distribution of the outcomes can be drawn in advance or (ii) it can be induced by looking at historical data (using this criterion to group instances) – decision-makers face a situation of risk (Sect. 51.7.1).

*Risk, Uncertainty, and Profit* has been the object of several interpretations by the literature. For instance, LeRoy and Singell [51.49] have interpreted Knightian uncertainty in light of the *agency theory* [51.50], stating that in some situations insurance markets can operate

thanks to the possibility to calculate and reduce *risk*, while in other situations they would collapse because of moral hazard and adverse selection (i. e., because of conditions of uncertainty). Also, *Langlois* and *Cosgel* [51.51, p. 457] interpret Knight's characterization of the firm as

“the system under which the confident and venture-some ‘assume the risk’ or ‘insure’ the doubtful and timid by guaranteeing the latter a specified income in return for an assignment of the actual results”

stressing the different risk aversion faced by the entrepreneur and the worker.

The entrepreneurship literature has argued that business decisions often deal with unique situations where objective probabilities or chance are immeasurable, and that entrepreneurial action is affected by uncertainty and ultimately involves strategies to cope with it [51.52, 53]. According to classical theories of entrepreneurship (e.g., [51.48, 52, 54]), entrepreneurial behavior is the fundamental engine that influences the environment through innovation. Schumpeter argues that the willingness to bear uncertainty is intrinsic to entrepreneurial activity and is given by the innovative act of creating new combinations. According to the Schumpeterian view, the role of entrepreneurial action in coping with uncertainty determines the success of the firm.

A variety of definitions of the concept of uncertainty exist in business studies literature. Uncertainty in Organization Studies is seen as lack of information for, and knowledge in decision making [51.55, 56]. It is also postulated as resulting from the indistinct and convoluted causal configuration underlying the internal operations of the firm, its environment, and the complex relationship between the firm and the environment [51.57]. Uncertainty is viewed as a product of unpredictability [51.58], environmental turbulence [51.59], and the complexity of influential variables [51.60]. Further, uncertainty is also a tangible facet of the external environment, and at the same time a perceptual attribute internal to managerial decision making [51.61].

In the Organization Studies, literature uncertainty [51.55, 56, 62, 63] is often found under the label of *environmental uncertainty* or *perceived environmental uncertainty*, the former referring to variation of conditions in the organizational environment, the latter to the ability of management in predicting such variation. According to transaction cost economics (TCE) [51.64], *environmental uncertainty* is one of sources – together with behavioral uncertainty and asset specificity – determining the choice for a firm to perform the transaction using the market (i. e., outsourcing the transaction) or hierarchy (i. e., internalizing the

transaction). Moreover, TCE has modeled behavioral uncertainty as a situation where the greater the asset-specificity of an investment, the greater the threat of opportunistic behaviors. Asset-specificity is a feature of all those investments that have value only within a specific transacting relationship, and lose value outside such transaction. Asset-specificity give rise to what in the literature is known as the *hold-up* problem meaning that, in a bilateral transaction, after one party has made the asset-specific investment, the other may take advantage of such specificity to appropriate of some rents expected to be earned with the investment through opportunistic behavior.

In the Agency Theory framework “uncertainty is viewed in terms of risk/reward trade-offs, not just in terms of inability to preplan” [51.50], while according to the Resource-Based View approach [51.65] a distinction between risk and uncertainty exists [51.65, p. 56]:

“[...] the fact that the future can never be known with accuracy means that the planning of business firms is based on expectations about the future which are held with varying degrees of confidence [...] *Uncertainty* refers to the entrepreneur's confidence in his estimates or expectations; *risk*, on the other hand, refers to the possible outcomes of action, specifically to the loss that might incurred if a given action is taken.”

### Uncertainty and Knowledge

In Economics models of uncertainty are usually incorporated into economic actors' decision-making [51.66–70] and uncertainty is generally defined as a *lack of knowledge* about the state in the future. *Keynes* [51.3] investigated the role of the quantity and quality of information owned by agents in their decision-making process and the role of incomplete information as being the core of uncertainty. The Keynesian conceptualization of uncertainty is not distant from the Knightian one. Indeed, the former believed that some social processes cannot be classified as *ergodic* or deterministic. These are the outcome of decision-making processes and thus behaviors are guided by two factors: the creation of premises through the imagination and the making of choices on the basis of nondeterministic forces like *animal spirits*. Arrow describes uncertainty as follows [51.3, pp. 33–34]:

“Uncertainty means that we do not have a complete description of the world which we fully believe to be true. Instead, we consider the world to be in one or another of a range of states. Each state of the world is a description which is complete for all relevant purposes. Our uncertainty consists in not knowing which state is the true one.”



From a micro perspective, if we look at theories of the internationalization of the firm – especially at those labeled as Process Theories – the impact of knowledge and learning gained through international experience emerges as crucial to the pace and direction of firms’ internationalization activities, especially those taking place after the first stages of international growth. The most famous Process Theory is the Uppsala Model developed by *Johanson* and *Wiedersheim-Paul* [51.72] and *Johanson* and *Vahlne* [51.73]; it theorizes that firms internationalize following a (linear) *learning process*: they first approach closer markets and subsequently more distant ones (distance in this stream of literature not only refers to geographic distance, but also to differences for instance in terms of culture, institutions, etc.). The level of commitment in a market (i. e., the *mode* of investing in that country) changes according to the *learning path*: the first internationalization stages are carried out only by engaging in (unsolicited) export activities, while the last ones entail establishing wholly owned subsidiaries. The Swedish authors advanced two basic assumptions:

- The *lack of knowledge* about foreign markets and operations is an important obstacle to the development of international operations.
- The *necessary knowledge* can be acquired mainly and preferably through operations abroad.

These hypotheses hold for the two directions of internationalization that were distinguished, (a) the increasing involvement of the firm in the individual foreign country – vertical expansion – and (b) the successive establishment of operations in new countries – lateral expansion. The determinant of the time order of such expansions and their geographical disposition is determined by the *psychic* distance between the home and the import/host country. *Psychic distance* is one of the pillars on which the Uppsala model is built and

refers to the “sum of factors preventing the flow of information from and to the market, and is often correlated with geographical distance” [51.72, p. 308]. Given the previous considerations, *Johanson* and *Vahlne* derived that internationalization is the product of a series of incremental decisions in small steps, rather than large foreign production investments at single points in time. In other words, internationalization is intended as the consequence of a process of incremental adjustment to changing conditions of the firm and its environment. Subsequent models about the internationalization of the firm, that is, the Innovation-Related Internationalization Models can be rightly regarded as behaviorally oriented theories and the organization’s internationalization process can be considered as influenced by (i) the lack of knowledge by the firm, especially experiential knowledge; and (ii) uncertainty associated with the decision to internationalize.

Figure 51.8 illustrates *Khaneman* and *Tversky*’s *phenomenology* of uncertainty [51.71]. The primary distinction refers to two loci to which uncertainty can be attributed: the external world and our state of knowledge. The second level distinguishes four prototypical variants of uncertainty that arise depending on the nature of the data that “the judge might consider in evaluating probability” [51.71, p. 152]. External uncertainty can be assessed in two ways: (i) a distributional mode, where the instance in question is part of a class of similar cases, for which the frequency distribution is known, or it can be estimated; or (ii) a singular mode, where probabilities are estimated by the propensity of that particular case to happen. Internal uncertainty can be either (a) reasoned or (b) introspective, but in both cases it reflects (partial) subjective ignorance, rather than disposition of external objects. The statement “Copenhagen is much colder than Milan” reflects a process of “sifting and weighing” [51.71, p. 154], while the statement “I think her name is Anna, but I am not sure” expresses a certain level of *confidence* based on an introspective judgment.

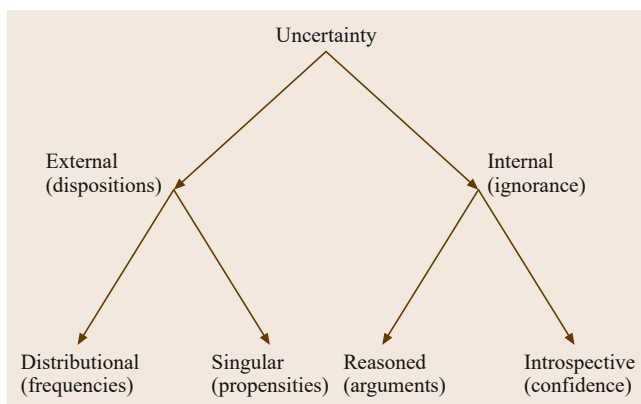


Fig. 51.8 Variants of uncertainty (after [51.71])

### Epistemic Issues About Probability Beyond Models of Risk and Uncertainty

The academic debate born during the 1920s and 1930s around probability and risk can be basically divided into two streams: *objectivists* and *subjectivists*. The formers are scholars arguing that probabilities are real, intrinsic in nature, hence discoverable through logic or statistical estimates. The latter are scholars asserting that probabilities are human beliefs [51.74]. Theories assuming that probability is merely a type (and not an object) of knowledge include the subjectivists, including *Savage* and *Friedman*, and some objectivists, including *Keynes*. Scholars that, on the opposite, suppose that probability exists as a part of external reality include, among

others, Knight and proponents of the rational expectations hypothesis [51.75]. In proposing a simple lexicon to discuss of probability, [51.76] argues that there are only two basic ideas of what probability is. One of these is the idea that the probability of an event is its long-run frequency of occurrence. Following Hacking's classification of probability types, a first conception is that probability is based on the proportion, percentage, or fraction of times that an event occurs in repeated trials: this is called *aleatory* probability (from the Latin word for a dice game, *alea*). The other conception of probability focuses on the uncertainty of the final outcome; hence, a probability is the degree of belief that one has in a hypothesis given some evidence. Because this conception depends on one's knowledge of the likelihood of an event, rather than solely on its relative frequency, it is called epistemic probability (from *episteme*, the Greek for knowledge). In addition to the aforementioned dichotomy, we can then distinguish between objective and subjective conceptions of probability. In the former, probabilities are unique and have the same value for all individuals, in the latter are individuals themselves assigning values to probabilities.

*Knight* distinguishes three types of probability situations [51.48, p. 225]:

1. *A priori probability*. Absolutely homogeneous classification of instances completely identical except for really indeterminate factors. This judgment of probability is on the same logical plane as the propositions of mathematics (which also may be viewed, and are viewed by the writer, as "ultimately" inductions from experience).
2. *Statistical probability*. Empirical evaluation of the frequency of association between predicates, not analysable into varying combinations of equally probable alternatives. It must be emphasized that any high degree of confidence that the proportions found in the past will hold in the future is still based on an a priori judgment of indeterminateness. Two complications are to be kept separate: first, the impossibility of eliminating all factors not really indeterminate; and, second, the impossibility of enumerating the equally probable alternatives involved and determining their mode of combination so as to evaluate the probability by a priori calculation. The main distinguishing characteristic of this type is that it rests on an empirical classification of instances.
3. *Estimates*. The distinction here is that there is no valid basis of any kind for classifying instances. This form of probability is involved in the greatest logical difficulties of all, and no very satisfactory discussion of it can be given, but its distinction from

the other types must be emphasized and some of its complicated relations indicated.

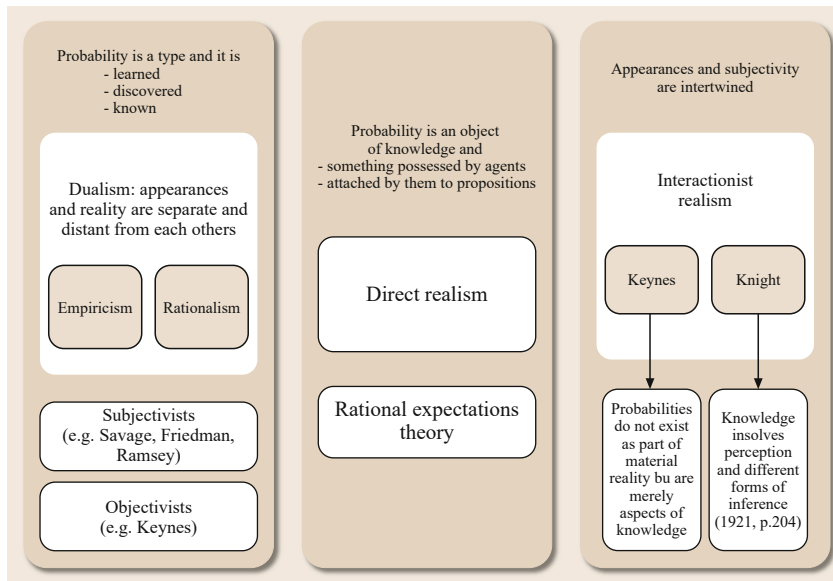
In (1) and (2), probability judgments are referred to as a priori probability and statistical probability, respectively. Concerning a priori probability, Knight stresses the "absolutely homogeneous classification of instances completely identical except for really indeterminate factors" [51.48, p. 224]; and in referring to statistical probability, *Knight* emphasizes that "any high degree of confidence that the proportions found in the past will hold in the future is still based on an a priori judgment of indeterminateness" [51.48, p. 225]. In the third type of probability situation the relevant judgment that is formed is referred to as an estimate. The first type of uncertainty is commonly referred to as risk, against which we can insure. Such a classification has been often translated by commentators into a differentiation between objective and subjective probabilities. For subjectivist theorists including, for example, Savage, Ramsey, and Friedman, probability cannot be known and it is the degree of belief in a given proposition or event, held by an actual individual at some specific point in time. On the contrary, scholars such as Lucas (in conversation with *Klamer* [51.77]) focus on the fact that what matters in terms of uncertainty is not that we do not know the probabilities but that we do not know the *classification* of outcomes.

With the aim of answering the question whether "are probabilities, as understood with economic analyses, a property of this external material reality, or are they only a property of knowledge?" *Lawson* [51.75, p. 40], works out different accounts in the notion of probability from the perspective of realism. The choice of such a stance is justified as useful for the need to state the presence of an [51.75, p. 39]

"objective material (physical and social) world which exists independently of consciousness (or of any individual consciousness in the case of the social) but which is knowable by consciousness."

Indeed, the author distinguishes amongst direct realism, dualism, and interactionist realism. Figure 51.9 and Table 51.1 illustrate Lawson's argument.

- Direct realism: According to such a perspective knowledge is reality.
- Dualism: Appearances and reality are separate and distant from each others. Inside dualism, we can distinguish among two streams of thought that are empiricism and rationalism. The former gives primary role to sensory experience in knowledge (e.g., Locke), while the latter gives central role to reason where first chaotic impressions are at a later mo-



**Fig. 51.9** Different notions of probability from the stance of realism (after [51.75])

**Table 51.1** Classification of prominent accounts of probability and uncertainty in economic analysis (after [51.75])

	Probability is a property of knowledge or belief	Probability is also an object of knowledge as a property of external material reality
Uncertainty corresponds to a situation of numerically measurable probability	Subjectivists (e.g., Savage, Friedman)	Proponents of the rational expectations view (e.g., Muth, Lucas)
Uncertainty corresponds to a situation of numerically immeasurable probability	Keynes	Knight

ment transformed and ordered by processes of the human reason (e.g., Kant).

- Interactionist realism: Appearances and reality are intertwined. Knowledge is an active process where thought, theory, and practice have equal weights in the process.

**Expected Utility Theory and Subjective Expected Utility Theory Models.** In the last 60 years, the leading theories of choice in economics and psychology have been the expected utility (EU) theory of [51.78] and the subjective expected utility (SEU) theory of *Savage Leonard* [51.79]. EU assumes that the probabilities of outcomes are known. Agents’ preferences are represented by real-valued utility functions where preferred choices correspond to higher utility, and the utility of a choice is the EU of expected possible outcomes weighted by the probability of their occurrence.

In SEU, probabilities are not necessarily objectively known, decision makers face uncertain states and are assumed to have subjective probabilities attached to these states. Under the SEU, axioms preferences can be represented by expected utilities that use subjective probabilities to weigh the probability of outcomes’ oc-

currence. The theory combines the von Neumann and Morgenstern EU approach with *De Finetti’s* [51.80] calculus of subjective probabilities. The key elements of EU theory are (1) a value function that is concave for gains, convex for losses, and steeper for losses than for gains, and (2) a nonlinear transformation of the probability scale, which overweight small probabilities and underweight moderate and high probabilities [51.81]. SEU theory was first developed by Savage (inspired by Ramsey and De Finetti), and then derived by Anscombe and Aumann in an approach that essentially combined EU and SEU. According to subjective probability theory – typically represented by the Bayesian approach – it is possible to assign numerical probabilities to virtually any proposition or event probability is the degree of belief in a given proposition or event held by an individual at a specific point in time. According to [51.75] for this group of scholars, probability is only epistemic, a property only of knowledge or belief, that does not necessarily has to correspond to external reality. As *Weatherford* [51.82, p. 226] asserted a “subjectivist [...] recognizes that his opinion is the final authority [...]. There is no correct reference class, since there is no correct probability.”

A person's subjective probability regarding the truth of a proposition (or the occurrence of an event) is revealed by the odds at which that person is exactly indifferent between betting for and against the proposition (or event) [51.83, p. 338]:

“For example, if a person is willing to accept to pay  $P^*$  for a gamble that pays  $S$  if proposition  $h$  is true and nothing if  $h$  is false, then for  $P^*/S$  to express the person's subjective probability it is necessary that the person be also willing to receive  $P^*$  for a gamble that involves a loss of  $S$  if  $A$  is true and nothing if  $A$  is false.”

This requires very precise beliefs, but presumably allows subjective probabilities to be assigned even to unique events [51.84]. According to Keynes, no proposition is subjectively probable, meaning that the likelihood for it to happen depends on our belief in it.

Radical Subjectivists, among which we can mention *Shackle* [51.85] and *Lachmann* [51.86], have a subjectivist view of expectations. The key point is that, for Shackle and his followers, imagination cannot be brought under the cover of reason: “expectation undermines the view of conduct as purely rational” [51.85, p. xvii]. According to such a view, agents vary not only in their tastes but also in their expectations, that is, in their visions of the future [51.87, p. 29]:

“In this view, the future is not so much unknown as it is nonexistent or indeterminate at the time of decision. The agent's task is not to estimate or discover, but to create. He must therefore exercise imagination.”

**The Concept of Ambiguity.** EU theory has been for several decades the dominant normative and descriptive model of decision-making under uncertainty, but at the same time a substantial body of evidence shows that decision makers systematically violate its basic assumptions [51.81].

The most famous challenge to SEU has been posed by the Ellsberg paradox. Nonetheless similar problems were argued earlier by *Knight* [51.48, pp. 218–219] as well as by *Keynes* [51.88]. Much empirical evidence *Camerer* and *Weber* [51.89], inspired by *Ellsberg* [51.90] and others, shows that people prefer to bet on events they know more about, even when their beliefs are held constant. In *Ellsberg's* [51.90, p. 657], ambiguity is the “quality depending on the amount, type, reliability, and *unanimity* of information,” or “Ambiguity is uncertainty about probability, created by missing information that is relevant and could be known” as defined by *Camerer* and *Weber* [51.89, p. 325]. According to these Authors it is misleading to suppose that ambi-

guity about outcomes and ambiguity about probabilities are parallel conditions or treatment variables [51.89, p. 331]:

“If people are averse to ambiguity about which outcome will occur, but outcome probabilities are known, then they are risk averse and consistent with EU. But if people are averse to ambiguity about the probability of an outcome, they are ambiguity averse and inconsistent with SEU. The two kinds of ambiguity are fundamentally different.”

**Ergodicity and Nonergodicity.** Ergodic theory has been explicitly developed in the theory of stochastic processes although the term derives from statistical mechanics [51.91]. Samuelson has argued that economics' claim to be scientific rests on the acceptance of the *ergodic hypothesis* [51.92]. To Keynes the source of uncertainty was in the nature of the real – nonergodic – world. It had to do, not only – or primarily – with the epistemological fact of us not knowing the things that today are unknown, but rather with the much deeper and far-reaching ontological fact that there often is no firm basis on which we can form quantifiable probabilities and expectations at all. Post-Keynesians economists typically distinguish between ergodic and non ergodic processes: the latter involving fundamental uncertainty and non ergodicity can be used to explain the existence of firm in the long run [51.93]. The fact that real social and economic processes are nonergodic is given by the pervasive presence of uncertainty.

*Davidson* [51.94] stresses the distinction between ergodic and nonergodic world and argues that uncertainty is associated with the latter situation. He concludes that the implication of this is that the inverse of knowledge, unknowledge, that is (fundamental) uncertainty, must be expounded in terms of the ergodic theory. The concept of ergodicity is described by *Davidson* [51.94, p. 6] as a situation in which

“the probability distribution of the relevant variable calculated from any past realization tends to converge to the probability function governing the current events and with the probability function that will govern future outcomes.”

The assumption of *bounded rationality* – conceived by *Herbert Simon* – refutes the assumption of rationality in the classical economic theory of the firm, arguing that its limits arise distinctly (i) when risk and uncertainty are introduced into the demand function, the cost function, or both; and (ii) when we assume that actors have incomplete information about all alternatives for their choices [51.95]. When uncertainty is introduced, a maximization behavior is replaced

by an approximation one, and actors do not optimize choices, but approximately optimize them: a satisfying behavior in decision-making is found to be more appropriate [51.95].

### 51.7.2 Models of Risk

There are numerous conceptions of risk both in the economics and business studies literature. According to most of the literature in economics, finance, and strategic management risk is associated to a variation in outcomes' performance.

*Luce* and *Raiffa* [51.96, p. 13] distinguish among three types of situations under which the decision-maker is in the realm of certainty, risk, or uncertainty:

1. Certainty is a situation where each action is known to lead invariably to a specific outcome.
2. Risk is linked to situations where each action leads to one of a set of possible specific outcomes, each outcome occurring with a known probability.
3. Uncertainty, where actions may lead to a set of consequences, but where the probabilities of these outcomes are completely unknown.

*Business risk* is defined to be the "risk inherent in the firm, independent of the way it is financed" [51.97, pp. 207–208] and it is related to the variability of net operating income or net cash flows.

*Financial risk* is defined as the added variability of the net cash flows of the owners of equity that results from the fixed financial obligation associated with debt financing and cash leasing [51.98]. Financial risk can be identified by the following equation [51.99, p. 561].

$$FR = \frac{\sigma_2}{cx - I} - \frac{\sigma_1}{cx}, \quad (51.11)$$

where  $\sigma_1$  is the net standard deviation of net cash flows without debt financing;  $\sigma_2$  is the standard deviation of cash flows with debt financing, but before the deduction of debt servicing payments,  $cx$  is the expected net cash flows without debt financing; and  $I$  are fixed debt servicing obligations.

The most important paradigm of risk emerged in the financial economics literature during the past quarter century is part of the set known in the literature as the *SBL model* (i. e., Sharpe–Lintner–Black model), known in the strategic management literature as the Capital Asset Pricing Model (CAPM). Such a model was developed by *Sharpe* [51.100], *Lintner* [51.101] and refined by *Black* [51.102] as an extension and simplification of earlier work by *Markowitz* [51.103]. The CAPM's main predictions and are summarized by *Fama* and *French* [51.104, p. 427]. Most of the studies using the CAPM [51.100, 101] framework employ measures of:

1. Systematic risk reflecting the sensitivity of the return on a firm's stock to general market movements.
2. Unsystematic risk refers to the extent to which general market movements cannot explain a firm's stock return.

Systematic and unsystematic risks are standard measures of risk for stock market return data. *Miller* and *Bromiley* [51.105] found that both types of risk, defined with accounting data, influenced performance. The definition of risk as unpredictable variation in business outcomes has to be found also in accounting literature in the form of return on assets, and return on equity. Hence, one measure of risk based on financial ratios is, for instance, the debt-to-equity ratio: a standard measure of corporate financial leverage reflecting a company's risk of bankruptcy [51.106, 107], and capital intensity, that is, the ratio of capital to sales.

Although the CAPM was developed explicitly for use in a finance context, it has been widely employed also in the field of strategic management, especially regarding issues related to corporate diversification strategy or in situations where maximization of stockholder wealth is taken as the primary objective of the firm. Variance is also a widely measure of risk used in the strategic management literature. Risk has been rarely addressed as a specific area of study in strategy formulation [51.108]. An analysis of the role of risk in strategic planning is given by *Gluck* et al. [51.109] with the identification of four phases in the development of strategic management within a firm.

In the acceptance that risk is variation from an expected probability distribution of outcomes, the relative magnitude of risk could be defined by the amount of dispersion in that distribution such as the standard deviation or variance:

1. *Variance* is a measure of the variability of a measured datum from the average value of the set of data;

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (xi - x) . \quad (51.12)$$

2. *Standard deviation* is a measure of the dispersion of a set of data from its mean. The more spread apart the data, the higher the deviation. Standard deviation is calculated as the square root of variance

$$\sqrt{\sigma^2 = \frac{1}{n} \sum_{i=1}^n (xi - x)} \quad (51.13)$$

Since distributions with different shapes and differing amounts of downside risk (the difference between

the actual return and the expected return) can have the same variance, measures such as *skewness* and *kurtosis* can be used to better quantify the risk that is not adequately described by variance alone. *Value at risk* (VaR) is a widely used risk measure of the risk of loss on a specific portfolio of financial assets. For a given portfolio, probability and time horizon, VaR is defined as a threshold value such that the probability that the mark-to-market loss on the portfolio over the given time horizon exceeds this value (assuming normal markets and no trading in the portfolio) is the given probability level. Given some confidence level  $\alpha \in (0, 1)$ , the VaR of the portfolio at the confidence level  $\alpha$  is given by the smallest number  $l$  such that the probability that the loss  $L$  exceeds  $l$  is not larger than  $(1 - \alpha)$

$$\begin{aligned} \text{VaR}_\alpha &= \inf \{l \in R: P(L > l) \leq 1 - \alpha\} \\ &= \inf \{l \in R: F_L(l) \geq \alpha\}. \end{aligned} \quad (51.14)$$

### 51.7.3 Models of Uncertainty

Several decades ago, Knight and Keynes, each in his own way, discussed uncertainty as a notion distinct from something else, which Knight called risk. The Keynesian approach to uncertainty comes straight from his studies on probability, we previously mentioned in the Treatise on Probability, and notion of logical probability.

#### Keynes on Uncertainty

The manner in which Keynes deals with uncertainty appears, on its face, quite similar to that of Knight. According to *Keynes* [51.88, p. 15]; the probability is the “degree of truth of a proposition contained in the evidence,” and the *weight of the argument* of a given proposition is the measure of the disbelief in the truth of such a proposition [51.88, p. 84]

“There appear to be four alternatives. Either in some cases there is no probability at all; or probabilities do not all belong to a single set of magnitudes measurable in terms of a common unit; or these measures always exist, but in many cases are, and must remain, unknown; or probabilities do belong to such a set and their measures are capable of being determined by us, although we are not always able so to determine them in practice.”

Uncertainty is linked to the concept of perfect knowledge and to the problem of investment decisions under uncertainty (the *weight of the arguments*).

In Keynes’s concept of uncertainty not only some premises may be unknown at the moment of decision but they may also actually be unknowable. Uncertainty

means the acknowledgement of the impossibility of dealing logically with this complexity

*Keynes* suggested that agents form their expectations based on how much *weight* they put on different possibilities of outcomes (i. e., they for subjective probabilities). In order to clarify his concept of uncertainty, we may quote a famous passage from *The General Theory of Employment, Interest, and Money* [51.3, p. 214]:

“The sense in which I am using the term is that in which the prospect of a European war is uncertain, or the price of copper and the rate of interest twenty years hence, or the obsolescence of a new invention, or the position of private wealth-owners in the social system in 1970. About these matters there is no scientific basis on which to form any calculable probability whatever. We simply do not know. Nevertheless, the necessity for action and for decision compels us as practical men to do our best to overlook this awkward fact and to behave exactly as we should if we had behind us a good Benthamite calculation of a series of prospective advantages and disadvantages, each multiplied by its appropriate probability, waiting to be summed.”

It seems that Keynes viewed the world as consisting of different degrees of uncertainty rather than in a dichotomy of uncertainty and probabilistic certainty. A related issue raised by *Coddington* [51.110] is whether or not certainty is obtainable. Moreover in Keynes’ account such certainty corresponds to knowledge. Keynes, however, does consider that knowledge and thus certainty are obtainable, he distinguishes between two types of knowledge: knowledge that is obtained directly and that which is obtained indirectly.

The distinction between risk and uncertainty in Keynes regards the distinction between short-term and long-term expectations. Unlike short-term expectations, long-term expectations (investments) do not entail continuous process of daily revision on the basis of actual market outcomes, rather involve long-term and largely irreversible commitments in advance of actual market outcomes. A risky situation is characterized by stable deterministic and stochastic components of which the decision maker has a very high degree of knowledge; all the range of future possible outcomes is known [51.111, p. 626],

“decision-makers are essentially backward looking, looking at past outcomes as a guide to future actions. Decision makers act on the assumption that the causal structure will remain fixed, at least in the short term, and form their expectations accordingly.”

An uncertain choice situation is that the probability distribution is partially, or totally unknown because only partial knowledge is possible for future outcomes and their likelihood. In such a situation, past offers little guidance to actual decision-making and firms have to form long-term expectations for investment decisions [51.111, p. 627]:

“Decision makers have to be forward-looking and *scientific*, in the sense of being willing to entertain alternative hypotheses about the causal structure and possible future outcomes, and able to interpret new information as confirming or disconfirming evidence that increases/decreases the degree of belief in any specific hypothesis.”

According to *Minsky* [51.31, p. 359] in its remarks upon receiving the Veblen-Commons award:

“The uncertainty that permeates the economics of Keynes and the economics of bounded rationality is due to the un-sureness about the validity of the model of the economy that enters in the decision process. Action involves a suspension of disbelief by both sides in the negotiations, and economic success fosters such a suspension. Government institution building can be interpreted as adding dimensions to the economy whose behaviors are not as uncertain as that of market-determined variables.”

According to *Minsky*, at a systemic level, money manager capitalism doubled with New Deal reforms have drastically increased the level of uncertainty.

#### Post-Keynesians on Uncertainty

*Dequech* [51.112] contends that uncertainty has proved to be a key concept in post-Keynesian economics. Important theoretical implications have stemmed from it:

1. It is the source for liquidity preference
2. Some post-Keynesians have derived a privilege of short-run analysis from uncertainty [51.113]
3. The possibility of structural breaks and sudden shifts in behavior [51.3, 114] and
4. The rejection for ergodicity [51.114].

The post-Keynesian Scholar argues that the rationale for the existence of the firm is not that it is efficient in the sense that it reduces transaction costs. Rather the institutional form of the firm is chosen to deal with uncertainty. Scholars of different schools of heterodox economic thought have identified situation of uncertainty of a more radical type, that is, characterized by the possibility of creativity and structural change, hence by “significant indeterminacy of the future” [51.112]. In a dynamic context, “the future cannot be anticipated

by a fully reliable probabilistic estimate because the future is yet to be created” [51.112, p. 41]. Such a kind of more radical uncertainty does not only refer to the lack of information needed to assess the probability distribution of future outcomes, rather to the impossibility of even imagining an event. This argument has clear antecedents in the work of *Shackle* [51.85], who is against the use of probability distributions, even subjective ones, in situations of fundamental uncertainty. A similar point has been raised against both the rational expectations hypothesis and SEU theory by *Bausor* [51.115], *Davidson* [51.116], and *Vickers* [51.113], among others.

According to *Dequech*, in his acceptance of uncertainty, Keynes refers both to situations of ambiguity and of fundamental uncertainty, without explicitly distinguishing between them. There is still disagreement around the interpretation of Keynesian uncertainty, based on the author’s treatise. Some commentators have argued that uncertainty refers to an absence of numerically determinate or even comparable probabilities, while, others contend that Keynesian uncertainty is measured by weight [51.112].

The concept of fundamental uncertainty claims that significant parts of economic decisions are made under conditions where the outcome of these decisions is not subject to a probabilistic calculus, but rather cannot be determined by scientific means. Thus, decisions under such circumstances are based on emotions or conventions. Typically, such conditions apply to decisions that involve a long time horizon and involve irreversible costs, what *Shackle* called *crucial experiments*. Fundamental uncertainty is a result of the fact that economic processes in the real world do not follow ergodic pattern. Under such circumstances no probability distribution for outcomes can be given. This inability to give probability does not merely reflect the limited knowledge or information processing abilities of humans, but is a reflection of the openness of the historical process in which human societies and economies evolve.

Scholars such as *Carabelli* [51.117], *Davis* [51.118], and *Arestis* [51.119] see the problem of agents’ uncertainty about each others’ expectations as being the fundamental source of general uncertainty [51.120]. Since we can never know for certain at what *degree* other people will be thinking about how average opinion will be forming its expectation of itself, we cannot ever know for certain what other peoples’ expectations of average opinion will actually be. This dependence of people for each other on the formation of their expectations, opens up the possibility of sudden mass changes of these expectations, and such an event goes under the name of *mob psychology*.

### Knight on Uncertainty

As already mentioned, one of the most important contributions to the conceptualization of uncertainty in economics is the seminal work of *Knight*, *Risk, Uncertainty, and Profit* [51.48]. Knight has both a conceptualization of what uncertainty is, and how individuals cope with it. Knight's theoretical framework begins by studying some with some very crucial points of the theory of thought. Among many other definitions [51.89] the economic understanding of risk, uncertainty, and ignorance has its origin in [51.48]. There has been a considerable discussion and disagreement about over the meaning of Frank Knight's risk and uncertainty [51.121], considering his two most famous contributions *Risk, Uncertainty and Profit*, and *Profit and Entrepreneurial Functions* [51.122]. A most common definition of risk is that it is related to outcomes that can be insured against, and uncertainty to outcomes that cannot be insured against [51.123]. Such an interpretation builds on the Knightian distinction of three possible future outcomes, and related probabilities: (i) a priori, (ii) statistic, and (iii) estimates which we have described in the previous section, hence on the measurability/immeasurability of probabilistic outcomes [51.48, p. 20, Emphasis author's own]:

“It will appear that a measurable uncertainty, or *risk* proper, as we shall use the term, is so far different from an unmeasurable that it is in effect not an uncertainty at all. We shall restrict the term *uncertainty* to cases of a nonquantitative type.”

Outcomes of the first (i) and second (ii) type can be grouped together as they are homogeneous instances, and hence, they can be insured against, while outcomes of the third type, of which we do not neither know the distribution (iii) nor we can draw it from historical data cannot be grouped, and insured against. Outcomes of the first and second type are risky, outcomes of the third type are uncertain. Outcomes subject to risk can be insured against, either through traditional insurance contracts or by holding a portfolio of stocks. *LeRoy* and *Singell* [51.49] offer a refinement on the insurance interpretation claiming that Knight anticipated the literature on the failure of markets as a result of adverse selection and moral hazard.

The second interpretation of Knight's distinction of risk and uncertainty relates to the difference between situations in which profit cannot exist (because of the presence of risk), and those in which profit can arise (thanks to the presence of uncertainty). *Knight's* aim in *Risk, Uncertainty, and Profit* [51.48] was to explain profit as the reward for bearing uncertainty [51.48, p. 232]:

“It is this true uncertainty which [...] gives the characteristic form of *enterprise* to economic organization as a whole and accounts for the peculiar income of the entrepreneur.”

A further interpretation is the one by *Langlois* and *Cosgel* [51.51], which claims that the Knightian distinction refers to states of the world that can be conceived and those that cannot. A risky decision is defined as a decision with a range of possible outcomes with a known probability for the occurrence of each state (e.g., a fair roulette game); or the probabilities are not precisely known and a decision has to be made under uncertainty (e.g., sport events and elections). In this sense, decisions under risk can be seen as a specific case of decisions under uncertainty with precisely known probabilities.

A practical example of the difference between Knightian risk and uncertainty is the one given by *Gueron-Quintana* [51.124, p. 10]. Suppose to throw a coin knowing that it is fair, and the unknown is whether the coin will land heads or tails. Since it is fair, we know that the odds for each flip to have either head or tail are 50-50. In such a case we know exactly the odds of each of the possible events: 50% heads and 50% tails, and we have this knowledge before starting the experiment. This is an example of Knightian risk. A second experiment involves flipping a coin that is no longer fair, furthermore the coin is replaced with a new (and unfair) coin after each flip. In this case we do not know the odds of obtaining heads, the only thing we know is that the coin will land either heads or tails. If we were thinking about flipping the coin 100 times, we could not (before we start the experiment) tell how many times the coin will land on heads. This is an example of Knightian uncertainty.

### Game Theory

Game theorists study uncertainty about the *strategic choices* of other players. They deal with uncertainty by reducing economy to a static equilibrium in which all economic activities take place at fixed one point in time. We can distinguish among three major stages in the development of game theory. The first one, classical game theory, is defined by John von Neumann and Oskar Morgenstern. The axioms for the concept of individual rational player making decisions in the face of certainty and uncertainty are developed. Such a player does not assume that the other players also act rationally. In contrast, modern game theory is defined by the Nash player who is not only rational but assumes that all players are rational to such a degree that they can coordinate their strategies so that a Nash equilibrium prevails. The more recent, third stage in the develop-



ment of game theory, new game theory, is defined by the so-called *Harsanyi player*. This player is rational but knows very little about the other players, e.g., their payoff functions or the way they form beliefs about other players' payoff functions or beliefs.

### Uncertainty in General Equilibrium Theory

The *General Equilibrium Theory*, as developed mainly by Arrow and Debreu in the 1950s, sets out to prove that the possibility of a competitive equilibrium in the economy does exist and that such an equilibrium is Pareto efficient. The analyses of Arrow and Debreu deal with uncertainty about the environment. The *world* is divided into two sets of variables: decision variables, which are control-led by economic agents, and environmental variables, which are not controlled by any economic agent. The *General Equilibrium Theory* was claimed by Arrow and Debreu both in a mathematically rigorous manner, but also rests on specific assumptions. Most importantly they assume the so-called dated, contingent commodities that allow for future markets for all goods, through which agents can determine their entire production and consumption plans, for they know the prices of all goods in all future periods, and they can insure them against all eventualities. Indeed, there exist markets for all actual and future goods because of the assumption that all economic actors share the same information.

Arrow himself sought the limitation of the General Equilibrium Analysis in light of informational asymmetries, leading to agency problems, moral hazard and adverse selection. Once asymmetric information is considered it leads to the arising of market failures hence Pareto-inefficient markets.

### New Institutional Economics

According to *Alchian* [51.125], uncertainty arises from at least two sources: imperfect foresight and human inability to solve complex problems containing a series of variables even when an optimum is definable. Under uncertainty each action is identified with a distribution of potential overlapping outcome. As Knight, Alchian sees uncertainty as the precondition for profits to arise. And coping responses are made by imitation of observed success; adaptive behavior via imitation provide opportunities for innovation. According to scholars of New Institutional Economics, such as Hodgson and Demsetz institutions have a *cognitive function* contribute reducing the mentioned complexity, and therefore ultimately to reduce perceived uncertainty.

### Transaction Costs Economics

TCE [51.64, 126, 127] theory assumes that firms' efficiency is maximized given a static set of knowledge,

technology and preference. Bounded rationality and opportunism are the two behavioral assumptions about human factors of transaction cost theory [51.128]. The former refers to the impossibility for individuals to retrieve information, and process it without error. The latter deals with the manifestation of opportunistic behaviors of two types: asymmetric information, and moral hazard. According to Williamson, the consequences of bounded rationality are less severe when transactions take place in a context with little environmental uncertainty. Williamson builds on Koopmans distinction among primary, secondary, and third type uncertainty. Primary uncertainty reflects a lack of knowledge about states of nature, such as the uncertainty regarding natural events, whereas secondary uncertainty refers to a lack of knowledge about the actions of other economic actors. Primary uncertainty also corresponds closely to state uncertainty as described by [51.61], in that both refer to the lack of knowledge about various states of nature. The third type of uncertainty might arise from opportunism, "self-interest seeking with guile" [51.129, p. 56].

Behavioral uncertainty relates to the informational problems that ensue from the coexistence of bounded rationality and opportunism [51.130]. "In circumstances where behavioral uncertainty is a pervasive and surrounds asset specific investments, then market based transaction costs are likely to be high" [51.130, p. 420]. Behavioral uncertainty in an international perspective arise from the inability of a company to predict the behavior of individuals in a foreign country. Hierarchical ownership conveys the right but not the means to control a foreign operation. Controlling foreign operations is a special skill that requires time to develop and refine. When a firm lacks such internal control mechanisms, it may reduce the chances of opportunistic behavior by shifting control to a foreign agent. Firms lacking international control related experience tend to prefer nonequity modes of entry.

*Behavioral Uncertainties*. They arise from the inability of a company to predict the behavior of individuals in a foreign country. According to transaction cost theory, behavioral uncertainty may lead to opportunistic behavior involving cheating, distortion of information, shirking of responsibility, and other forms of dishonest behavior. In order to minimize opportunisms, a company has to develop some type of control mechanisms, such as internal control, achievable through hierarchical ownership that gives the firm a legal right to control the actions of foreign-based employees. However, hierarchical ownership conveys the right but not the means to control a foreign operation. Controlling foreign operations is

a special skill that requires time to develop and refine. When a firm lacks such internal control mechanisms, it may reduce the chances of opportunistic behavior by shifting control to a foreign agent. Firms lacking international control related experience tend to prefer nonequity modes of entry. Behavioral uncertainties may be an especially important influencing factor for small and medium enterprises (SMEs), they tend to rely on the managerial abilities of one/two entrepreneurs. SMEs may not have the ability or willingness to establish a competent managerial control structure in another country and in most cases will not have the ability to send their own people to a foreign country for any extended period of time [51.131]. Therefore, behavioral uncertainties may discourage SMEs from organizing foreign operations in a hierarchical form.

**Environmental Uncertainties.** They are created by the target market environment, and refer to the risks associated with the host country, for example the ability to enforce contracts and control or other types of political and legal risks [51.64]. In other words if a company desires increased control, it has to commit additional resources. However, by committing additional resources, a firm increases its exposure to external environmental risks. In countries with high environmental uncertainty, companies may be better off selecting nonequity, low investment entry modes. By following a low-resource commitment strategy in an uncertain market, a company can retain flexibility and, if the need arises, switch partner organizations or exit the market entirely, if the situation so dictates.

### The Subjectivist School

As *Hey* [51.132] argues subjective expected utility theory, which is intrinsically bound up with subjective probability theory, is the very “foundation stone of the Economics of Uncertainty” [51.132, p. 130]. In addition, *Diamond* and *Rothschild* [51.133] provide a collection of thirty papers in a volume entitled *Uncertainty in Economics*, all of which associate uncertainty with the Ramsey/Savage subjectivist view of probability. And in their expository survey of the “analytics of uncertainty and information” this view of uncertainty is also adhered to by [51.68]. Agents are optimal forecasters, and, in the individual’s mind personal (as Savage calls them) probabilities regarding future prospects at the moment of choice govern future outcomes. These subjective probabilities need not coincide with objective distributions, even if well-defined objective distributions happen to exist.

### The Rational Expectations School

During the 1970s and 1980s, significant theoretical contributions have been made in the general area of uncertainty and expectations covering the modeling of individual decision-making under uncertainty to the implications of rational expectations in temporary general equilibrium and macro models.

Theory of Rational Expectations [51.134, 135] argued that Keynesian economics was fundamentally flawed in that it relied on the assumption of adaptive or exogenous expectations. In Sargent’s argument, *bounded rationality* means that the agents in the economy are unsure about the degree of rational belief that is warranted in the model that they use at any time to guide their action. Thus, in Sargent’s artificial world, intractable uncertainty is pervasive because the agents in the model need to learn the properties of the model from experience. The self-seeking agents are uncertain (or unsure) in their knowledge artificial world, intractable uncertainty is pervasive because the agents in the model need to learn the properties of the model from experience. The self-seeking agents are uncertain (or unsure) in their knowledge of the economy and they accept that others are also unsure. Sargent’s definition of uncertainty is similar to the Minskyan one.

The theory of rational expectations deals with the uncertainty of future events by assuming that agents can anticipate rationally the choices of other agents using the information they hold from the observation of past behavior of the agent. The models assume that economic actors behave as if they know the structure of the economy so they can deduce optimal forecasts despite the ongoing changes in the economy. If no objective probabilities can be calculated, the expectations of agents are modeled by using Bayesian decision theory, which operates with subjective probabilities.

Therefore, the model used to deal with situations of uncertainty does not change fundamentally for economists even if they assume the absence of objective probabilities because agents can attach subjective probabilities to outcomes, provided that actors share the same information and the same subjective probabilities. *Bayesian rationality* can be integrated into static economic analysis. This claim has been empirically challenged with the argument that the degree of foreknowledge and rationality attributed to agents in these advanced economic models becomes increasingly sophisticated and it becomes more and more unlikely that economic actors understand all relevant variables of the model properly. But this in itself does not yet constitute a theoretical challenge that would affect the theoretical validity of economic decision-making models that deal with uncertainty.

## References

- 51.1 G. Magnus: What this Minsky Moment means, *Financial Times*, August 23 2007, retrieved on September 20 2014 from <https://next.ft.com/content/ddb7842c-50c2-11dc-86e2-0000779fd2ac>
- 51.2 G. Magnus: *The Credit Cycle and Liquidity: Have We Arrived at a Minsky Moment? Economic Insights* – By George (UBS Investment Research, London 2007)
- 51.3 J.M. Keynes: *The General Theory of Employment Interest and Money* (Macmillan, London 1964), originally published in 1936
- 51.4 H.P. Minsky: Banking and industry between the two wars: The United States, *J. Eur. Econ. Hist.* **13**, 235–272 (1984)
- 51.5 H.P. Minsky: The financial instability hypothesis: An interpretation of Keynes and an alternative to “standard” theory, *Challenge* **20**(1), 20–27 (1977)
- 51.6 P. Davidson: *John Maynard Keynes* (Palgrave Macmillan, Basingstoke 2007)
- 51.7 J.M. Keynes: *Essays in Persuasion* (W.W. Norton, New York 1963)
- 51.8 J.M. Keynes: The general theory of employment, *Q. J. Econ.* **51**(2), 209–223 (1937)
- 51.9 H.P. Minsky: *Can “It” Happen Again?: Essays on Instability and Finance* (ME Sharpe Armonk, New York 1982)
- 51.10 H.P. Minsky: Capitalist financial processes and the instability of capitalism, *J. Econ. Issues* **14**, 505–523 (1980)
- 51.11 F. Ferrara: Moneta endogena, disponibilità di credito e preferenza per la liquidità, *Studi e Note di Econ.* **1**, 87–109 (1998)
- 51.12 H.P. Minsky: *John Maynard Keynes* (Columbia Univ. Press, New York 1975)
- 51.13 R. Bellofiore: L’ipotesi dell’instabilità finanziaria e il ‘nuovo’ capitalismo, The complexity of financial crisis in a long-period perspective: Facts, theory and models workshop, 2009, Siena (University of Siena 2009) pp. 1–21
- 51.14 R. Bellofiore, P. Ferri (Eds.): *The Economic Legacy of Hyman Minsky* (E. Elgar, Northampton 2001)
- 51.15 E. Tymoigne: The Minskyan System, Part I: Properties of the Minskyan Analysis and How to Theorize and Model a Monetary Production Economy, The Levy Economics Institute of Bard College Working Paper No. 452, (The Levy Economics Institute of Bard College, Annandale-on-Hudson 2006)
- 51.16 H.P. Minsky: *Stabilizing an Unstable Economy* (Yale Univ. Press, London 1986)
- 51.17 H.P. Minsky: Financial instability, the current dilemma, and the structure of banking and finance. In: *Compendium of Major Issues in Bank Regulation*, ed. by Committee on Banking, Housing, and Urban Affairs (US Government Printing Office, Washington, DC 1975) pp. 310–353
- 51.18 D. Carson (Ed.): *Banking and Monetary Studies* (R.D. Irwin, Homewood 1963)
- 51.19 M. Kalecki: *Selected Essays on the Dynamics of the Capitalist Economy 1933–1970* (Cambridge Univ. Press, Cambridge 1971)
- 51.20 M. Passarella: Hyman P. Minsky e l’ipotesi di Instabilità Finanziaria, Lecture, Corso di Economia dei Mercati Monetari e Finanziari (2010), retrieved on 20 Sept 2014 from: <http://www.marcopassarella.it/wp-content/uploads/slides/lezioni-parte-4-varese.pdf>
- 51.21 G.P. Szego, K. Shell (Eds.): *Mathematical Methods in Investment and Finance* (American Elsevier, New York 1972)
- 51.22 I. Fisher: The debt-deflation theory of great depressions, *Econometrica* **1**, 337–357 (1933)
- 51.23 I. Fisher: *Booms and Depressions: Some First Principles* (Adelphi Company, New York 1932)
- 51.24 I. Fisher: *The Purchasing Power of Money; Its Determination and Relation to Credit Interest and Crises* (Macmillan, New York 1911)
- 51.25 L. Sau: La deflazione da debiti di Irving Fisher nell’era della globalizzazione, *Riv. Ital. Degli Econ.* **10**, 443–458 (2005)
- 51.26 I. Fisher: Reflation and stabilization, *Ann. Am. Acad. Polit. Soc. Sci.* **171**, 127–131 (1934)
- 51.27 R. Guttmann: Asset bubbles, debt deflation, and global imbalances, *Int. J. Polit. Econ.* **38**, 46–69 (2009)
- 51.28 J. Peck, N. Theodore, N. Brenner: Postneoliberalism and its malcontents, *Antipode* **41**, 94–116 (2010)
- 51.29 T.I. Palley: Financialization: What it is and why it matters, The Levy Economics Institute Working Paper No. 525 (The Levy Economics Institute of Bard College, Annandale-on-Hudson 2007)
- 51.30 M. Szabó-Pelsöczy (Ed.): *The Future of The Global Economic and Monetary System* (Institute for World Economics of the Hungarian Academy of Sciences, Budapest 1990)
- 51.31 H.P. Minsky: Uncertainty and the institutional structure of capitalist economies: Remarks upon receiving the Veblen-Commons award, *J. Econ. Issues* **30**, 357–368 (1996)
- 51.32 J. Kregel: Resolving the US financial crisis: Politics dominates economics in the New Political Economy, *PSL Q. Rev.* **64**, 23–37 (2011)
- 51.33 E. Stockhammer: Financialization and the global economy, *Polit. Economy Research Institute Working Paper No. 242* (2010)
- 51.34 G. Epstein (Ed.): *Financialization and the World Economy* (Edward Elgar, Cheltenham Glos 2006)
- 51.35 C.A. Williams, P. Zumbansen: *The Embedded Firm: Corporate Governance, Labor, and Finance Capitalism* (Cambridge University Press, Cambridge 2011)
- 51.36 W. Lazonick: The theory of the market economy and the social foundations of innovative enterprise, *Econ. Ind. Democr.* **24**, 9–44 (2003)
- 51.37 W. Lazonick: The new economy business model and the crisis of US capitalism, *Cap. Soc.* (2009), doi:10.2202/1932-0213.1062, article 4
- 51.38 E.R. Freeman, W.M. Evan: Corporate governance: A stakeholder interpretation, *J. Behav. Econ.* **19**,

- 337–359 (1991)
- 51.39 M.M. Blair: Rethinking assumptions behind corporate governance, *Challenge* **38**, 12–17 (1995)
- 51.40 R.E. Freeman: *Strategic Management: A Stakeholder Approach* (Cambridge University Press, Cambridge 2010)
- 51.41 E.R. Freeman, W.M. Evan: Corporate governance: A stakeholder interpretation, *J. Behav. Econ.* **19**, 337–359 (1991)
- 51.42 T.J. Palley: America's Exhausted Paradigm, *Real World Econ. Rev.* **50**, 52–73 (2008)
- 51.43 J.B. Foster, R. McChesney: Monopoly–finance capital and the paradox of accumulation, *Mon. Rev.* **61**, 203–227 (2009)
- 51.44 D.M. Kotz: The financial and economic crisis of 2008: A systemic crisis of neoliberal capitalism, *Rev. Radic. Polit. Econ.* **41**, 305–317 (2009)
- 51.45 P. Sweezy, H. Magdoff: *Stagnation and the Financial Explosion (Economic History as it Happened, Vol. IV)* (Monthly Review, New York 1987)
- 51.46 F.A. Hayek: *Individualism and Economic Order* (University of Chicago Press, Chicago 1948)
- 51.47 J. Beckert: What is sociological about economic sociology? Uncertainty and the embeddedness of economic action, *Theory Soc.* **25**, 803–840 (1996)
- 51.48 F.H. Knight: *Risk, Uncertainty and Profit* (Hart, Schaffner and Marx, New York 1921)
- 51.49 S.F. LeRoy, L.D. Singell: Knight on risk and uncertainty, *J. Polit. Econ.* **95**, 394–406 (1987)
- 51.50 K.M. Eisenhardt: Agency theory: An assessment and review, *Acad. Manag. Rev.* **14**, 57–74 (1989)
- 51.51 R.N. Langlois, M.M. Cosgel: Frank Knight on risk, uncertainty, and the firm: A new interpretation, *Econ. Inq.* **31**, 456–465 (1993)
- 51.52 J.A. Schumpeter: *The Theory of Economic Development. An Inquiry into Profits, Capital, Credit, Interest, and the Business Cycle* (Transaction Publishers, London 1934)
- 51.53 Forrester: *World Dynamics* (Wright-Allen, Cambridge 1971)
- 51.54 I.M. Kirzner: *Perception, Opportunity, and Profit: Studies in the Theory of Entrepreneurship* (Univ. Chicago Press, Chicago 1979)
- 51.55 R.B. Duncan: Characteristics of organizational environments and perceived environmental uncertainty, *Adm. Sci. Q.* **17**, 313–327 (1972)
- 51.56 P.R. Lawrence, J.W. Lorsch: Differentiation and integration in complex organizations, *Adm. Sci. Q.* **12**, 1–47 (1967)
- 51.57 D. Collis: The strategic management of uncertainty, *Eur. Manag. J.* **10**, 125–135 (1992)
- 51.58 R.M. Cyert, J. March: *A Behavioral Theory of the Firm* (Blackwell Publishers, Cambridge, 1992)
- 51.59 F. Emery, E. Trist: The causal texture of organizational environments, *Hum. Relat.* **18**, 12–32 (1965)
- 51.60 J.R. Galbraith: *Designing Complex Organizations* (Addison-Wesley, Reading 1973)
- 51.61 F.J. Milliken: Three types of perceived uncertainty about the environment: State, effect, and response uncertainty, *Acad. Manag. Rev.* **12**, 133–143 (1987)
- 51.62 W.R. Dill: Environment as an influence on managerial autonomy, *Adm. Sci. Q.* **2**, 409–443 (1958)
- 51.63 J.D. Thompson: *Organizations in Action* (McGraw-Hill, New York 1967)
- 51.64 O.E. Williamson: *The economic Institutions of Capitalism: Firms, Markets, Relational Contracting* (Collier Macmillan, London 1985)
- 51.65 E.T. Penrose: *The Theory of the Growth of the Firm* (Blackwell, Oxford 1959)
- 51.66 K.J. Arrow: Limited knowledge and economic analysis, *Am. Econ. Rev.* **64**, 1–10 (1974)
- 51.67 T.C. Koopmans: *Three Essays on the State of Economic Science* (McGraw-Hill, New York 1957)
- 51.68 J. Hirshleifer, J.G. Riley: The analytics of uncertainty and information—an expository survey, *J. Econ. Lit.* **17**, 1375–1421 (1979)
- 51.69 P. Milgrom, J. Roberts: Limit pricing and entry under incomplete information: An equilibrium analysis, *Econometrica* **50**, 443–459 (1982)
- 51.70 M.J. Machina: Decision-making in the presence of risk, *Science* **236**, 537–543 (1987)
- 51.71 D. Kahneman, A. Tversky: Variants of uncertainty, *Cognition* **11**, 143–157 (1982)
- 51.72 J. Johanson, F. Wiedersheim-Paul: The internationalization of the firm – Four Swedish cases, *J. Manag. Stud.* **12**, 305–323 (1975)
- 51.73 J. Johanson, J.-E. Vahlne: The internationalization process of the firm—A model of knowledge development and increasing foreign market commitments, *J. Int. Bus. Stud.* **8**, 23–32 (1977)
- 51.74 G.A. Holton: Perspectives: Defining risk, *Financ. Anal. J.* **60**, 19–25 (2004)
- 51.75 T. Lawson: Probability and uncertainty in economic analysis, *J. Post Keynes. Econ.* **11**, 38–65 (1988)
- 51.76 I. Hacking: All kinds of possibility, *Philos. Rev.* **84**, 321–337 (1975)
- 51.77 A. Klamer: *The New Classical Macroeconomics: Conversations with the New Classical Economists and their Opponents* (Wheatsheaf Books, Brighton 1984)
- 51.78 J. Von Neumann, O. Morgenstern: *Theory of Games and Economic Behavior* (Princeton Univ. Press, Princeton 1947)
- 51.79 L.J. Savage Leonard: *The Foundations of Statistics* (Wiley, New York 1954)
- 51.80 B. De Finetti: *Problemi di Optimum, Problemi di Optimum Vincolato* (Istituto italiano degli attuari, Roma 1937)
- 51.81 A. Tversky, D. Kahneman: Advances in Prospect Theory: Cumulative representation of uncertainty, *J. Risk Uncertain.* **5**, 297–323 (1992)
- 51.82 R. Weatherford: *Philosophical foundations of probability theory* (Routledge Kegan, London 1982)
- 51.83 J. Runde: On Popper, probabilities, and propensities, *Rev. Soc. Econ.* **54**, 465–485 (1996)
- 51.84 A. Riabacke, R. Ari: Managerial decision making under risk and uncertainty, *Int. J. Comput. Sci.* **32**, 453–459 (2006)
- 51.85 G.L.S. Shackle: *Epistemics and Economics: A Critique of Economic Doctrines* (Cambridge Univ.

- Press, Cambridge 1972)
- 51.86 L.M. Lachmann: From Mises to Shackle: An essay on Austrian economics and the kaleidic society, *J. Econ. Lit.* **14**, 54–62 (1976)
- 51.87 R.N. Langlois (Ed.): *Economics as a Process: Essays in the New Institutional Economics* (Cambridge Univ. Press, New York 1990)
- 51.88 J.M. Keynes: *A Treatise on Probability* (MacMillan, London 1921)
- 51.89 C. Camerer, M. Weber: Recent developments in modeling preferences: Uncertainty and ambiguity, *J. Risk Uncertain.* **5**, 325–370 (1992)
- 51.90 D. Ellsberg: Risk, ambiguity, and the Savage axioms, *Q. J. Econ.* **75**, 643–669 (1961)
- 51.91 W. Parry: Ergodic theory. In: *The New Palgrave: A Dictionary of Economics*, ed. by J. Eatwell, M. Milgate, P. Newman (Palgrave Macmillan, Basingstoke 1987)
- 51.92 R.W. Clower (Ed.): *Monetary Theory: Selected Readings* (Penguin books, Harmondsworth 1969)
- 51.93 S.P. Dunn: Fundamental Uncertainty and the Firm in the Long Run, *Rev. Polit. Econ.* **12**, 419–433 (2000)
- 51.94 P. Davidson: Sensible expectations and the long-run non-neutrality of money, *J. Post Keynes. Econ.* **10**, 146–153 (1987)
- 51.95 H.A. Simon: Theories of bounded rationality, *Decis. Organ.* **1**, 161–176 (1972)
- 51.96 R.D. Luce, H. Raiffa: *Games and Decisions: Introduction and Critical Survey* (Wiley, New York 1957)
- 51.97 J.C. Van Horne: *Fundamentals of Financial Management* (Prentice–Hall, Englewood Cliffs 1974)
- 51.98 A. Barges: *The Effect of Capital Structure on the Cost of Capital: A Test and Evaluation of the Modigliani and Miller Propositions* (Prentice–Hall, Englewood Cliffs 1963)
- 51.99 S.C. Gabriel, C.B. Baker: Concepts of business and financial risk, *Am. J. Agric. Econ.* **62**, 560–564 (1980)
- 51.100 W.F. Sharpe: Capital asset prices: A theory of market equilibrium under conditions of risk, *J. Finance* **19**, 425–442 (1964)
- 51.101 J. Lintner: The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets, *Rev. Econ. Stat.* **47**, 13–37 (1965)
- 51.102 F. Black: Capital market equilibrium with restricted borrowing, *J. Bus.* **45**, 444–455 (1972)
- 51.103 H.M. Markowitz: Portfolio selection, *J. Finance* **7**, 77–91 (1952)
- 51.104 E.F. Fama, K.R. French: The cross-section of expected stock returns, *J. Finance* **47**, 427–465 (1992)
- 51.105 K.D. Miller, P. Bromiley: Strategic risk and corporate performance: An analysis of alternative risk measures, *Acad. Manag. J.* **33**, 756–779 (1990)
- 51.106 G.J. Hurdle: Leverage, risk, market structure and profitability, *Rev. Econ. Stat.* **56**, 478–485 (1974)
- 51.107 A.C. Shapiro, S. Titman: An integrated approach to corporate risk management, *Revolut. Corp. Finance* **3**, 251–265 (1986)
- 51.108 I.S. Baird, H. Thomas: Toward a contingency model of strategic risk taking, *Acad. Manag. Rev.* **10**, 230–243 (1985)
- 51.109 F.W. Gluck, S.P. Kaufman, A.S. Walleck: Strategic management for competitive advantage, *Harv. Bus. Rev.* **58**, 154–161 (1980)
- 51.110 A. Coddington: Deficient foresight: A troublesome theme in Keynesian economics, *Am. Econ. Rev.* **72**, 480–487 (1982)
- 51.111 G. Fontana, B. Gerrard: A post Keynesian theory of decision making under uncertainty, *J. Econ. Psychol.* **25**, 619–637 (2004)
- 51.112 D. Dequech: Fundamental uncertainty and ambiguity, *East. Econ. J.* **26**, 41–60 (2000)
- 51.113 D. Vickers: *Economics and the Antagonism of Time: Time, Uncertainty, and Choice in Economic Theory* (Univ. Michigan Press, Ann Arbor 1994)
- 51.114 T. Lawson: Uncertainty and economic analysis, *Econ. J.* **5**, 909–927 (1985)
- 51.115 R. Bausor: The rational–expectations hypothesis and the epistemics of time, *Camb. J. Econ.* **7**, 1–10 (1983)
- 51.116 P. Davidson: Is probability theory relevant for uncertainty? A post Keynesian perspective, *J. Econ. Perspect.* **5**, 129–143 (1991)
- 51.117 A. Carabelli: Keynes on probability, uncertainty and tragic choices, *Pap. Polit. Econ.* **30**, 187–226 (1998)
- 51.118 J.B. Davis: *Keynes's Philosophical Development* (Cambridge University Press, Cambridge 1994)
- 51.119 P. Arestis: Post–Keynesian economics: Towards coherence, *Camb. J. Econ.* **20**, 111–135 (1996)
- 51.120 J.B. Rosser: Alternative Keynesian and post Keynesian perspectives on uncertainty and expectations, *J. Post Keynes. Econ.* **23**, 545–566 (2001)
- 51.121 G.T. Brooke: Uncertainty, profit and entrepreneurial action: Frank Knight's contribution reconsidered, *J. Hist. Econ. Thought* **32**, 221–235 (2010)
- 51.122 F.H. Knight: Profit and entrepreneurial functions, *J. Econ. Hist.* **2**, 126–132 (1942)
- 51.123 J.F. Weston: The profit concept and theory: A restatement, *J. Polit. Econ.* **62**, 152–170 (1954)
- 51.124 P.A. Guerron–Quintana: Risk and Uncertainty, *Bus. Rev.* **Q1**, 9–18 (2012)
- 51.125 A.A. Alchian: Uncertainty, evolution, and economic theory, *J. Polit. Econ.* **58**, 211–221 (1950)
- 51.126 R.H. Coase: The nature of the firm, *Economica* **4**, 386–405 (1937)
- 51.127 O.E. Williamson: *Markets and Hierarchies: Analysis and Antitrust Implications: A study in the Economics of Internal Organizations* (The Free Press, New York, 1975)
- 51.128 O.E. Williamson: Markets and hierarchies: Some elementary considerations, *Am. Econ. Rev.* **63**, 316–325 (1973)
- 51.129 O.E. Williamson: Economic organization: The case for candor, *Acad. Manag. Rev.* **21**, 48–57 (1996)
- 51.130 S. Dunn: *The 'Uncertain' Foundations of Post Keynesian Economics: Essays in Exploration* (Routledge, London 2010)
- 51.131 J.W. Lu, P.W. Beamish: The internationalization and performance of SMEs, *Strateg. Manag. J.* **22**, 565–586 (2001)

- 51.132 J.D. Hey: Whither uncertainty?, *Econ. J.* **93**, 130–139 (1983)
- 51.133 P. Diamond, M. Rothschild: *Uncertainty in Economics* (Academic Press, New York 1978)
- 51.134 R.E. Lucas: Expectations and the neutrality of money, *J. Econ. Theory* **4**, 103–124 (1972)
- 51.135 R.E. Lucas, T. Sargent: After Keynesian macroeconomics, *Ration. Expect. Econ. Pract.* **1**, 295–319 (1981)

## 52. Application of Models from Social Science to Social Policy

Eleonora Montuschi

The use of models in social science is now widely acknowledged, and well beyond cosmetic or illustrative purposes. However, the details and the mechanics of their use still prove hard to pin down. Equally, the usefulness of social scientific models in social practice and intervention is often challenged by a number of contentious and recurrent issues. One of these issues is ontological: How do model descriptions and aspects of social reality relate to each other? Often the descriptions offered by models are thin and unrealistic. Can we (and how much) learn about what goes on in the real-social world by analyzing the way/s that world gets described or explained by a model? A second issue is methodological: why using models when we can design experiments in the social world that are able, with some rigor, to inform us on *what works*? Nowadays there is an established trend to prefer the results achieved, for example, by well-conducted randomized control trials, by many considered the golden rule to doing good and useful social science.

In this chapter, we will first show some of the limitations and costs of using models in representing real-life situations, and suggest some strate-

Ever since it was acknowledged that models do not only perform a cosmetic or a heuristic function but, more substantially, can capture parts or features of the real world and fulfil an explanatory function, a vast philosophical literature has been developed on the many ways model ontology and reality relate to each other [52.1–4] (This change in perspective originates in the shift from the syntactic to the semantic view of theories in the philosophy of science, between the 1960s and the 1970s.). Models can simplify, approximate, idealize, or abstract from target systems (e.g., the behavior of a gas, the trajectory of a planet, the decision of a board committee, long-term average inflation rates, the migration movements of a population in a suc-

52.1	<b>Unrealistic Assumptions</b> .....	1105
52.1.1	Quality of Life.....	1105
52.1.2	QALY: What Are We Measuring?.....	1106
52.1.3	Unrealistic Utility Assumptions.....	1107
52.2	<b>Real Experiments, Not Models Please!</b> .....	1110
52.2.1	Causal Models: How Can They Come to the Rescue .....	1111
52.2.2	Class Size Reduction Programmes.....	1111
52.2.3	TINP and BINP .....	1112
52.2.4	Children Mortality and Inflicted Death .....	1114
52.3	<b>Conclusions</b> .....	1115
	<b>References</b> .....	1116

gies by which we can still formulate informative inferences from model to target system. We will then point out some of the virtues and benefits of using models (particularly causal models) when what is at stake is not only answering the question *what works* in policy terms, but also *why it works* – or, even more interestingly, why it does not work in given circumstances.

cession of historical periods) that often appear complex and difficult to be described or comprehended as they stand. Despite differences in construction, one feature that all models – be it formal/quantitative, or qualitative, or a combination of the two – have in common is that of being representational devices [52.5]. Models' representational power is displayed by establishing appropriate relationships with the target system (this is a theory, or a phenomenon, depending on the models concerned). Appropriate in what sense? How are these relationships established?

*The World in the Model* – the allusive title of a recent book by Morgan [52.6] – points us in two equally suggestive directions of thought. It refers to the world

the model creates, out of its own premises, variables, operations and conditions. Or it questions what of the world as we know it (or don't know it) gets captured (portrayed, described, explained) by a model.

Undoubtedly some models create ontologies that take shape out of their own principles and goals. But when, for example, we try to represent the real functioning of, say, a magnetic field, or a chemical reaction, or a decision on whether to implement a certain educational policy, how far do these ontologies travel from model to target?

A problem that is consistently pointed out in the vast and heterogeneous literature on models is that a good number of them offer descriptions of imaginary situations (model ontologies) which, compared to the target situations (real-world ontologies), are both thin (they are built on very few details) and, more worryingly, unrealistic (the details they are built on are untrue). These are models that routinely use a mixture of formal and natural language and that are built in such a way as to provide their results by deductive inference. It is, however, pointed out that thin and unrealistic assumptions prove necessary to these models in order to obtain the degree of rigor they aspire to, namely the type of rigor achieved deductively by means of controlled variables and operationalized language [52.7]. Therefore, thin and unrealistic assumptions are not a problem for the models as such. They nonetheless become a problem when a “result that must occur given characteristics different from those in the target inform conclusions about what will happen in the target.” [52.7, p. 3]. The problem just described takes then the form of the following question: Can thin and unrealistic premises in the model give rise to realistic conclusions concerning the target? When the target is *the world* (real, occurring situations in the world we live in and that we try to explain, predict, etc.) then the problem becomes deep indeed. This does not imply that unrealistic assumptions prevent us from learning something valuable about targets, but how exactly does this happen requires clarification.

In the first half of this chapter, we will pay specific attention to this problem, by showing the virtues as well as the limits of using the imaginary to model the real. The domain of reality where this problem will be tested is that of social settings. Modeling decisions, individual behaviors, social conducts, and social causes inevitably confront us with the issue of how much and how well does a model construction guide us into de-

scribing, understanding, and predicting what goes on in the complex realms (and often the vagaries) of real decisions, behaviors, etc. – as the first example analyzed in the following section will prove. If models in social science are to be explanatory and not just illustrative, if what they describe is to be of real use and guide in intervening on problematic or contentious social situations, the nature of their assumptions and the quality of their inferences deserve careful analytic scrutiny.

Learning how to master the use of models in social analysis will also bring us to address another important issue. Nowadays there is a consistent movement toward the use of experiments in the social world, and particularly certain types of experiments (e.g., RCTs) are considered to be the best methodology to understand and intervene on social problems. We intend to show how social experiments on their own do not often secure the understanding and the effective answers they promise to offer, and how the appeal to adequately formulated models (and a good use of them) is able to explain why. In the second half of the chapter, we will deal with how a particular category of models – causal models – relates to some types of social experiments, and with what consequences. We will briefly describe what causal models are, and we will then detail what we can learn from using them that we cannot apprehend by following a strictly experimental methodology. A number of examples will assist us in this task. They will point out:

1. How causal models can identify what necessary factors are decisive in increasing the chances of success of a course of action or an intervention
2. How using these models makes us realize how the results of an experiment can work in different contexts, or identify the reasons why they do not/cannot work in some
3. What is entailed by the expression *why something happens the way it happens*, and why causal model are well equipped to achieve this goal.

The overall aim of this chapter is not to provide a comprehensive taxonomy of social models, nor to offer a literature review (a task well pursued in Chap. 42 in this handbook). Here we focus on a selected range of issues that make us reflect on the *use* and *usefulness* of models (or at least some types of models) in social practice.



## 52.1 Unrealistic Assumptions

Cartwright claims that at least some models function like *Galileian thought experiments*. With these models, arguably, unrealistic assumptions are a necessity, not a hindrance. What is a Galileian thought experiment? In *Cartwright's* own words [52.7, p.4]:

“A Galileian thought experiment isolates a single factor to observe its natural effect when it operates *on its own*, and *without impediment*. [...] In a Galileian thought experiment it is the principles built into the model that determine what the effect must be. [...] the thought experiment has only the factors in it that we put there. So we can be sure that confounders are absent but we cannot be sure the effect is right because that depends on the principles we provide in the model.”

Two bodies move inertially only in the absence of forces, only if they move in a Euclidean framework, only if they travel on geodesics (shortest distance between two points), etc. The loss of skills during periods of unemployment makes unemployment persist in the future if skills matter to productivity, if the creation of new jobs is motivated by firms' expected profit, if loss of skills is predominantly due to unemployment and not to other factors, etc. This model was studied by *Pissarides* in [52.8]; discussed in [52.7].

So, there is no guarantee that we learn the right lesson *vis a vis* the factors singled out by the model. Nonetheless, these models have a virtue, namely that of eliminating all sorts of intervening circumstances and interferences real situations are frothed with. And this somehow breaks it even *vis a vis* real experiments, where also we are not sure whether an effect is due to natural law or to intervening confounding factors.

However, the problem with this way models function is, if we follow Cartwright, that they find themselves by necessity introducing much more unrealistic assumptions than those required to isolate the factor from confounders. So, for example, in order to isolate a particular cause we need to produce a suitable background for that cause to operate appropriately (causes do not work in isolation). Or else, things are to be presented in rather particular ways within the model if calculations are to be made possible, and often this requires mathematically tractable descriptions which take into account almost exclusively the results that the model is equipped to handle. So these models end up being over-constrained. They are to include all the conditions and factors that the model requires in view of successfully isolating its factors. This indeed becomes a problem when we include all these conditions and factors in the deduction of the results in the model: learning

a lesson that goes outside the rather special framework set out by the model and meets its target proves difficult.

Part of the problem, Cartwright argues, is that we wrongly try to learn a lesson directly from the model, and from within its own boundaries and premises. We expect that a model behaves like a fable, from which a *moral* can be extracted. For example, in a fable like the one constructed by Lessing – “a marten eats the grouse; a fox throttles the marten; the tooth of the wolf, the fox” (Lessing 1759, Sect. I, p.73; quoted in [52.9, p. 39]) – the moral can be *read out* of the fable itself – *the weaker is always prey of the stronger*. With models we find ourselves in a different situation. The move from the model to the target is not written in the model, nor should be provided by it. And it is not obvious, nor automatic. We might need a good deal of interpretation, theory, empirical work, etc. to fill in the gap between what is stated in the model and what the model, in its own language, points out about its target (outside the model). *Filling the gap*, and in ways that cannot be straightforwardly inferred from stated premises, is the way models pursue the task of letting us learn a lesson that goes beyond their boundaries. That is why it has been suggested [52.9] that we should think of them as if they were parables – where morals are not written within them but where “the prescription for drawing the right lesson must be supplied from elsewhere” (from Greek *parabolein* → to set beside; *para* = besides; *ballein* = to throw).

Besides, when we refer to a *moral*, there are two aspects to consider: what the moral says literally (the weaker is prey of the stronger) and how we value what it says (it is wrong that the stronger takes advantage of the weaker). If the former aspect is difficult to infer directly from the model, the second is even more difficult – as we will point out by means of the example to follow.

### 52.1.1 Quality of Life

By means of this first example I will analyze a model of measurement of a social concept that is particularly relevant in terms of both research design and policy implications. I refer to the model of QALY (Quality Adjusted Life Year) that is meant to assess the worth of a treatment/intervention in the context of the *quality of life* of a patient, and by comparison with other patients' life expectancies. Models of this sort are certainly timely if we consider that, at least in modern Western society, there has been a considerable progress in cures, care, and medical technology, as well as improvements in early diagnosis of potentially terminal illnesses, or of those pathological conditions possibly leading to lethal

illnesses. As a consequence, there has been an increase in the occurrence of so-called chronic or degenerative diseases, which modern medicine is able to treat and, within limits, to control by postponing the moment of acute crisis and eventual terminal resolution.

Due to this emergent scenario, a new social attitude toward the process of illness and death has come forth. *Taking care* of the extension of time that modern medicine is able to secure to more or less severely diseased people has become a form of social duty for institutions and social organizations in modern Western society. Social investment on care leads to improved conditions of care, and better conditions of care mean increased opportunities for survival.

This new scenario, however, clashes with the well-known and established fact of the limited resources that can be allocated to care/cure in our societies. So the *quality* of life as extended by social and medical progress is to be assessed in the context of how limited resources can be best distributed in view of maximizing health and health improvements. Cost–utility analyses and econometric tools have been devised to create standard measurements that can account for both quantity and quality of life. Since the 1970s, measuring the quality of life has become a constant topic of interest for those social scientists involved in fields of research such as epidemiology, or health economics, with potential applications in practice and policy [52.10, 11]. Models of measurements have been since used widely in health-care policy making and by institutions such as the National Institute of Clinical Excellence (NICE) in the UK.

So, how can we calculate how much quality one’s life possesses which makes it, objectively and reliably, both worth living and investing resources on? QALYs are an interesting case for what we are discussing in this chapter, as the models on which these measures are based put forward a series of problems akin to those we have discussed in the preceding section.

### 52.1.2 QALY: What Are We Measuring?

QALY stands for *Quality Adjusted Life Year*. It is a formal tool that allows to quantify the health benefits consequent to some treatment/intervention, and to make comparisons with other treatments/interventions. How does it do that? By valuing health states in terms of utilities, and by valuing life-years in terms of preference weights. Health states are weighted by associating utility scores to them. *Alan Williams* summarises the QALY measurement rationale as follows [52.12]:

“The basic assumption is that it takes a year of healthy life expectancy to be worth 1, but a year

of unhealthy life expectancy is regarded as worth less than 1. Between 1 and  $-1$  there are a number of intermediate states of health (death = 0), each of which is given a value. Its precise value is lower the worse the quality of life of the unhealthy person (which is what the *quality adjusted* bit is all about).”

The efficacy of health-care interventions is quantified not only on the basis of life expectancy (how many extra years an intervention can predictably grant) but also on the quality of life that those very interventions possibly achieve. For example, an intervention that secures 10 years of extra life at full health would have a QALY value of 10. An intervention that improves quality of life from 0.5 to 0.8 for a person with a life expectancy of 30 years, would have a QALY value of 9. [0.3 (0.8–0.5) multiplied by 30] [52.13].

How is a *state of health* defined? Typically by parameters (or dimensions) and levels. For example, *Rosser* and *Kind* [52.14] define a *state of health* on the basis of two parameters: an objective invalidity and a subjective pain. Objective invalidity is subdivided into eight levels, and subjective pain into four. Both invalidity and pain are not referred to specific diseases (see Table 52.1).

How are states of health valued? As we saw above, utilities are attached to health states, based on prefer-

**Table 52.1** Classification of states of sickness (after [52.14, p. 349])

Disability	
1.	No disability
2.	Slight social disability
3.	Severe disability and/or slight impairment of performance at work. Able to do all housework except very heavy tasks
4.	Choice of work or performance at work very severely limited. Housewives and old people able to do light housework only but to go out shopping
5.	Unable to undertake any paid employment. Unable to continue any education. Old people confined to home except for escorted outings and unable to do shopping. Housewives able only to perform a few simple tasks
6.	Confined to chair or to wheelchair or able to move around in the home only with support from an assistant
7.	Confined to bed
8.	Unconscious
A.	No distress
B.	Mild distress (slight pain which is relieved by aspirin)
C.	Moderate distress (pain which is not relieved by aspirin)
D.	Severe distress (pain for which heroin is prescribed)

ences for different states, and in such a way that the most preferred (or most desirable) will receive greater weight (1 = full health; 0 = death; 0 to 1 = not full health; -1 = bad/worse/worst health). The QALYs are then calculated by multiplying the utility score by the time spent in each state.

Who attaches utility scores to health states? The values are normally established via interviews. The interviewees are either members of the public (who are asked to imagine being in a particular health state), or patients (who are experiencing, or who have actually experienced particular health states). There is a lively debate on who is in a better position to be interviewed for the purpose of calculation. In Rosser and Kind's model, the sample of interviewees is mixed – about 70 subjects including patients, doctors and nurses, and a number of healthy people. During interviews subjects were asked first to mark each state according to the severity of each state (including death), and to express numerically comparative judgements among states – for example, how many times a person in the state  $x$  is *more ill* than a person in the state  $y$ , etc. Then, they were asked to indicate (by being given appropriate statistical information concerning the prognosis of patients) the proportion of resources to allocate for the cure of the various states.

The scores and data emerging from these calculations are to be used first of all to establish the quality of life of a patient as the result of a cure – as in the intentions of its creators as well as of its supporters (e.g., Williams [52.15] as quoted in [52.10]). For example, a patient who without treatment would continue to live for a further 20 years in state 4A in Table 52.1 would enjoy 19.28 QALY. If it is assumed that treatment  $x$  will facilitate a complete cure and add a further 10 years to the patient's life, then a further 10.72 QALY would be yielded. If we assume that an alternative treatment  $y$  would give the patient 35 years of life but with slight disability the QALY can be used to assist deciding which treatment to use [52.10, pp.32–33].

Second, the outcomes of calculations are meant to assist in deciding in what way available resources for treatment could be equally and efficiently distributed. Once the QALY value of a health-care intervention is calculated and its cost is known, it is in fact possible to calculate the cost per QALY of each intervention and provide a direct comparison between interventions (generally high priority is given to health-care activity where the cost per QALY is as low as it can be). Overall, an evaluation on the basis of QALY takes into account the costs of a range of interventions relative to the changes in terms of quality of life and to the projections of life expectancy as a consequence of the interventions themselves.

On the basis of all these considerations, QALY measurements are taken to be as much objective a tool as possible to decide, in real-life situations, how limited resources can be distributed in the community of people in need of treatment and care. However, they also raise a number of controversial issues. Some have to do with the methods of calculations of quality adjustment factors, others with the assumptions made by the underlying model of quality these measurements entail. In what follows I will focus on the latter, as the problems raised by the model's assumptions reflect some of the concerns expressed in the previous section.

### 52.1.3 Unrealistic Utility Assumptions

The utility model informing QALY measurements assumes that there are *facts* about human life that can arguably be used as criteria to decide whether a life is more worth living than another (or worth living at all). *Good health* is one of them: It is a fact that good health is a highest ranked preference among people. What the model implicitly asks us to do, then, is first of all to assume a type of *ideal life*, on the assumption that there is one that is the most preferable, in the absence of any intervening circumstances. This is the *healthy life*. Secondly, on the basis of this type of life the model asks us to choose between types of real life, where health approximates by a multitude and variety of degrees to the parameter set up by the ideal life. This second assumption (types of lives are equally comparable) becomes, for example, explicit when, once confronted by the alternative between a brief but healthy life and a long but unhealthy one we are led by the model's calculations to prefer the former. (Of course the extent to which this is true will depend on the weights.)

Now, these assumptions work for the purpose of the model (they allow to calculate QALY), but at a cost. As I said, the model seems to advocate, as a term of comparison, a picture of *ideal life* construed on the basis of measurable and comparable *facts*. However, in order for such a picture to work in the way the model requires, a series of further assumptions must be put in place that seem to make the model *overconstrained*, in the sense explained in the first section. For example, it is assumed that what matters is not just life but a trade-off between life and quality; that the just distribution of resources maximizes the thing that matters (qualys); and that a combination of real people's preferences plus more *objective* facts such as invalidity and pain is the way to decide on the quality of life (at least for the purposes, set out by the model, of deciding on the distribution of resources).

Equally, the questionnaires the interviews are formulated on assume that: Individuals' bias toward

certain kinds of diseases or handicaps are excluded (weights are assigned by the people interviewed on degrees of invalidity and pain independently of types of illness); individuals' beliefs (religious, moral, or other) concerning, for example, the absolute value of human life do not interfere with assessment, or can be set aside for the sake of policy deliberation; individuals are, or can be forced to be (by means of policy) *altruistic*, in the sense that they should not expect an action to be performed in their own interest should it prove detrimental to others (e.g., they would not undertake some medical treatment which is not widely available because some other people rather than themselves might benefit from that very treatment, according to qaly criteria, or from other treatments at equal cost, etc.). One might indeed argue that there is a difference between interpreting the model from the side of the policy that it is meant to guide or from the side of the individuals that a policy is meant to protect; and that the reasonableness of a policy, or the practical issue of how we get everybody to agree on a course of action, should be treated separately from the model. Nonetheless, in order to calculate benefit in terms of QALYs the model needs to assume that individuals choose within the constraints described above.

All this means that an overconstrained model ensures that its results follow, but only under the rather specific and often unusual conditions the model requires for reaching those results. And "[...] unrealistic assumptions that overconstrained the results are a problem for learning lessons that apply elsewhere [...]]" [52.7, p. 4].

There is a further assumption that detrimentally forces yet more constraints on the model. QALYs, as we said above, measure benefits in terms of utilities, and the utilities in QALYs' terms are the extra years of life, adjusted for quality, granted to people by allowing certain treatments. Benefit (this is the model assumption) is what the model isolates in order to study its effects without confounding factors. Obviously, when it comes to health, benefits are important, and QALYs have an important role to play in guiding how to distribute limited resources. But if the model is to focus exclusively on benefit it must introduce extra assumptions to allow a correct analysis and delivery of outcomes; and this gets in the way when we then come to *read out a moral* from the model that has an impact on real target situations outside the model.

One of these extra assumptions is the neglect of fairness. Benefitting from treatment is not the same as deserving treatment. In real-life situations individuals (healthy or not) weigh quality not just on their desire for health but on their desire to live, admittedly as long and as well as possible, and to a larger or lesser extent

independently of intervening circumstances. A focus on QALYs seems to obscure fairness: Giving priority to those who better benefit from treatment is not necessarily the fairest choice. It might make us discriminate the old for the sake of the young, for example, or more generally those in a better position to benefit from treatment [52.16, p. 196].

When we use a model to guide us in real-life situations, the results we aim at should prove to be right both in the sense of being correct and in the sense of being just (of the type: the stronger abuses the weaker – and it is bad that it does so). Instead the *moral* we read out of an overconstrained model seems to confuse what is right in terms of the models' assumptions with what is right for reasons that might go well beyond the boundaries of the model.

Such confusion might also give the impression, or the illusion, that the model's scope is wider than it is. For example, could the model be used to assist in the thorny and controversial debates on euthanasia? Indeed, QALYs are meant to offer guidance in assessing what makes a life *bearable*. The state of death is very much part of the benefit calculus. In the original matrix of Rosser and Kind's model (Table 52.1), states such as *being confined to bed with moderate pain* or *being on a wheelchair with intense pain* were marked on average by the interviewees at the same level as the state of death. States like *permanent loss of consciousness* or *being confined to bed with intense pain* received a value even inferior to that given to death. In such cases, it would appear well justified to terminate somebody's life, both from the side of policy (distribution of resources) and from that of the interest of the individual in question (bearability of one's life). Allowing an individual in that kind of pain to terminate his or her own life (or helping him to achieve this) would be based on an objective assessment of that life's quality (or lack of it).

*Barrie* [52.17, p. 1]:

"In a climate where evidence-based decisions are valued so highly, numerical measures that provide a guide to quality may be more widely used and trusted than general ethical concerns about the elderly or very sick that cannot be expressed in the same quantifiable terms."

However, social debates surrounding euthanasia teach us that decisions about the right to die, when the desire to live has ceased, can hardly be made on the basis of a concept of quality as calculated in terms of a benefit for the people whose right to die is at stake. QALYs calculations seem to assume an equivalence between the fact that some health states make somebody's life so unbearable that he/she is better off dead than

alive and the idea that actively pursuing the termination of somebody's life is more beneficial than any of those states [52.17, p. 2]. 0 utility score or negative scores indicate that death is preferable, but do not indicate that death has to be procured or induced. The fallacy in the underlying reasoning consists of making us infer from a state of fact (supposedly backed up by empirical evidence: We know how to calculate how bearable a life is) a state of principle (it is acceptable to allow somebody to die, or to terminate his/her life). If the model is to inform a decision regarding the right to die, a whole series of factors and conditions, well beyond the boundaries of the model, must be put in place. Let us see what some of them are by means of a discussion of some real cases.

In 1973, George Zygmanski was involved in a motorcycle accident near his house in New Jersey [52.18, pp. 177–178]. As a consequence, he was confined to a hospital bed, paralyzed from the neck down, with no prospect of recovery and in severe pain. This is a typical case that in terms of QALYs would receive a utility score below 0 (e.g., in Table 1 and 2 in Rosser and Kind's model above). George Zygmanski told his doctor and his brother Lester that he did not want to prolong his life in such a state. In fact, he begged them both to kill him. One day Lester went to the hospital smuggling in a pistol. He asked George once again whether he wanted his pain to be ended. George, who could not speak any longer, because of an operation performed in order to ease his breathing, nodded. Lester shot him.

Several levels of questioning become relevant in this episode. First, a factual-empirical level: Was George right in preferring death to life? This is a question that, as mentioned above, could be put in terms of QALYs. At a different level we can ask: Was George right in asking to have his life terminated? Here a QALY calculation runs short of answers for the host of questions that come to the fore. Did Lester have sufficient evidence to be able to assess objectively, in a "once and for all" manner, his brother's physical and emotional condition? Is such evidence to be used as the only deciding factor in a plea for death? Was Lester entitled to act upon his brother's request? Or was he just a murderer? Should a doctor have taken the responsibility of carrying out the termination of George's life by less extreme means (e.g., by giving George a lethal injection)? Were the actors involved in this episode violating a moral code, or any particular moral rule (e.g., it is wrong to kill) – besides violating legal rules? How do the two sets of rules interact with each other? If, for example, George Zygmanski lived, rather than in New Jersey, in the Netherlands (where doctors are by law entitled to help their patients to die), would this very fact lift moral responsibility, and bypass moral judgement for all and

each of the actors involved? These, and more, are the questions that make the debate on euthanasia so controversial.

A second real case involves an infant known to the public as *Baby Jane Doe* who was born on 11 October 1983 in New York suffering from multiple defects, including spina bifida (a broken and protruding spine), hydrocephaly (excess fluid on the brain) and microcephaly (an abnormally small brain; in this case part of the little girl's brain was actually missing, as a CAT-scan revealed). The baby's parents were told that immediate surgery on the spina bifida abnormality would have given their baby a 50–50 chance of surviving until her twenties. However, even in the case that she did survive that long, their daughter would have never have a chance to lead a normal life: She would be severely mentally retarded, physically impaired, paralyzed, epileptic, confined to bed, and in constant risk of serious diseases and infections. Without surgery, the baby would have died in 2 years.

In the face of all these, the parents chose not to let the doctors proceed with the surgery. They reasoned in intuitively QALYs terms: a 50% chance of life extension with severe disability (supposedly below the 0 score) achieved by means of risky surgical procedure is worse than a 100% chance of expected 2 years with severe disability with no surgical intervention.

The questions that can be addressed in this second case entail an even more complicated set of variables than the previous case. First, in asking whether the baby's parents were right, and justified, in making their decision, we have to take into account that the decision was made *on behalf of* their baby daughter. We should also acknowledge the fact that the parents' decision was probably made taking into account their own future – practical as well as emotional – in the upbringing of such a severely disabled individual. Secondly, in asking how and how far did the factual evidence offered by the doctors influence the parents' decision, the extent of the disability at stake added strain and uncertainty to the decision-making process. As a matter of fact, the medical profession was split over the evidence: For some doctors the baby's condition was not as hopeless as had been presented to the parents. In this uncertain scenario, Lawrence Washburn, a lawyer associated with some conservative right-to-life groups, brought the case to the Supreme Court in New York State to be reassessed, and a federal investigation was issued against the hospital to determine whether there had been discrimination against a handicapped person.

The relevant question seems here to be the following: Should everything possible be done in order to save a severely disabled life? Or else, should the severely disabled be left untreated since there is no ben-

efit in prolonging their lives? These questions cannot be appropriately addressed by resorting to QALY calculations (or in this case, DALY calculations: Disability Adjusted Life Year). Health benefits are important, as we noted above, but if the model QALYs are based on is to provide a guide to make decisions in real-social cases, issues of fairness should also be considered. If we want a moral of the type *the weaker is prey of the stronger* to mean that *the weaker should be defended against/protected from the stronger* we ought to import a story into the model (or put the model in the context of a story) such that fairness to the weaker becomes as important to the deduction of the model's outcomes as it is benefit. Prioritizing treatment on those who are most likely to benefit from it is not necessarily the fairest thing to do. This however does not necessarily mean that QALYs are unusable, it means that "they cannot entirely determine which decision is the right one," a limitation that has not always been acknowledged by QALYs' sympathizers [52.16, p. 196].

The two cases just described put us in front of real dilemmas: How can I put an end to the suffering of my paralyzed brother? Should I let my severely handicapped daughter die? Faced with the complexity of such questions, the *moral* that is supposedly written into models à la Rosser and Kind appears thin, besides being dubious. The objective conditions of a state of health combined with the benefit of treatment are not sufficient/adequate to conclude that a certain course of action is to be pursued. The *moral* of the model is not simply the logical conclusion of a deductive inference. To continue with the same example, the concept of a *bearable life* brings us to question, by adopting a vocabulary that not only includes reference to the biology of individuals, whether for a human being all we need to assess, when we consider bearability, is its QALY value.

*Quality of life* in the model is not equivalent to *quality of life* in the target system: It is a specific construct of the model that is simultaneously constrained (and over-constrained) by the model's chosen assumptions. The *world in the model* in the case of QALY perfectly exemplifies the double ontological regime that was pointed out at the beginning of this chapter (the model creates a world out of its own premises; the world outside the model is selectively targeted via the focus created by the model's assumptions).

What use can be made of models like Rosser and Kind's then? What reliable guidance do/can they offer to making objective decisions? Their contribution can be significant (once their formal framework has been checked for rigor and internal validity) provided that (1) their limited domain of application is acknowledged; (2) their contribution to decision making is not taken to be *ready-made* (a direct consequence of the *objectivity* of the cost-benefit analyses of the model). A lot of further work is needed to fill the gap between a model and the world the model is asked to help us with. A complex background of social, ethical, legal considerations is to be spelled out. Scientific considerations also enter this background – in the case of medicine, for example, by offering evidence and degrees of care that help individuals to weigh reasons for and against certain options, and by taking on some of the responsibility in suggesting certain options as the most or the least favorable.

Unrealistic assumptions in the models do not necessarily detract from the objectivity of decisions based on such models. They however both give boundaries to what we can be objective about, and force us to acknowledge the complex and controversial variety of issues brought forward by the particulars of real-life situations (as the cases described above point out) that the model intends to target [52.19].

## 52.2 Real Experiments, Not Models Please!

So far in this chapter we have showed the limits of using models in representing target systems (in our examples, real-life situations) and the costs in terms of including assumptions that make models work by becoming untrue to the systems they are meant to target. In the light of what we pointed out, would then rescuing their applicability and good use be too demanding a task? Are we, all in all, better off without them?

There is nowadays an established trend to prefer the results achieved by well conducted experiments in social science, and to trust their outcomes over those pursued by other methodologies. In particular, random-

ized control trials are by many considered the golden rule to doing good social science, namely science that produces results that are not only theoretically adequate, but reliable and useful when made available in practice (e.g., when informing a policy program).

A randomized controlled trial (RCT) is an experiment in which investigators randomly assign eligible subjects (or other units of study, e.g., classrooms, clinics, playgrounds) into groups. Each of the groups receives or does not receive one or more interventions (e.g., a particular treatment). Then the results are compared, and if the observed outcome is statistically

significant, then it can be concluded that it has indeed been caused by the experimenters' manipulation, that is, there is a high probability that the intervention actually works. Blind procedures (single, double, triple, to even quadruple) are often used to control bias.

Experiments of the type of RCTs ably answer the question *what works* that seems to be the question that most concerns the field of policy making. However, it is often the case that restricting policy making to answering this question ends up not only in policy failure, but also in an inability to understand and explain why something (e.g., an intervention, a treatment, etc.) does not work, and then possibly put it right. Here is where models can help, in particular causal models. Being able to formulate *why it works* questions give policies a better chance to work better, or to work at all. Let's see what these models consist of and – by means of a few examples – how and where they prove most useful.

### 52.2.1 Causal Models: How Can They Come to the Rescue

What is a causal model? Described in general terms, it is a representational tool that describes causal relations among a set of variables. These models suggest hypotheses about the presence of, and direction of influence between, these variables by mapping them by means of a variety of descriptive devices, such as path diagrams, flow graphs, or causal pies – to name just a few (see below for some graphic representations).

The definition of causal models, as set out above, appears though obscure if relevant bits of terminology are not explained. What counts as *causal relation*? What are the *variables* that enter such a relation? Where do the *hypotheses* that models suggest come from, or where are they grounded? Without entering too much in detail, to understand the functioning and the purpose of these models the following is, at the very least, essential.

First, what are the items we call *variables* in a causal relation? In the social sciences the relations could be kinds of situations (e.g., mothers' education and child survival, size of classes and children performance), or a type of *population* (e.g., children in a school, an audience in a theatre, the over-sixty in good health in a national statistic), or indicators (e.g., aggressive behavior, social isolation, fear of failure in the context of child abuse, or consumer price index, money supply, consumer confidence, retail trade sales in the context of the state of an economy). Variables can be depicted at different levels (individual, social, ecological) or can be described (and indeed constructed) by means of different theoretical tools, or concepts. Understand-

ing how they are depicted and/or described is important to understand and evaluate the causal relation in which they find themselves associated.

Second, a *causal relation* is not simply an association of variables. The goal of a causal model is not only to express co-variation, but to evaluate how and why co-variation occurs. This is what constitutes *the hypothesis* put forward by the model, that is a suggestion of why variable  $x$  influences variable  $y$  and how the way  $x$  and  $y$  are related contributes to a certain outcome (output  $z$ ).

So third, the hypothesis is set out to understand what makes the variables be relevant to each other and draw a structure that the model is meant to represent. The hypothesis is formulated by means of a series of assumptions grounded in some background knowledge, or some theory (when present). The better established is the background knowledge, or the better accepted the theory the model appeals to, the better grounded are the assumptions that help formulating the model's hypothesis. Neither background knowledge nor theory are guarantors that the hypothesis is correct (once and for all). There is no well-rehearsed recipe to build up a good causal model. This does not mean that there is no justification, or degrees of confidence, for how a model identifies or describes a causal relation. Typically (though not necessarily) the causal relations in these models are represented in a statistical form by means of systems of equations. However, it takes a good deal of judgment, detailed information, and awareness of local factors – some or none of which, often, available in advance, or in abstract/general terms – to achieve any degree of justification. Model building, especially in the social domain of inquiry, where theories are scanty and information routinely disputed, is itself hypothetical, and tentative.

So why are these models useful? Where and how are they most useful?

### 52.2.2 Class Size Reduction Programmes

In the 1990s there was growing concern about the poor performance of primary school children in California [52.20, 21]. A program was implemented to reduce class size in view of improving children performance. Not only it appeared plausible that children in small classes are more closely and easily looked after, and therefore learn more and better. There was also evidence that this is actually the case. The evidence was provided by the results of the STAR program, run in Tennessee starting in 1985. A 4 year RCT, involving 79 schools (around 7000 children from kindergarten to third grade) looked at how small(er) classes might lead to improved academic performance. Class size was brought down from an average of 22–25 children to an

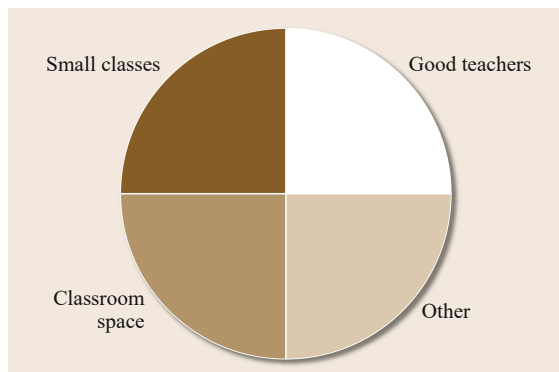
average of 13–17, and children and teachers were randomly assigned to the smaller classes. The results of the RCT established that reducing class sizes improved a number of educational outcomes in children.

California endorsed these results, and in 1996 fed them into issuing a policy of class-size reduction that, however, resulted in failure. There were a number of reasons that accounted for the failure. The reduction program was implemented over too short a period (from spring to the following autumn). California needed to hire extra teaching staff, and given the short time of implementation the quality of teachers was not carefully assessed. Smaller classes meant more space that Californian schools did not have, which also meant that space was taken away from other activities in children’s curricula (with consequent decay in other areas of children’s education, such as music, arts, physical activities).

Was there any way to rescue the California program, and the intuitively plausible idea behind class-size reduction? We might think that, in order to achieve that, more experiments were needed, over longer periods, in different places in California, or in places other than California – in view of gaining more data and better evidence. But perhaps a change in strategy would be more appropriate, and more efficient in policy terms: What if the program needed a good causal model? Let us explore what this second option involves, and why it might have provided better results [52.22, 23].

First, if we compare the Tennessee and the California programs, two factors consistently emerge that, by being present in the former context and absent in the latter, contributed to either success or failure: space and quality of teaching. These two factors can be mapped into a number of purpose-built causal models. One of them is a *causal pie* showed in Fig. 52.1.

In this model, each slice representing a factor must be present if we want to have a pie (i. e., make the pol-



**Fig. 52.1** An example of pie diagram for improved learning scores (original drawing by Alex Marcellesi)

icy – reduce class size – hit the target – improved child education). The point of the pie model might go missing if we think of factors in terms of slices (indeed, even if we *eat* one slice there is still a lot of pie left). Perhaps a better representation of factors in the pie model is in terms of ingredients: all the ingredients are equally necessary to bake our pie; should one of the ingredients be missing then we end up with no good pie, or even worse, no pie at all [52.22, p. 63]. Of course, the pie model can be extended to include more factors than the two pointed out above, with the result of showing more accurately what is required for predicting whether a policy will work.

Another way of representing how the necessary factors interact and with what results is using *causal path* models.

In this type of model arrows link factors in cause-to-effect relations that include both positive and negative outcomes (Fig. 52.2). It can also be used to represent side effects to the policy to implement (which is something that causal pies are ill equipped to account for: these can just show what it takes for a pie to be baked, independently of whether I could get a sore tummy after eating it).

The graphic resolution of these models does not prevent from representing them quantitatively (e.g., Monte Carlo simulation).

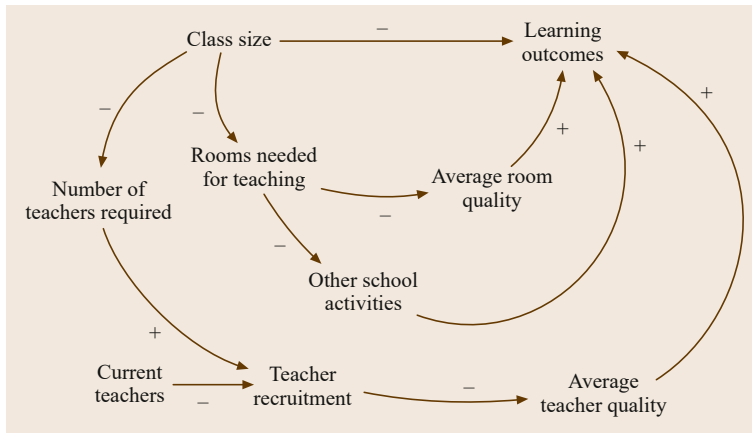
Good causal models will at the very least pick what factors are needed to get a policy to achieve (or maximize the chances to achieve) a certain outcome, and what other factors (positive or negative) might a policy bring into a purported setting of implementation, with what either beneficial or detrimental outputs. The point of using these models is to understand, in advance of implementing a policy, what is needed to increase the chances for a policy to work in specific settings or circumstances. The fact that a policy worked *there* is no evidence, or reason, that it would work *here*, let alone *anywhere* [52.22, Part II] – as the following example will clearly show.

### 52.2.3 TINP and BINP

RCTs provide strong evidence for the conclusions concerning a study population. In this sense, it is said that well-conducted RCTs are *internally valid*: on the basis of agreed upon premises certain conclusions must consistently follow for the target population. The agreed upon standards are usually identified by a number of formal requirements, which are taken to secure the consistency of results (the reasons why they obtain and why they are considered to be valid).

RCTs are often also expected to be *externally valid*: The *same intervention* adopted for the study population





**Fig. 52.2** A path diagram for improved learning scores (original drawing by David Lane)

has the *same result* when used with a new population deemed sufficiently similar. However, how can we decide whether the new population is sufficiently similar? Often apparent similarities can be misleading, and yet there is little guidance as to how to apply the findings of an RCT across contexts. Using causal models can be of some help here. Let us see how.

Children malnutrition in developing countries is, and has been for a while, a huge, pressing challenge. Globally an estimated 165 million children under 5 year of age are stunted (i. e., height-for age below), and more than 90% of them live in Africa and Asia. An estimated 101 million are underweight (weight-for-age below) [52.24].

Several international programs have been launched over the years to tackle the problem. One of them is the Bangladesh Integrated Nutrition Project (BINP), implemented in 1995 by the World Bank. It was a growth-monitoring pilot program, modeled on an acclaimed successful predecessor, the Indian Tamil Nadu Integrated Project (TINP).

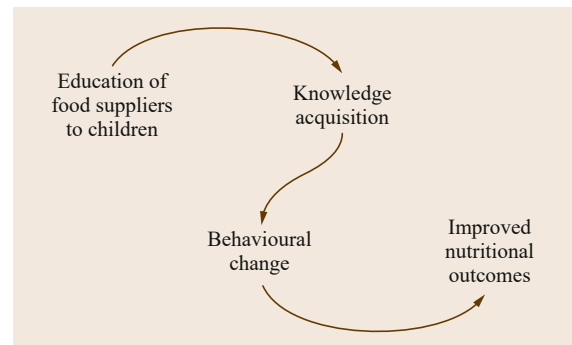
What these programs integrate are nutritional education, food supplementation and health measures, with a particular emphasis on the education aspect. They admitted children who were growing insufficiently (after being monitored at purpose-built local weighing centers) or showed evident signs of malnutrition. They also targeted the children's mothers, as well as mothers-to-be, in the belief that "ignorance, rather than poverty, was to blame for poor nutrition" [52.25, p. 5]. Both programs covered rural areas which had on average the worst nutritional records among children.

However, unlike TINP, the Bangladesh pilot program turned into failure when implemented at a national scale (and it reached almost negligible success at local scale too). After 6 years, and despite documented evidence that the educational message had reached the project population target, malnutrition did not fall at

any significant rate [52.26]. On the basis of evidence of unsatisfactory impact the World Bank discontinued the project.

Why things went wrong in Bangladesh, and not in Tamil Nadu? Were Bangladesh and Tamil Nadu not *sufficiently similar*?

To answer these question let us look at the project design in terms of a casual pathway model (of the type discussed with class-size reduction).



The causal chain, that reconstructs the rationale underlying the nutritional program, immediately reveals what went wrong in rural Bangladesh: here mothers are not the principal food providers. Men go to the market to buy food, and mothers-in-law administer food for everybody in the household.

Educated Bangladeshi mothers could not practice their newly acquired nutritional knowledge to make an impact on their children growth.

A further problem emerges if we look at the causal chain that goes from supplementary food to improved growth in children. In Bangladesh the food provided by the program was not often *supplementary*: Due to the very impoverished conditions of families, it was given *in substitution of* any food, if not ultimately given to somebody else than children (the *leakage effect*). This

shows at least two things: Even a well formulated causal chain should be open to revision and re-formulation in the light of emerging local or circumstantial factors; and the language used to categorize the intervening variables in the chain should be formulated at the appropriate level of abstraction to pick the relevant aspects that set the causal chain in motion, independently of the contextual multiple realizability of the categories themselves.

In summary, good causal models might provide the missing link between internal and external validity, and be a useful guide in moving from *there* to *everywhere* to *here*, in Cartwright-Hardie terminology. They can help us see that the *same* is not the *same* in different contexts, and understand why. They are also flexible enough to endorse the emerging dynamics of the systems they endeavour to understand.

#### 52.2.4 Children Mortality and Inflicted Death

There is one word missing from our discussion: *mechanism*. So far we have referred to causal chains, causal links, causal paths – none of which are equivalent to causal mechanisms. Mechanisms better resonate with the word *structure* that we encountered while describing the hypothesis that informs the working of a causal model. How do mechanism and structure go together [52.27, 28]?

Imagine a black box. What we call causes enter the box, and what exit the box are the effects. But what happens inside the box? What makes the causes *cause* the effects? Said plainly, mechanisms are what explain why *C* makes *E* happen. They account for what happens inside the black box. This is why they are sometimes called *generative* mechanisms. Causal models model these mechanisms – if they are to be used explanatorily to say why certain variables are related in a particular way and direction. They illustrate the rationale underlying a causal chain, or link, etc.

What about *structure*? A structure is a relatively stable arrangement of components that brings about a particular outcome. It is this type of arrangement that in philosophy of biology is often called *mechanism*. And it is this arrangement that a model tries to reproduce, in view of mapping why something happens the way it happens. A mechanism is what makes us see what a structure can achieve, what is its output. Conversely a structure is what allows a mechanism to function the way it does.

The expression *why something happens the way it happens* requires some unpacking. To say why something happens we need a mechanism able to explain why an identified structure of variables achieves an out-

come. To explain the way it happens we need, besides a mechanism, a number of further conditions (often local, or specific to circumstances) that account for the fact that something actually happens, and in the way it does (in the circumstances). Modeling a causal mechanism should then include the structure, its functioning, and the constellation of *support factors* that jointly account for the achievement of an outcome [52.22, p. 44].

Let us look at an example similar in some respects to the one discussed in the previous section [52.29, pp. 21–23, pp. 162–65]. In a seminal article Caldwell famously argued that mothers' education is [52.30, p. 408]:

“the single most significant determinant of these marked differences in child mortality. These are also affected by other socio-economic factors but no other factor has the impact of maternal education [...]”

Why? Caldwell does not only rely on empirically established statistical correlations between rates of mortality among children and the level of education of their mothers. He tries to unravel the mechanism that relates the two variables. Mothers' education is not as such a *cause* of anything, we need to explain how it comes to perform this causal role vis a vis a specific outcome (decreased mortality rates in children).

Caldwell argues that educated mothers are more capable to interact with “the modern world” (doctors and nurses) and been listened to; they become less “fatalistic” about illnesses and more pro-active (looking for alternatives in child care); and their education impacts on “the traditional balance of familial relationships with profound effects on child care” [52.30, pp. 409–410]. These are reasons that we would certainly expect to find in a good causal model explaining the role of education in child mortality decline, but on their own they do not constitute a *model* of why this is the case. To build a model, what is also needed is good knowledge of the causal context where those reasons make sense as contributing factors, and a good story (= a meaningful hypothesis) of how they interact and jointly make it believable that education can perform such a causal role. In the process of building such a model, we might well realize first that maternal education requires a host of further factors (socioeconomic, environmental, biological, and ecological) to make the mechanism effective, besides being believable as a model. Second, it also requires that the consequences of maternal education on decreased child mortality (better interaction with modern world, less fatalism about illnesses, beneficial changes in familiar relations) identify a stable enough arrangement of factors that can explain why in a certain

context maternal education is a believable/justifiable cause of decreased child mortality. Building causal models is no doubt complex, building good ones unforgettingly demanding!

And yet, building good causal models proves essential not only for social science knowledge practice, but also for its impact on policy making and social intervention. To give an example of such use and usefulness we can take a final look at the following area of application.

Inquiries in child abuse deaths normally take the approach of blaming human error or fault [52.31]. However these enquiries have not led to any substantial improvement in professional practice. This, according to Munro, is due to the fact that inquiries into inflicted deaths on children take human error not as the starting point, but as a conclusion. However, errors often occur within systems that are ill-equipped for preventing them. A *system approach* is then a more appropriate line of enquiry. This consists in tracing the causal chain that led to a death not so much back to the individual, but to the structural failure of the system that explains why the individual committed the error. This is not to abrogate responsibility but to understand it better, both at individual and system level, and to act on those aspects of the system that allow for mistakes to emerge.

Looking deeply into the causes of error entails identifying the interacting factors, the organizational context and the available resources that made an outcome possible: in other words, it entails an understanding of causes in the context of the social structure and underlying mechanism where individuals operate.

In adopting a system approach “the focus is on the interaction of the different layers of the system so that a more vivid picture is drawn of how the particular case fits into its context” [52.31, p. 12]. The *vivid picture* is what, using a different language, a *model* can provide by mapping the underlying structure of the case investigated and by means of it explaining its causal mechanism.

Modeling a case according to this approach (by means of whatever methods appear appropriate in mapping the context) can then prove to be a crucial help not

only in understanding what happened, but also in trying to prevent it from happening again.

Eight-year-old Victoria Climbié’s death in 2000 is an instructive case. Victoria was tortured and killed by her guardians, suffering 128 injuries during months of abuse and neglect. She was repeatedly hit with shoes, coat hangers, a hammer, a bike chain. In her last days she was made to sleep in winter in an unheated bathroom “bound hand and foot inside a bin bag, lying in her own urine and faeces” [52.32, p. 13]. In the end Victoria died of lung, heart, and kidney failure. During some of the period of her abuse she was seen by four local authorities’ social workers, two protection teams from the Metropolitan Police, a specialist from NSPCC (National Society for the Prevention of Cruelty to Children) and staff of two hospitals where she was admitted as a consequence of severe and suspicious injuries.

What went wrong, it is asked in the Report into her death? “A lack of good practice” is the answer, and a consequent breakdown of the entire system [52.32, p. 15]. Therefore the report’s recommendations are “clear accountability” (who is responsible, especially at the level of senior and experienced managers) and closer monitoring of compliance with principles and guidelines of good practice. Would these measures, had they been in place, “*detected* the poor quality of the service offered to Victoria?” This question, argues Munro, cannot be answered appropriately (and indeed it is not addressed as such in the report) without a “clearer understanding of the factors influencing the poor quality work” of those who were responsible for the failings in this case [52.31, p. 14]. Munro [52.31]:

“Until we understand why those errors looked the reasonable thing to do to the individuals at the time, we cannot devise solutions that ensure that, in the future, they will be more likely to opt for the right course of action.”

If devising good causal models that help us detecting *why something went wrong* (not only *what went wrong*) and designing safer systems of protection might even minimally contribute to prevent the death of a child, this seems a task well worth pursuing.

## 52.3 Conclusions

If models are to guide us in making good decisions, or implementing good social interventions, how can we handle the unrealism of the assumptions on which some of them are based? If models are imaginary structures from which we cannot immediately derive a *moral* concerning their target systems, how much supplementary work (and of what type) is needed on them to become an

actual help in learning about real situations? Do models constitute a good methodology of social intervention or policy, or are experiments better in providing socially useful information? Should experimental methodology be found lacking in some respects, could models (at least some types of models) be used, and be useful, to tell us where experiments go wrong? The examples pre-

sented in this chapter offered some suggestions as to how to answer these questions, and handle some of the controversial theoretical issues that emerge when models are used in social practice. The overall aim was to show that, if we learn how to use models and what to expect from their use, they are a real asset both for social scientific research and for social policy.

**Acknowledgments.** Nancy Cartwright, Eileen Munro, and Jeremy Hardie have provided most of the ex-

amples and case studies presented in this chapter, and which have been part of discussion in the research project *Evidence for Use* (Centre for Philosophy of Natural and Social Science, London School of Economics and Political Science).

The quality of life example is discussed in E. Montuschi, *Oggettività ed Evidenza Scientifica* (Carocci, Roma 2001), Chapter 5.

I am grateful to Federica Russo for reading and advising on the pre-final draft of this chapter.

## References

- 52.1 R. Frigg: Models and fictions, *Synthese* **172**(2), 251–268 (2010)
- 52.2 M.S. Morgan, M. Morrison: *Models as Mediators* (Cambridge Univ. Press, Cambridge 1999)
- 52.3 R. Giere: *Scientific Perspectivism* (University of Chicago Press, Chicago 2006)
- 52.4 U. Mäki: On the method of isolation in economics. In: *Idealization IV: Intelligibility in Science*, Poznań Studies in the Philosophy of the Sciences and the Humanities, ed. by C. Dilworth (Rodopi, Amsterdam 1992)
- 52.5 S. French: Models. In: *Encyclopedia of Philosophy and the Social Sciences*, ed. by B. Kaldis (Sage, Thousand Oaks 2013)
- 52.6 M.S. Morgan: *The World in the Model: How Economists Work and Think* (Cambridge Univ. Press, Cambridge 2012)
- 52.7 N. Cartwright: Models: Parables v fables. In: *Beyond Mimesis and Convention. Representation In Art and Science*, Boston Studies in the Philosophy of Science, Vol. 262, ed. by R. Frigg, M. Hunter (Springer, Dordrecht 2010) pp. 19–31
- 52.8 C.A. Pissarides: Loss of skill during unemployment and the persistence of employment shocks, *Q. J. Econ.* **107**(4), 1371–1391 (1992)
- 52.9 N. Cartwright: *The Dappled World* (Cambridge Univ. Press, Cambridge 1999)
- 52.10 A. Edgar: Measuring the quality of life. In: *Facing Death*, ed. by P. Badham, P. Ballard (University of Wales Press, Cardiff 1996)
- 52.11 E. Montuschi: Modelling objective decisions. In: *Objectivity and Scientific Evidence*, ed. by E. Montuschi (Carocci, Rome 2011), Chap. 5, in Italian
- 52.12 A. Williams: The value of QALYs, *Health Soc. Serv. J.* **3**, 3–5 (1985)
- 52.13 UK Clinical Ethics Network: *Ethical Considerations, Maximizing Welfare/Benefit* [http://www.ukcen.net/index.php/ethical\\_issues/resource\\_allocation/ethical\\_considerations3](http://www.ukcen.net/index.php/ethical_issues/resource_allocation/ethical_considerations3) (2011)
- 52.14 R. Rosser, P. Kind: A scale of valuations of states of illness: Is there a social consensus?, *Intern. J. Epidemiol.* **7**(4), 347–358 (1978)
- 52.15 A. Williams: Economics, QALY and medical ethics: A health economics perspective, *Health Care Anal.* **3**(3), 221–226 (1995)
- 52.16 J. Broome: *Ethics Out of Economics* (Cambridge Univ. Press, Cambridge 1999)
- 52.17 S. Barrie: QALYs, euthanasia and the puzzle of death, *J. Med. Ethics* (2014), doi:[10.1136/medethics-2014-102060](https://doi.org/10.1136/medethics-2014-102060)
- 52.18 P. Singer: *Practical Ethics*, 2nd edn. (Cambridge Univ. Press, Cambridge 1993)
- 52.19 E. Montuschi: *The Objects of Social Science* (Continuum press, London, New York 2003)
- 52.20 F. Mosteller: The tennessee study of class-size in the early school grades, *Crit. Issues Child. Youth* **5**(2), 113–127 (1995)
- 52.21 C. Thompson: *How to Tell when Efficacy will NOT Translate into Effectiveness*, Contingency and Dissent in Science, Technical Report 03/09. (Centre for the Philosophy of Natural and Social Science, London 2009)
- 52.22 N. Cartwright, J. Hardie: *Evidence-Based Policy, A Practical Guide to Doing it Better*, Part II. (Oxford Univ. Press, Oxford 2012)
- 52.23 N. Cartwright: Will your policy work? Experiments vs. models, unpublished manuscript (2013)
- 52.24 UNICEF-WHO-The World Bank: *Levels and Trend in Child Malnutrition* (New York 2012) [http://www.who.int/nutgrowthdb/jme\\_unicef\\_who\\_wb.pdf](http://www.who.int/nutgrowthdb/jme_unicef_who_wb.pdf)
- 52.25 H. White: Theory-based impact evaluation. Principles and practice, *J. Dev. Eff.* **1**(3), 271–284 (2009)
- 52.26 Save the Children: *Thin on the Ground. Questioning the evidence behind World-Bank funded community nutrition projects in Bangladesh, Ethiopia and Uganda* (Save the Children, London 2003)
- 52.27 D. Little: Causal mechanisms. In: *The Sage Encyclopedia of Social Science Research Methods*, Vol. 1, ed. by M.S. Lewis-Beck, A. Bryman, T.F. Liao (Sage, Thousand Oaks 2004)
- 52.28 P. Hedstrom, R. Swedberg: *Social Mechanisms: An Analytic Approach to Social Theory* (Cambridge University Press, Cambridge 1999)
- 52.29 F. Russo: *Causality and Causal Modelling in the Social Sciences. Measuring Variations* (Springer, New York 2009)
- 52.30 J. Caldwell: Education as a factor in mortality decline: An examination of Nigerian data, *Popul. Stud.* **3**(3), 395–413 (1979)
- 52.31 E. Munro: A system approach to investigating child abuse deaths, *Br. J. Soc. Work* **35**(4), 531–546 (2005)
- 52.32 Department of Health: *The Victoria Climbié Inquiry* (Department of Health, London 2003)

# Models and

## 53. Models and Moral Deliberation

Cameron Shelley

It is clear that models embody or encode information about moral values and moral conduct that is frequently important in moral deliberation, that is, the process of solving moral problems. However, there is a diversity of views on how models perform this function. In part, this diversity is due to the well-known diversity of views on the concept of model itself. Naturally, scholars with different views of what a model is produce different accounts of their place in moral deliberation. As a result, the shared involvement of models in these accounts has been largely unnoticed. The purpose of this chapter is to review the main, varying accounts of models and model-based reasoning in moral deliberation. These accounts include models as rules, as mental models, schemata, analogies, empathy, and role models. These accounts emphasize different aspects of moral deliberation. Rule-based accounts tend to emphasize morally generalized information concentrated in a set of rules and a cognitive style based on calculation. Other accounts, such as analogies, empathy and role models, tend to emphasize morally particular

53.1	<b>Rules</b> .....	1118
53.2	<b>Mental Models</b> .....	1119
53.3	<b>Schemata</b> .....	1121
53.4	<b>Analogy</b> .....	1122
53.5	<b>Empathy</b> .....	1124
53.6	<b>Role Models</b> .....	1125
53.7	<b>Discussion</b> .....	1126
	<b>References</b> .....	1127

information spread out throughout a large set of source analogs, and reflect the emotional aspects of moral deliberation. Most accounts concentrate on information originating with the deliberators, although role models, conversely, emphasize models that originate outside the deliberators themselves. Hopefully, this chapter invites further work on the relationships among the accounts reviewed within.

Models play an important role in moral deliberation. For example, someone with a sick friend lacking health insurance may consider stealing drugs in order to help out. Reasoning with models may play a role in this deliberation: Would it not be right to steal a loaf of bread if the friend were starving? The latter situation serves as a model for the former in this case.

The purpose of this chapter is to review recent literature on the topic of models in moral deliberation. This objective immediately raises two concerns. First, the term *model* is quite flexible, that is, many things have been counted as models. Thus, the relevant literature includes approaches that bear only a kind of family resemblance to one another; cf. [53.1] on *moral hypotheses*. The approach in this chapter is to accept

this situation at face value and not attempt to restrict the subject matter beyond this accepted practice.

The second concern is what constitutes a *moral deliberation*. Here, the approach is more strict. A moral deliberation is a cognitive process whereby an individual attempts to solve a moral problem. A moral problem is a decision problem in which at least some options for action are morally permissible whereas others are not. The aim of deliberation is to identify the option that is morally preferable.

One effect of this characterization of moral deliberation is that it excludes moral theorizing, that is, reasoning about moral principles or moral values as such. The focus of this chapter is instead on models in individual decision-making. Of course, people may

reason about moral principles in the process of deliberation, but that is a complication that will be set aside here.

Another effect is that thought experiments are also excluded. Many well-known thought experiments, e.g., the trolley problem, rely on models. However, since

these experiments primarily concern the nature or content of moral intuitions or principles, they are not considered here.

The sections below deal with accounts of models in moral reasoning based on rules, schemata, mental models, analogies, empathy, and role models.

## 53.1 Rules

The view that moral deliberation is grounded in rule following is a significant part of modern ethical philosophy. *Immanuel Kant* [53.2], for example, argued that moral behavior requires people to form categorical rules of action and then abide by them. His famous statement of this view is, “Act only according to that maxim by which you can at the same time will that it should become a universal law.”

One of Kant’s examples of this process is a person who sets out to obtain money through a false promise of repayment. To test the morality of this intention, it is first framed as a maxim: if I falsely promise repayment, then I can get the money I want. Next, the maxim is reframed as a universal rule by removing the subjective element: If anyone falsely promises repayment, then they can get money.

Unfortunately, this rule cannot be willed as a universal law, that is, a rule that everyone may follow. If everyone felt free to obtain money by false promises, then no one would willingly lend money since they could not trust the promises of others who follow the same rule. In effect, a world in which everyone feels free to make false promises is a world where promising achieves nothing [53.3].

Kant does not argue that people ordinarily reason in this way when faced with moral problems. Instead, it is a recommendation about how people can distinguish moral intentions from others. However, it is clear that rule following is the foundation of moral deliberation in his view.

*John Stuart Mill* [53.4] also held that ordinary moral deliberation is comprised of rule following. As is well known, Mill held that actions are moral exactly when they conform to the Greatest Happiness Principle: “actions are right in proportion as they tend to promote happiness, wrong as they tend to produce the reverse of happiness.” Mill claims that this principle distinguishes right actions from others but does not hold that people actually reason in such terms when faced with moral problems.

However, he argues that people think in terms of *secondary principles* that aid them in making moral decisions. These he describes as *intermediate gener-*

*alizations* and as working in moral reasoning as landmarks and guideposts do in navigation. The difficulty of making calculations directly on the basis of the Greatest Happiness principle means that people need ready-made rules to follow in ordinary matters of conduct. It is only when such secondary principles lead to contradictions or other difficulties that the primary principle becomes crucial for practical decision making. Mill does not furnish any detailed examples but does continue on to cite common feelings about moral problems. For example, in the matter of justice, he says that, “it is mostly considered unjust to deprive any one of his personal liberty, his property, or any other thing which belongs to him by law” [53.4, Chap. 5]. This rule seems to be an example of a secondary principle that people apply in moral deliberations. Mill then explores its relationship to the primary principle.

Like Mill, *W. D. Ross* [53.5] held that moral deliberation consists in the following rules. Unlike Mill, Ross did not agree that there is a primary principle, such as the principle of Greatest Happiness, that reconciles conflicts among rules. Instead, Ross argued that people follow a set of what he called *prima facie duties* or rules of right conduct. Although he did not furnish an exhaustive set of these rules, he identified seven important ones:

1. Fidelity: People should keep promises and be honest.
2. Reparation: People should make amends for wrongs done to others.
3. Gratitude: People should repay others who have done them good turns.
4. Nonmaleficence: People should not harm others either physically or psychologically.
5. Beneficence: People should help others improve their well-being.
6. Self-improvement: People should improve their own well-being.
7. Justice: People should be fair and distribute benefits and burdens equitably.

In some moral decisions, these rules can be applied straightforwardly. For example, the decision to bribe

a police officer to fix a ticket can be considered to be a contravention of the duty to justice as it seeks to make the fixer an exception to traffic law.

In other situations, however, different duties will suggest incompatible courses of action. Consider an episode of the TV show *Seinfeld* where Jerry and his friends are shown a baby they consider ugly [53.6]. Although the duty of fidelity suggests that they should admit that the baby is ugly, the duty of nonmaleficence suggests that they should say that the baby is cute so as to avoid offending the happy new parents. Jerry and his friends decide that it is a *must-lie* situation in which the duty of nonmaleficence outweighs the duty of fidelity.

Ross' ideas about prima facie duties have been developed in the form of defeasible deontic logics. A deontic logic is a formal logic of permissions and obligations. A defeasible logic is one in which arguments proceed nonmonotonically, that is, where conclusions may be withdrawn or rejected in the face of new considerations.

For example, *Asher* and *Bonevac* [53.7] develop such a logic precisely in order to account for moral dilemmas where different prima facie duties entail contradictory actions. They consider an action from the *Odyssey* in which Odysseus orders Neoptolemus to trick Philoctetes out of a bow that the Greeks need to obtain victory against the Trojans. Neoptolemus obeys but regrets his action and later returns the bow to Philoctetes. Neoptolemus' reasoning could be understood as follows:

1. Neoptolemus has a prima facie obligation to obey Odysseus' order.
2. The order implies that Neoptolemus has an obligation to trick Philoctetes.
3. Neoptolemus tricks Philoctetes and takes the bow.
4. Neoptolemus has a prima facie obligation of fidelity, which implies that he has an obligation not to trick Philoctetes.
5. Odysseus' order is unjust, which implies that Neoptolemus has done an injustice to Philoctetes in obeying it.
6. In light of this second obligation and his unjust act, Neoptolemus has an obligation of reparation to Philoctetes, which implies that he should return the bow.
7. Neoptolemus returns the bow to Philoctetes.

The defeasibility of obligations consists of how conclusions about them (e.g., 2 above) can be withdrawn in the light of further evidence (e.g., 4 and 5 above) [53.8]. In other words, it is the arguments for each obligation that are defeasible, rather than the obligations themselves.

These deontic logics were not developed as cognitive models. Nevertheless, they clearly suggest how rule following might be employed as an account of moral deliberation. This impression is reinforced by the fact that such logics have been developed in the Artificial Intelligence community as accounts of moral reasoning for agents in general [53.9, 10].

## 53.2 Mental Models

A related tradition concerning reasoning about permissions and obligations is centered on mental models [53.11]. This research concerns how people understand and reason about states of the kind *If A, then C*, where *A* and *C* are about things that people ought to be doing or ought not to be doing. An overview of this research is provided by *Byrne* [53.12], which is summarized here.

A fundamental result in this literature is the *Wason selection task* [53.13]. In this task, people are given a rule in an if-then form, such as *If a card has a vowel on one side, then it has an even number on the other side*. They are then given four cards from a deck with numbers on one side of each card and letters on the other, e.g., *A, B, 2, 3*. The task is to select the cards that would indicate that the rule is violated. The correct answer is cards *A* and *3*. Notoriously, most people select the *A* but not the *3*.

Performance on this task changes distinctly when the if-then rule concerns permissions and obligations. Consider a rule such as *If Paul rides a motorbike, then he must wear a helmet* where *must* is understood as a legal obligation. In the analogous selection task, people are asked to select violations of this rule from the possible cases:

1. Paul rode a motorbike.
2. Paul did not ride a motorbike.
3. Paul wore a helmet.
4. Paul did not wear a helmet.

In this version, most people select both 1 and 4.

Since the two tasks have the same structure, the difference in performance demands explanation. On the mental models view [53.14], the explanation is that people tend to represent the tasks differently in their

cognition. In the abstract case, people consider only cases allowed by the rule. In the example above, that would be cards with vowels marked on one side and even numbers on the other side. Thus, the possibility of cards with vowels on one side but odd numbers on the other side is not considered.

In the deontic case, people consider both cases permitted by the rule and cases forbidden by the rule. In the motorbike example, people consider the case in which Paul rides a motorbike and wears a helmet, and also the case in which Paul rides a motorbike and does not wear a helmet. With this representation in mind, it is easier for people to realize the relevance of the condition in which Paul did not wear a helmet to violations of the rule.

A similar disparity arises in the application of the deductive argument forms *modus ponens* and *modus tollens*. These forms of argument can be schematized as follows:

- *Modus ponens*:

$$\frac{\text{If } A, \text{ then } B \\ A}{\text{Therefore, } B}.$$

- *Modus tollens*:

$$\frac{\text{If } A, \text{ then } B \\ \text{Not-}B}{\text{Therefore, not-}A}.$$

Both argument forms are deductions involving if-then rules. However, as with the Wason selection task, people's performance in applying these rules is quite different in assertive and deontic cases.

In the assertive case, people are given a rule such as *If Nancy rides a motorbike, then she goes to the mountains*, and a statement *Nancy rides a motorbike*, and are asked to identify what follows. Most people correctly infer that Nancy goes to the mountains. In other words, they apply a *modus ponens* argument. However, when given the statement *Nancy does not go to the mountains*, most people say nothing follows when it actually follows that *Nancy does not ride a motorbike*. In other words, they do not apply a *modus tollens* argument.

In the deontic case, however, people readily apply both forms of argument. When given the rule *If Paul rides a motorbike, then he must wear a helmet* and the statement *Paul rides a motorbike*, most people infer that *Paul wears a helmet*. When given the statement that *Paul does not wear a helmet*, most people say that it follows that *Paul did not ride a motorbike*. By the same token, when given the statement *Paul rides a motorbike*

*and does not wear a helmet*, most people agree that Paul broke the rule.

As before, the explanation for this disparity between assertive and deontic cases concerns mental models. In the assertive case, people think explicitly only about the case allowed by the rule whereas, in the deontic condition, they think explicitly about both the permitted and the forbidden conditions.

Sometimes, if-then rules can convey *counterfactual obligations*. For example, if-then rules can convey obligations that do not actually hold. Consider the rule, *If Mary had lived in the 1700s, then she would have had to have been chaperoned*. This rule concerns Mary, who did not live in the 1700s, and an obligation concerning her that would have applied the event that she did.

As Byrne [53.12, p. 84] points out, the mental model of counterfactual obligations is richer even than indicative ones. In an indicative, deontic rule, the permitted and forbidden cases are represented explicitly, e.g., *If your parents are elderly, then you have to care for them*:

- Your parents are elderly and you care for them (permitted).
- Your parents are elderly and you do not care for them (forbidden).

In a counterfactual, deontic rule, the permitted and forbidden cases are represented, but the additional fact about Mary is represented as well, e.g., *If Mary had lived in the 1700s, then she would have had to have been chaperoned*:

- Mary lived in the 1700s and she was chaperoned (permitted).
- Mary lived in the 1700s and she was not chaperoned (forbidden).
- Mary does not live in the 1700s and she is not chaperoned (factual).

Since the rule is understood as counterfactual, it prompts people to represent the factual condition explicitly.

The mental models account of moral reasoning is similar to the rule-based account in the obvious sense that both hold moral knowledge to be encoded in the form of if-then rules with deontic content. Reasoning with moral knowledge is equated with the construction of arguments in which these rules are obeyed or violated.

An important difference between the two is that the rule-based account provides a mechanism for the conduct of moral decision-making, that is, following the rules. The mental models account is restricted to explanation of how people understand deontic, if-then rules. It does not explicate how people might employ rules in the course of moral deliberations.



### 53.3 Schemata

Johnson [53.15] allows that some moral deliberation takes the form of rule following. However, he disputes that rule following comprises an accurate or adequate account of the phenomenon. Instead, he argues that imagination of a certain sort is applied when people interpret and explore morally problematic situations. More specifically, he argues that people apply schemata derived from bodily metaphors when engaged in moral deliberation.

Moral schemata come in families circumscribed by the metaphors on which they rest. One such family rests on the image MORAL INTERACTIONS ARE COMMODITY TRANSACTIONS, or *moral accounting*. The underlying metaphor is of transactions in the moral domain being modeled on financial transactions. In financial transactions, people exchange goods and money in order to build up wealth. In the moral domain, people exchange actions that enhance or detract from well-being. On this understanding, having a duty to someone is like owing that person a debt. Having a right is like a debt that someone owes you (i. e., a credit). This manner of moral reasoning turns up in expressions such as, *Larry owed him a debt of gratitude*, and *Thomas paid for his mistake*.

The projection of financial accounting with moral accounting is summarized in Table 53.1 [53.15, p. 42].

The moral accounting metaphor structures a conceptual system of rights and duties that inform moral decision-making.

Now, in order to make this system operative, a means for keeping accounts must be added. To accomplish this, Johnson proposes that accounting concepts are added to the financial domain to comprise what he calls the *commodity transaction* domain. This domain includes concepts of valuing actions and comparing evaluations. Such a system allows people to calculate that when providing a creditor with eight sheep it is appropriate to discharge a debt of one cow, for example.

**Table 53.1** Projection of financial transactions as moral actions

Financial domain	Moral domain
Wealth	Well-being
Getting money	Achieving a purpose
Earning money	Achieving a purpose by honest toil
Payment	Actions that increase well-being
Debts	Duties
Letters of credit	Rights
Debtor	Person with a duty
Creditor	Person with IOU
Inexhaustible credit	Inalienable rights
Contract	Exchange of rights

Commodity transaction concepts are added to the moral domain to comprise the *moral transaction domain*. The augmented domain incorporates concepts of a balance of rights and duties and transactional justice.

The projection of the commodity transaction domain to the moral transaction domain is summarized in Table 53.2.

The concept of transactional justice, grounded in the commodity transaction metaphor, allows people to weigh rights and duties, and to discriminate just outcomes from unjust ones.

With this conceptual grounding in view, people construct schemata that they may apply to make moral judgments in particular situations. Each schema represents a set of conditions that are met in a situation that motivates a decision on how to act morally. Johnson [53.15, pp. 47–49] elaborates five of these schemata as follows. (The expressions *something good* and *something bad* refer to things of positive and negative utility or worth.):

1. *Reciprocation: One good turn deserves another:*
  - Event: *A* gives something good to *B*.
  - Judgment: *B* owes something good to *A*.
  - Expectation: *B* should give something good to *A*.
  - Moral inferences: *B* has a duty to give something good to *A*. *A* has a right to receive something good from *B*.
  - Commercial inference: *B* pays *A* for getting something good by giving something good of equal price.
  - Example: *You have been so good to me, how can I repay you?*

**Table 53.2** Projection of commodity transactions to moral transactions

Commodity transaction	Moral transaction
Commodities	Actions, states
Utility of commodities	Moral worth of actions or states
Wealth, money	Well-being
Accumulation of wealth	Increase in well-being
Profitable	Moral
Unprofitable	Immoral
Giving/taking wealth	Performing moral/immoral actions
Account of transactions	Moral account
Balance of accounts	Moral balance of actions
Debt	Owing increase in well-being to others
Credit	Others owe increase in well-being to you
Fair exchange/payment	Justice

2. *Retribution: You will get what's coming to you:*
  - Event: *A* gives something bad to *B*.
  - Judgment: *B* owes something bad to *A*.
  - Expectation: *B* should give something bad to *A*.
  - Moral inference: *B* has the right to give something bad to *A*. *A* has a duty to receive something bad from *B*.
  - Commercial inference: *B* pays *A* back for receiving something bad by giving *A* something bad.
  - Example: *I will pay you back for that insult!*
3. *Restitution: I will make it up to you:*
  - Event: *A* gives something bad to *B*.
  - Judgment: *A* owes something good to *B*.
  - Expectation: *A* should give something good to *B*.
  - Moral inferences: *A* has a duty to give something good to *B*. *B* has the right to receive something good from *A*.
  - Commercial inference: *A* pays *B* by giving something good.
  - Example: *You owe me an apology for your rudeness!*
4. *Revenge: An eye for an eye, a tooth for a tooth:*
  - Event: *A* gives something bad to *B*. *A* will not give something good to *B*.
  - Judgment: *A* owes something bad to *B*.
  - Expectation: *B* should take something good from *A*.
  - Moral inferences: *A* has a duty to give something good to *B*. *B* has a right to receive something good from *A*.
5. *Altruism/charity: What a saint!:*
  - Commercial inference: *B* exacts payment from *A*.
  - Example: *I will make you pay for what you did!*
  - Event: *A* gives something good to *B*. *B* cannot give something good to *A* in return.
  - Judgment: *B* owes something good to *A*.
  - Expectation: *B* does not give something good to *A*. *A* accumulates moral credit even without a debtor.
  - Moral inferences: *A* has no duty to give something good to *B*. *B* has no right to receive something good from *A*.
  - Commercial inference: *B* receives something good from *A* without incurring a debt.
  - Example: *That was a selfless act.*

The charity schema is different because it results in an accounting imbalance. However, the donor may hope for rewards from parties other than the recipient of the charity, such as an enhanced reputation or good karma.

The moral accounting metaphor is not the only important metaphorical basis for moral deliberation. There is also what Johnson calls the EVENT STRUCTURE metaphor. On this metaphor, achieving a goal is like motion along a path. A right is like a right of way, a path that presents no obstacles. A duty is a requirement to cede right of way to others [53.15, pp. 42–43]. This metaphor also informs a set of schemata that people employ to make moral decisions. Several further moral schemata are discussed in [53.16].

## 53.4 Analogy

Perhaps the clearest application of models to moral deliberations comes with analogical or case-based reasoning. In such reasoning, instances of previously considered moral problems are applied to current ones. Such reasoning has long been a subject of psychological research [53.17].

For example, consider a moral analogy used during deliberations over the Cuban Missile Crisis in 1962 within the Kennedy administration [53.18]. Senior officials in the administration considered several ways to respond to the arrival of Soviet nuclear missiles in Cuba. One such option was to launch a surprise attack on Cuba in order to destroy the missiles before they could be used. Although initially favored by Robert Kennedy, the option was abandoned, in part, because of an analogical argument made by CIA Director John McCone and Under Secretary of State George

Ball. They argued that a surprise attack on Cuba would be morally akin to the surprise attack on Pearl Harbor by Japanese forces that brought the USA into the Second World War. The latter was admitted by all to be a saliently immoral act and the analogy helped to change the minds of several hawks in the administration, including Robert Kennedy.

In the analogy, the known or source (or base) analog is the Japanese attack on Pearl Harbor. The problematic or target analog is a surprise attack by American forces on Cuba. There are clear similarities between the two cases, including that both involve attacks without a declaration of war and both concern nations that are already unfriendly towards each other. Of course, there are notable differences. For example, Pearl Harbor was a sneak attack by an authoritarian regime against a democratic one, whereas a sneak attack by the USA

against Cuba would be the reverse. Also, some similarities appear unimportant, such as the fact that both attacks would be launched against islands.

Psychological research sheds some light on this situation. Analogical reasoning in general typically involves at least two processes, selection of source analogs and alignment of source and target analogs. The selection process is dominated by overt similarities. In this case, that would include the sneaky and the military nature of the attacks, and the fact that both attacks would be launched against islands. The alignment process is dominated by the construction of systemic mappings between source and target analogs [53.19]. The alignment between the two analogs could be captured roughly as in Table 53.3.

The analogy is represented as follows [53.20]. The top three rows are *attribute mappings*, that is, simple items from the source and target analogs that are aligned in the analogy. The middle three rows are *relational mappings*, that is, relations between the attributes of each analog that are aligned in the analogy. The bottom two rows are *system mappings*, that is, relations between the relations of each analog that align in the analogy.

The system mappings structure the entire analogy and capture the thrust of the comparison. The first system mapping suggests that Japan surprised the USA at Pearl Harbor in order to carry out its attack. Likewise, the USA would surprise Cuba in order to carry out its attack. The second system mapping suggests that for Japan to surprise the USA implies that Japan stands for the values of a dictatorship. Likewise, for the USA to surprise Cuba implies that the USA stands for the values of a dictatorship. This representation roughly captures the sentiment of George Ball who argued that a surprise attack [53.18, p. 63]:

“would cut directly athwart everything we have stood for in our national history, and condemn us as hypocrites in the opinion of the world.”

**Table 53.3** Analogy between Pearl Harbor attack and attack by USA on Cuba. The top three rows are attribute mappings, the middle three relational mappings, and the bottom two system mappings

Source (Pearl Harbor)	Target (Cuba)
Japan	USA
USA	Cuba
dictatorship	dictatorship
attack(Japan,USA)	attack(USA,Cuba)
surprise(Japan,USA)	surprise(USA,Cuba)
stand-for(Japan,dictatorship)	stand-for(USA,dictatorship)
in-order-to(surprise,attack)	in-order-to(surprise,attack)
imply(surprise,stand-for)	imply(surprise,stand-for)

As noted above, people are more likely to be influenced by an analogy if that analogy exhibits *systematicity* or *structural coherence*. This implies that the attributes and relations of each analog are included in, and structured by, the system predicates, and that the mappings are one-to-one and between similar items.

The influence of emotions in analogical deliberations is also illustrated in the Pearl Harbor example. *Tierney* [53.18] notes that the surprise attack on Pearl Harbor had become a canonical example in American culture of criminal conduct in warfare. Thus, the idea that the USA should engage in similar behavior was somewhat shameful. This aspect of the analogy reflects the element of hypocrisy that Ball refers to in the USA taking such an action. At a more personal level, the notion of imitating the actions of the Japanese high command was distasteful to President Kennedy. He had fought against the Japanese in the Second World War and regarded the prospect of another such conflict with displeasure. Moreover, the prospect of playing the role of *Tojo* was loathsome to him.

On the theory of emotional coherence [53.21], such emotional valences play an important role in deliberation. The negative emotional valence attached to surprise military attacks in the Pearl Harbor case would attach to the surprise-attack-on-Cuba scenario as well. That negative valence would, in turn, prompt decision makers to view it negatively as well.

The role of analogies in the larger picture of moral deliberation has long been a matter of debate. In the past, analogies have been viewed as a positive and indispensable part of moral deliberation. More recently, analogies have largely been viewed as inferior to rule-based deliberations because of their relative lack of generality and thus have been demoted to a secondary role [53.22]. Cognitive models of analogical deliberation place it in a variety of relationships with other forms of deliberative reasoning.

For example, *Dehghani* et al. [53.23] place analogies in a subordinate role relative to rule-based reasoning in their Moral Decision-Making (MoralDM) simulation. MoralDM contains two modules for deliberating about moral problems. The first is a first-principles reasoning module that implements a qualitative, utilitarian calculus. In short, it evaluates potential actions or nonactions based on its evaluation of the utility of their outcomes. The second is an analogical reasoning module that constructs analogies, if possible, from a knowledge base, analogies that are systematic and deal with a matter of similar magnitude to the problem at hand. The analogy module allows the system to make evaluations based on nonutilitarian grounds and to provide solutions to problems that the rule-based module is unable to solve.

Another simulation incorporating analogies and rules is W.D., named after W.D. Ross [53.10]. W.D. is primarily an implementation of Ross's notion of defeasible duties, discussed above, and thus is primarily a rule-based system. However, the problem of deciding when some duties supercede over others is treated not as an intuitive judgment but as a matter of case-based reasoning. More specifically, where the order of duties is not clear, W.D. reasons by identifying decisions that are most consistent with a set of solved cases. The method is inspired by Rawls's [53.24] account of reflective equilibrium and involves making generalizations drawn from solved cases, which are then tested by comparison to further cases un-

til a conclusion is reached. Although not intended as a cognitive model, W.D. illustrates how analogical deliberation may be viewed by scholars interested in ethical reasoning.

On Thagard's [53.21] coherence theory, analogical coherence is one among several sorts of considerations that participate in ethical deliberations. Such deliberations bring many kinds of mental representations into play, such as rules, explanations, means-ends relationships, along with analogies. Decisions are made based on resolving the course of action that best coheres with all such elements as they appear in the problem. On this account, analogies play not a subordinate role but are simply one consideration amongst others.

## 53.5 Empathy

Empathy appears to be an obvious example of the deployment of models in moral deliberation. Informally, empathy involves placing oneself *in another's shoes*, that is, understanding how another person is feeling by imagining how it might feel to be in that person's position. Empathizing with someone may then create the sense that one is obligated to act, e.g., to help out. Thus, empathy is both model-based and moral in nature.

Explicating this characterization requires some caution. The notion of empathy is a relatively recent one, reaching back to accounts of *sympathy* given by David Hume and Adam Smith [53.25], but has been defined and adapted in a variety of ways. It has developed variously in different fields, including philosophy [53.26], social psychology [53.27], developmental psychology [53.28], cognitive science [53.29], and neuroscience [53.30]. Any treatment of empathy must begin by specifying how the term is to be understood.

In broad terms, empathy is one among many ways that people have of understanding or *reading* the minds of others [53.31]. One such way is to theorize about or *mentalize* what other people are thinking. This activity is focused on the beliefs and attitudes of others instead of on their immediate feelings. Since empathy is focused on feelings, it should be distinguished from detached consideration of the perspectives of others.

Also, empathy should be distinguished from *sympathy*. Both concern feelings and one's relation to other people. However, sympathizing with someone does not necessarily mean feeling the way the other person does [53.32]. It is possible to sympathize with a person who is angry or depressed, for instance, without feeling angry or depressed oneself.

Finally, empathy should be distinguished from emotional contagion [53.32]. For example, a study of people

on the Facebook social network showed that users exposed mainly to positive messages mainly produced positive messages, whereas people exposed mainly to negative messages mainly produced negative messages [53.33]. It seems that the users in this experiment just took on the mood of the producers of the messages that they read. Simple emotional contagion lacks the explicit identification of one person with another individual that defines empathy.

So, empathy can be taken more narrowly as understanding how another person feels by sharing a similar feeling through explicit identification with that person. In this narrow sense, empathy with another person can be modulated in various ways [53.32]. In other words, empathy with another person may be facilitated or inhibited by the following four factors:

1. Characteristics of feeling: It may be easier to empathize with someone experiencing a negative rather than a positive feeling, an intense rather than mild feeling, and feelings involving primary rather than secondary emotions.
2. Relationship with target: It may be easier to empathize with a person whom the empathizer likes or cares for, and for a person who is not upset with the empathizer.
3. Characteristics of the empathizer: Empathy seems to vary with the age, gender, and past experience of the empathizer. Men, for example, seem less likely than women to empathize with people who are perceived to deserve their suffering.
4. Situational factors: It is easier to empathize with someone when the cause of their feeling is known or evident, and salient. Someone crying for an ap-

parently trivial reason, for example, may not elicit empathy.

These characteristics of empathy suggest that it involves more than a shared feeling with another person. Empathy comprises an assessment of a situation in general and the place of the empathizer within it in particular.

*Barnes and Thagard* [53.26] argue that empathy is realized through analogy. That is, empathy involves a systematic mapping between the empathizer (the source) and the target. Consider the experience of watching a figure skater fall during a televised performance. Few viewers will have ever attempted a triple Salchow or anything resembling it but many will empathize with the skater nevertheless. Often, it is written all over their faces. This empathy may arise because viewers are able to construct an analogy with a different episode involving themselves, e.g., falling down during a foot race at school. A representation of this analogy can be seen in Table 53.4.

In this analogy, the empathizer constructs a systematic set of mappings in which a personal experience resulting in disappointment is compared with the experience observed for the skater. The empathizer recalls an episode of falling and injuring a knee during a race, resulting in losing the race. In one of the mappings, disappointment felt by the empathizer is mapped to the disappointment displayed by the skater. Moreover, the mapping allows the empathizer to transfer the effect of that recollected disappointment to the scene on television.

This view of empathy helps to account for some of its features noted above. It is easier to map simi-

**Table 53.4** Analogy between a racer and a figure skater

Source (foot race)	Target (figure skating)
racer	skater
track	ice
knee	groin
race	competition
disappointment	disappointment
hit(racer,track)	hit(skater,ice)
injure(racer,knee)	injure(skater,groin)
lose(racer,race)	lose(skater,competition)
feel(racer,disappointment)	display(skater,disappointment)
cause(hit,injure)	cause(hit,injure)
cause(injure,lose)	cause(injure,lose)
cause(lose,feel)	cause(lose,display)

lar feelings, for example, and feelings that are more intense. Also, people can empathize only if they have some past experience for which a strong analogy can be constructed.

On this description, empathy is not itself a form of moral deliberation [53.34]. That is, feeling empathy for someone need not change the attitudes or actions of the empathizer [53.35, pp. 327–335]. It is possible, for example, for a juror to feel empathy for an accused person without denying that the behavior of the accused was wrong and criminal. However, empathy does sometimes move people to compassion, that is, feeling sorry for a distressed person and offering assistance. People are more likely to give money to beggars, perhaps, if they can imagine themselves in the others' place. It is this connection with compassion, and perhaps other affective dispositions to assist others, that gives empathy a role in moral deliberation.

## 53.6 Role Models

The term *role model* first appears in the scholarly literature in 1957 in connection with the practice of medical students who often choose a senior figure in the profession to imitate and to use as a standard for evaluating their own performance [53.36]. However, the concept was quickly generalized to any individual who served as a reference that others might emulate or measure themselves against. In the literature, a role model can refer to a person who serves as a model with respect to a single role, e.g., doctoring. A person who serves as a reference individual for general purposes could be considered a *hero* or *idol* instead. However, the distinction is not always observed.

Typically, a role model is a person who is regarded by the modeler as someone who has attained higher sta-

tus or position, possesses superior knowledge or skills, and has achieved more than the modeler. Modelers may set themselves the goal of emulating the behavior of the model, of internalizing the attitudes of the model, or both [53.37]. It is also possible for people to adopt inferior persons as role models in the event they are seeking to save themselves from moral degradation or immoral deeds [53.38].

In the social psychology literature, the study of role models has concentrated on the impact of choosing a model on the socialization of the modeler, ways in which the comparison affects the modeler's self-assessment and conduct [53.38]. For present purposes, role models are of interest when modelers choose them for their excellence in moral values or conduct.

The process of deliberating by applying a role model seems to be analogical [53.39], that is, modelers place themselves in analogical correspondence to the model, as discussed above. The stronger the analogy, the more apt the role model.

Some moral role models exhibit an excellence of character that the modeler seeks to emulate. For example, Rosa Parks is widely considered to be a role model for people who advocate for civil rights. She refused to vacate her seat in the whites-only area of a city bus in the 1955 Montgomery bus boycott. The incident is still held up today as an exemplar of a courageous person protesting social injustice [53.40].

Rosa Parks could serve as a role model for a young American black woman seeking to protest inequality for black Americans. However, she could also serve as a role model for people of other ethnic backgrounds who are concerned about other social injustices. For ex-

ample, a white male might take part in a public rally to advocate for a path to citizenship for American immigrants from Latin America, in spite of the disapproval of his neighbors. The two episodes do not much resemble each other but the protestor may take the view that if Rosa Parks can brave the disapproval of racist police and citizenry, then he can endure the frowns of his neighbors as well. The systemic coherence between the source and target still makes this mapping a strong one. Also, the analogy transfers an emotional sense of resolve that is important in galvanizing action.

In some ways, a role model is the reverse of empathy. In the case of empathy, a person imagines what it would be like to be in another person's place. In the case of role models, a person imagines what it would be like if another person were in the modeler's place. Thus, it is not surprising that both sorts of deliberation should be similar, that is, analogical, in character.

## 53.7 Discussion

The use of models is clearly important to moral deliberation and it can be represented in many different ways. These ways include rules, schemata, mental models, analogies, empathy, and role models. The diversity among these approaches illustrates both the complexity of moral deliberation and the breadth of the concept of *model*. In view of the discussion above, it is instructive to consider this diversity further.

One obvious difference in accounts of models in moral deliberation is the emphasis they place on the role of emotions. Approaches based on rules, schemata, and mental models emphasize the cognitive component of moral deliberation with models. That is, the primary concern is with the concepts that the deliberator has and uses in the process of deliberation. Approaches based on analogies, empathy, and role models, however, emphasize more of the affective aspect of moral deliberation, though without neglecting its conceptual side. Some of this diversity may harken to the view, perhaps originating in the Enlightenment, that moral deliberation should be a matter of dispassionate calculation. As a psychological matter, moral deliberation is often emotionally charged, and no complete account of it can overlook this fact.

Similarly, accounts of moral deliberation vary according to the importance they grant to universal versus particular considerations. In the rules approach, for example, deliberations qualify as moral because they employ a small set of rules that concern moral values. The rules are combined dynamically in various ways in order to derive conclusions about moral problems that

work to guide the modeler. The rules are universal in the sense that they can be applied to many different situations.

In the analogical approach, models may be drawn from a large population of source analogs that each apply to perhaps relatively few situations. These analogs are particular in the sense that the guidance they provide is more limited. Also, analogs do not combine with the same readiness and flexibility as rules. Approaches based on rules and mental models fall on the universal side of this spectrum, whereas analogies, empathy, and role models fall on the particular side, with schemata being somewhere in between.

Another source of diversity among accounts here concerns the location of resources for deliberation. Most accounts are *internalist* in the sense that they represent deliberators as reliant solely on resources internal to them during the process of deliberation. The rule account, for example, assumes that the deliberator has rules sufficient to reach a conclusion, with the alternative being a simple failure to resolve the moral problem at all. The schemata, mental models, and analogies accounts are similar in this regard.

Other accounts are *externalist* in the sense that they represent deliberators as relying also on resources that are external to them. The role model account, for example, assumes that people consider the qualities, feelings, and actions of other people in the course of deliberation. Similarly, people who empathize with others look outside themselves for the means to resolve moral problems they face.

As Magnani [53.1] notes, resources that support moral deliberation, *moral mediators*, are not limited to other people and would extend to our technological means and to cultural resources. For example, Dehghani et al. [53.41] explore how people use analogies with cultural narratives to make prosocial moral choices. Stories of the moral decisions made by cultural heroes illustrate

for people how moral problems may be appropriately resolved. In effect, the heroes function sometimes as fictional role models. The external dimension of models in moral deliberation could be further explored.

**Acknowledgments.** Thanks to Paul Thagard for discussion of earlier drafts of this chapter.

## References

- 53.1 L. Magnani: *Morality in a Technological World: Knowledge as Duty* (Cambridge Univ. Press, New York 2007)
- 53.2 I. Kant: *Groundwork of the Metaphysics of Morals* (Harper Row, New York 1964), originally published 1785, translated by H. J. Paton
- 53.3 C. Shelley: On the impermissibility of telling misleading truths in Kantian ethics, *Open J. Philos.* **2**, 89–91 (2012)
- 53.4 J.S. Mill: *Utilitarianism*, 7th edn. (Longman's Green, London 1879)
- 53.5 D. Ross: *The Right and the Good* (Oxford Univ. Press, Oxford 2002), ed. by P. Stratton-Lake, originally published 1930
- 53.6 L. David, J. Seinfeld, P. Mehlman, C. Leifer (Writers): *Seinfeld: The Hamptons*, Motion Picture (Sony Pictures Television 1994), directed by T. Cheronos
- 53.7 N. Asher, D. Bonevac: Prima facie obligations, *Studia logica* **57**(1), 19–45 (1996)
- 53.8 H. Praken, M. Sergot: Dyadic deontic logic and contrary-to-duty obligations. In: *Defeasible Deontic Logic*, ed. by D. Nute (Kluwer Academic, Dordrecht 1997) pp. 223–262
- 53.9 M. Anderson, S.L. Anderson: Machine ethics: Creating an ethical intelligent agent, *AI Magazine* **28**(4), 15–26 (2007)
- 53.10 M. Anderson, S.L. Anderson, C. Armen: Towards machine ethics: Implementing two action-based ethical theories, *Proc. AAAI 2005 Fall Symp. Mach. Ethics*, Cryst. City (2005) pp. 1–7
- 53.11 R.M. Byrne, P.N. Johnson-Laird: 'If' and the problems of conditional reasoning, *Trends Cogn. Sci.* **13**(7), 282–287 (2009)
- 53.12 R.M. Byrne: Thinking about what \_should\_ have happened. In: *The Rational Imagination*, (MIT Press, Cambridge 2005) pp. 69–98
- 53.13 P.C. Wason: Reasoning. In: *New Horizons in Psychology*, ed. by B. Foss (Penguin, Harmondsworth 1966) pp. 135–151
- 53.14 P.N. Johnson-Laird: *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness* (Cambridge Univ. Press, Cambridge 1983)
- 53.15 M. Johnson: *Moral Imagination: Implications of Cognitive Science for Ethics* (Univ. Chicago Press, Chicago 1993)
- 53.16 G. Lakoff, M. Johnson: *Philosophy in the Flesh* (Basic, New York 1999)
- 53.17 A. Markman, D. Medin: Decision making. In: *Steven's Handbook of Experimental Psychology*, Vol. 2, ed. by D. Medin (Wiley, New York 2002) pp. 443–445
- 53.18 D. Tierney: Pearl Harbor in reverse: Moral analogies in the Cuban Missile Crisis, *J. Cold War Stud.* **9**(3), 49–77 (2007)
- 53.19 D. Gentner: Structure-mapping: A theoretical framework for analogy, *Cogn. Sci.* **7**(2), 155–170 (1983)
- 53.20 K. Holyoak, P. Thagard: *Mental Leaps: Analogy in Creative Thought* (MIT Press, Cambridge 1995)
- 53.21 P. Thagard: *Coherence in Thought and Action* (MIT Press, Cambridge 2000)
- 53.22 A.R. Jonsen, S. Toulmin: *The Abuse of Casuistry: A History of Moral Reasoning* (Univ. of California Press, Oakland 1988)
- 53.23 M. Dehghani, K. Forbus, E.K. Tomai: An integrated reasoning approach to moral decision making. In: *Machine Ethics*, ed. by M. Anderson, S.L. Anderson (Cambridge Univ. Press, Cambridge 2011)
- 53.24 J. Rawls: Outline for a decision procedure for ethics, *Philos. Rev.* **60**(2), 177–197 (1951)
- 53.25 A. Coplan, P. Goldie: Introduction. In: *Empathy: Philosophical and Psychological Perspectives*, ed. by A. Coplan, P. Goldie (Oxford Univ. Press, Oxford 2011) pp. ix–xlvii
- 53.26 A. Barnes, P. Thagard: Empathy and analogy, *Dialectic: Can. Philos. Rev.* **36**(4), 705–720 (1997)
- 53.27 J. Decety, A.N. Meltzoff: Empathy, imitation, and the social brain. In: *Empathy: Philosophical and Psychological Perspectives*, ed. by A. Coplan, P. Goldie (Oxford Univ. Press, Oxford 2011) pp. 58–81
- 53.28 N. Eisenberg, T.L. Spinrad, A. Morris: Empathy-related responding in children. In: *Handbook of Moral Development*, ed. by M. Killen, J.G. Smetana (Psychology Press, New York 2014) pp. 184–207
- 53.29 A. Goldman: Two routes to empathy: Insights from cognitive neuroscience. In: *Empathy: Philosophical and Psychological Perspectives*, ed. by A. Coplan, P. Goldie (Oxford Univ. Press, Oxford 2011) pp. 31–44
- 53.30 M. Iacoboni: *Mirroring People: The Science of Empathy and how We Connect with Others* (Picador, New York 2008)
- 53.31 T. Singer: The neuronal basis and ontogeny of empathy and mind reading: Review of the literature and implications for future research, *Neurosci. Biobehav. Rev.* **30**, 855–863 (2006)
- 53.32 F. de Vignemont, T. Singer: The empathic brain: How, when and why?, *Trends Cogn. Sci.* **10**(10), 435–441 (2006)

- 53.33 A.D. Kramer, J.E. Guillory, J.T. Hancock: Experimental evidence of massive-scale emotional contagion through social networks, *Proc. Natl. Acad. Sci. USA* **111**(24), 8788–8790 (2014)
- 53.34 J.J. Prinz: Is empathy necessary for morality? In: *Empathy: Philosophical and Psychological Perspectives*, ed. by A. Coplan, P. Goldie (Oxford Univ. Press, Oxford 2011) pp. 211–229
- 53.35 M. Nussbaum: *Upheavals of Thought: The Intelligence of Emotions* (Cambridge Univ. Press, Cambridge 2001)
- 53.36 H. Zuckerman: The role of the role model: The other side of a sociological coinage. In: *Surveying Social Life: Papers in Honor of Herbert H. Hyman*, ed. by H.H. Hyman, H.J. O’Gorman, C. Bay (Wesleyan Univ. Press, Middletown 1988) pp. 119–144
- 53.37 D.E. Gibson: Role models in career development: New directions for theory and research, J. Vocat. Behav. **65**, 134–156 (2004)
- 53.38 P. Lockwood, Z. Kunda: Outstanding role models: Do they inspire or demoralize us? In: *Psychological Perspectives on Self and Identity*, ed. by A. Tesser, R.B. Felson, J.M. Suls (American Psychological Association, Washington 2000) pp. 147–171
- 53.39 K.J. Holyoak, P. Thagard: The analogical mind, *Am. Psychol.* **52**(1), 35–44 (1997)
- 53.40 P. Callero: *The Myth of Individualism: How Social Forces Shape our Lives* (Rowman Littlefield, Plymouth 2013)
- 53.41 M. Dehghani, D. Gentner, K. Forbus, H. Ekhtiari, S. Sachdeva: Analogy in moral decision making. In: *New Frontiers in Analogy Research*, ed. by B. Kokinov, K.K. Holyoak, D. Gentner (New Bulgarian Univ. Press, Sofia 2009)



## About the Authors



### Mark Addis

St Mary's University  
School of Arts and Humanities  
Twickenham, UK  
[mark.addis@stmarys.ac.uk](mailto:mark.addis@stmarys.ac.uk)

Chapter G.33

Mark Addis is Professor of Philosophy at Birmingham City University. He is also a Research Associate at the Centre for Philosophy of Natural and Social Science at the London School of Economics and has been a visiting professor at the Department of Culture and Society at Aarhus University. His current research interests include the philosophies of mind and science.

### Atocha Aliseda

Chapter C.10

For biographical profile, please see the section "About the Part Editors".

### Francesco Amigoni

Chapter G.36

For biographical profile, please see the section "About the Part Editors".

### Margherita Arcangeli

Humboldt-Universität of Berlin  
Department of Philosophy  
Berlin, Germany  
[margheritarcangeli@gmail.com](mailto:margheritarcangeli@gmail.com)



Chapter D.21

Margherita Arcangeli graduated in Philosophy at the Università di Roma Tre (BSc 2006, MSc 2008) and obtained her PhD (2011) at the Institut Jean-Nicod, where she was also a postdoctoral researcher. Her areas of research are philosophy of mind, philosophy of science, epistemology, philosophy of language, and aesthetics. Two of her more specific topic-areas are imagination and thought experimentation.



### Cristina Barés Gómez

Universidad de Sevilla  
Grupo de Investigación en Lógica,  
Lenguaje e Información  
Sevilla, Spain  
[crisbares@gmail.com](mailto:crisbares@gmail.com)

Chapter C.14

Cristina Barés Gómez received her PhD from the University of Seville in 2012. She has worked at the National Research Center (CSIC), Spain, and is member of the Logic, Language and Information group of Seville University. She works on natural language negation, the transmission of knowledge through language (evidentiality), abduction, and defeasible reasoning.



### Alessandra Basso

University of Helsinki  
Department of Political and Economic  
Studies  
Helsinki, Finland  
[alessandra.basso@helsinki.fi](mailto:alessandra.basso@helsinki.fi)

Chapter D.19

Alessandra Basso is a PhD student at the Academy of Finland Centre of Excellence in the Philosophy of the Social Sciences. Her research focuses on the philosophy of measurement and the philosophy of the social sciences, in particular of economics.

### William Bechtel

University of California San Diego  
Department of Philosophy and Center for  
Circadian Biology  
La Jolla, USA  
[wbechtel@ucsd.edu](mailto:wbechtel@ucsd.edu)



Chapter F.27

William Bechtel is a philosopher of science who focuses on cell and molecular biology and neuroscience, especially circadian biology, and cognitive science. His research has examined biologists' reasoning strategies in developing mechanistic explanations and constructing computational models to understand their often complex, dynamic behavior. Recently he addressed how scientists employ diagrams and graphs in such reasoning.

**Mathieu Beirlaen**

Ruhr University Bochum  
Institute for Philosophy II  
Bochum, Germany  
[mathieubeirlaen@gmail.com](mailto:mathieubeirlaen@gmail.com)



Chapter C.11

Mathieu Beirlaen received his PhD from Ghent University in 2012. He has worked at Ghent University, at the National Autonomous University of Mexico, and is currently working as a postdoctoral researcher in the Workgroup for Non-Monotonic Logics and Formal Argumentation, Ruhr University Bochum. He has worked on philosophical logic at its intersection with meta-ethics, philosophy of science, and formal argumentation.

**Alisa Bokulich**

Boston University  
Center for Philosophy and History of  
Science  
Boston, USA  
[abokulich@bu.edu](mailto:abokulich@bu.edu)



Chapters A.4, H.41

Alisa Bokulich is Professor of Philosophy at Boston University and Director of the Center for Philosophy and History of Science, where she organizes the Boston Colloquium. She is Associate Member of Harvard's History of Science Department and Series Editor for *Boston Studies in the Philosophy and History of Science*.

**Tibor Bosse**

VU University Amsterdam  
Department of Computer Science  
Amsterdam, The Netherlands  
[t.bosse@vu.nl](mailto:t.bosse@vu.nl)



Chapter I.50

Tibor Bosse has been an Assistant Professor in the Agent Systems Research Group at the Department of Computer Science at VU University Amsterdam since 2006. His research focuses on computational (agent-based) modeling and simulation of social and cognitive processes. The models developed are used for both theoretical purposes and for practical applications, for instance, in the domains of ambient intelligence and serious gaming.

**Juliana Bueno-Soler**

University of Campinas  
School of Technology  
Limeira, Brazil  
[juliana@ft.unicamp.br](mailto:juliana@ft.unicamp.br)



Chapter C.15

Juliana Bueno-Soler received her PhD from the University of Campinas in 2009. She has worked at the Federal University of ABC in São Paulo and is currently Assistant Professor in Differential Calculus and Statistics at the University of Campinas. Her research now focuses on non-classical logics and probability theory.

**Angelo Cangelosi**

Plymouth University  
Centre for Robotics and Neural Systems  
Plymouth, UK  
[acangelosi@plymouth.ac.uk](mailto:acangelosi@plymouth.ac.uk)



Chapter F.28

Angelo Cangelosi is Professor of Artificial Intelligence and Cognition and the Director of the Centre for Robotics and Neural Systems at Plymouth University (UK). His main research expertise is on language grounding and embodiment in humanoid robots, developmental robotics, human-robot interaction, and neurobotics. He is Editor-in-Chief of the *IEEE Transactions on Autonomous Development*, and of *Interaction Studies*.

**Walter Carnielli**

University of Campinas  
Centre for Logic, Epistemology and the  
History of Science  
Campinas, Brazil  
[walter.carnielli@cle.unicamp.br](mailto:walter.carnielli@cle.unicamp.br)



Chapter C.15

Walter Carnielli received his PhD in Mathematics from the University of Campinas in 1982. He has worked as researcher at the University of California in Berkeley, USA and at the Universities of Münster and Bonn in Germany. He is the author of circa 80 scientific papers and 20 books and works on theory and application of contemporary logic, focused on foundations of reasoning.

**Antonio Cicchetti**

Mälardalen University  
Department of Innovation, Design, and  
Engineering  
Västerås, Sweden  
[antonio.cicchetti@mdh.se](mailto:antonio.cicchetti@mdh.se)



Chapter G.32

Antonio Cicchetti is Associate Professor in Software Engineering at Mälardalen University, Sweden. His research interests include software engineering of industrial systems by means of model-driven engineering techniques. In particular, he investigates several research problems related to the design of domain-specific languages, multi-view-based systems, model transformations, and model version management.

**Marcelo E. Coniglio**

University of Campinas  
Centre for Logic, Epistemology and the  
History of Science  
Campinas, Brazil  
[coniglio@cle.unicamp.br](mailto:coniglio@cle.unicamp.br)



Chapter C.15

Marcelo Esteban Coniglio obtained his PhD in Mathematics from the University of São Paulo in 1997. In 1998, he joined the University of Campinas (UNICAMP) faculty. In 2004 he obtained his Habilitation (Livre-Docência) in Logic from UNICAMP, and since 2013 he has been Full Professor in the Department of Philosophy of IFCH/UNICAMP. His current research focuses on non-classical logics.

**Ralf F.A. Cox**

University of Groningen  
Department of Psychology  
Groningen, The Netherlands  
[r.f.a.cox@rug.nl](mailto:r.f.a.cox@rug.nl)



Chapter F.30

Ralf Cox received his PhD in Social Sciences (Cum Laude) at the Radboud University Nijmegen in 2007. He is currently employed at the University of Groningen. His research focus is on complex dynamical systems in psychology. Specifically he studies how self-organization, embodiment, and interaction-dominant dynamics underlie cognitive and motor performances and developmental processes.

**Edoardo Datteri**

Università degli Studi di Milano-Bicocca  
Dipartimento di Scienze Umane per la  
Formazione "R.Massa"  
Milano, Italy  
[edoardo.datteri@unimib.it](mailto:edoardo.datteri@unimib.it)



Chapter G.37

Edoardo Datteri is a researcher in the philosophy of science. His main interests include the epistemology, methodology, and history of robotic simulations in artificial intelligence, biorobotics, and bionics. He has also worked on the modeling and implementation of anticipation-based algorithms of visuomotor coordination for humanoid robots at the BioRobotics Institute of the Scuola Superiore Sant'Anna, Pisa.

**Paul Davidsson**

Malmö University  
Department of Computer Science  
Malmö, Sweden  
[paul.davidsson@mah.se](mailto:paul.davidsson@mah.se)



Chapter G.35

Paul Davidsson is Professor of Computer Science at Malmö University, Sweden. He received his PhD in 1996 from Lund University, Sweden. Davidsson is currently the Director of the Internet of Things and People Research Centre. His research interests include agent technology, simulation, information systems, and data mining. Major application areas are transport systems, building automation, and energy systems.

**Ruud J.R. Den Hartigh**

University of Groningen  
Department of Psychology  
Groningen, The Netherlands  
[j.r.den.hartigh@rug.nl](mailto:j.r.den.hartigh@rug.nl)



Chapter F.30

Ruud J.R. Den Hartigh received his PhD from the University of Groningen and the University of Montpellier in 2015, with the distinction Cum Laude. He is currently Assistant Professor in the Department of Psychology at the University of Groningen. His primary research focus is on capturing performance processes in sports and beyond by applying the dynamical systems approach.

**Alessandro Di Nuovo**

University of Enna "Kore"  
Faculty of Engineering and Architecture  
Enna, Italy  
[alessandro.dinuovo@unikore.it](mailto:alessandro.dinuovo@unikore.it)



Chapter F.28

Alessandro Di Nuovo obtained the Laurea (2005) and the PhD (2009) degrees in Informatics Engineering from the University of Catania. Currently, he is Senior Lecturer with Sheffield Hallam University and Assistant Professor with the University of Enna 'Kore'. Previously, he was Research Fellow with Plymouth University (2012–2015). His current research specializes in cognitive robotics models and their social applications.

**Santo Di Nuovo**

University of Catania  
Department of Education  
Catania, Italy  
[s.dinuovo@unict.it](mailto:s.dinuovo@unict.it)



Chapter F.28

Professor Santo di Nuovo has Laurea degrees in Philosophy and Psychology. Since 1990 he has been Full Professor of Psychology at the University of Catania. He is Head of the Department of Education. He is President of the Academy of Fine Arts of Catania and Vice-president of the National Conference on Academic Psychology. His main research interests are experimental psychology, artificial intelligence, methodology and assessment, and clinical-rehabilitative and forensic psychology.

**Gordana Dodig-Crnkovic**

Chapter G.32

Chalmers University of Technology  
Department of Applied Information  
Technology  
Göteborg, Sweden  
[dodig@chalmers.se](mailto:dodig@chalmers.se)

Gordana Dodig-Crnkovic is Professor of Computer Science who is sharing her time between Chalmers University of Technology, University of Gothenburg and Mälardalen University, Sweden. Her research interests include computing paradigms, natural computing, social computing and social cognition, info-computational models, foundations of information, computational knowledge generation, computational aspects of intelligence and cognition, theory of science/philosophy of science, and computing and philosophy.

**Matthieu Fontaine**

Chapter C.14

Universidad Nacional Autónoma de México (UNAM)  
Instituto de Investigaciones Filosóficas  
Ciudad de México, Mexico  
[fontaine.matthieu@gmail.com](mailto:fontaine.matthieu@gmail.com)



Matthieu Fontaine obtained his PhD from the Université Lille 3 – Charles-de-Gaulle, France, where he worked under the supervision of Professor Dr Shahid Rahman. He has been postdoctoral fellow at the Universidad Nacional Autónoma de México for 2 years. He is specialized in the philosophy of logic and argumentation. His work more specifically focuses on fictionality, hypothetical objects, and abduction.

**Joachim Frans**

Chapter E.24

Vrije Universiteit Brussel  
Centre for Logic and Philosophy of  
Science  
Brussels, Belgium  
[joachim.frans@vub.ac.be](mailto:joachim.frans@vub.ac.be)



Joachim Frans studied philosophy at Vrije Universiteit Brussel and Ghent University. Currently, he is a member of the Centre for Logic and Philosophy of Science at Vrije Universiteit Brussel, where he is completing a PhD on explanation in mathematics.

**Roman Frigg**

Chapter A.3



London School of Economics and Political  
Science  
London, UK  
[r.p.frigg@lse.ac.uk](mailto:r.p.frigg@lse.ac.uk)

Roman Frigg is Professor of Philosophy in the Department of Philosophy, Logic and Scientific Method, Director of the Centre for Philosophy of Natural and Social Science (CPNSS) at the London School of Economics and Political Science. He holds a PhD in Philosophy from the University of London and Master's degrees in Theoretical Physics and Philosophy from the University of Basel, Switzerland.

**Tjerk Gauderis**

Chapter C.12



Ghent University  
Centre for Logic and Philosophy of Science  
Gent, Belgium  
[tjerk.gauderis@ugent.be](mailto:tjerk.gauderis@ugent.be)

Tjerk Gauderis obtained his PhD at Ghent University in 2014 with a dissertation on the formation of hypotheses which incorporated published papers in the fields of logic (abduction), philosophy of science, artificial intelligence, epistemology, and the history of physics. He is currently Associated Postdoctoral Fellow of Ghent University and an IT professional who specializes in computer aided-design, artificial intelligence, and machine learning.

**Axel Gelfert**

Chapter A.1

National University of Singapore  
Dept. of Philosophy  
Singapore, Singapore  
[axel@gelfert.net](mailto:axel@gelfert.net)



Axel Gelfert received his PhD in History and Philosophy of Science from the University of Cambridge in 2006. He is currently an Associate Professor in the Department of Philosophy at the National University of Singapore. He has authored two books.

**Charlotte Gerritsen**

Chapter I.50

Netherlands Institute for the Study of  
Crime and Law Enforcement  
Amsterdam, The Netherlands  
[cgerritsen@nscr.nl](mailto:cgerritsen@nscr.nl)



Charlotte Gerritsen is a postdoctoral researcher at the Netherlands Institute for the Study of Crime and Law Enforcement (NSCR). Her research explores possibilities of applying techniques from the area of artificial intelligence, such as agent-based modeling and simulation, to the field of criminology. She has published over 40 scientific papers for the leading international conferences and in journals in this area.

**Valeria Giardino**

Chapter E.22

UMR 7117 CNRS – Université de Lorraine  
Laboratoire d'Histoire des Sciences et de  
Philosophie – Archives Henri-Poincaré  
Nancy Cedex, France  
[valeria.giardino@univ-lorraine.fr](mailto:valeria.giardino@univ-lorraine.fr)

Valeria Giardino obtained her PhD from the University of Rome 'La Sapienza'. She has held a Marie Curie Fellowship in Paris and was a visiting researcher at Columbia University, the University of Seville, and the Freie Universität in Berlin. She is researcher at the Archives Henri Poincaré in Nancy. Her main research is on diagrammatic reasoning and the cognitive foundations of mathematics.

**Fernand Gobet**

Chapter G.33

University of Liverpool  
Department of Psychological Sciences  
Liverpool, UK  
[fernand.gobet@liv.ac.uk](mailto:fernand.gobet@liv.ac.uk)

Fernand Gobet received his PhD in 1992 from the University of Fribourg, Switzerland. He is Professor of Cognitive Psychology at the University of Liverpool. His main research interests are the psychology of expertise, the acquisition of language, computational modeling, and scientific discovery. He has written ten books.

**William Goodwin**

Chapter H.40

University of South Florida  
Department of Philosophy  
Tampa, USA  
[wgoodwin@usf.edu](mailto:wgoodwin@usf.edu)



William Goodwin is Associate Professor of Philosophy at the University of South Florida. He works in the field of philosophy of science, with particular interests in chemistry, climatology, Kuhn, and Kant.

**Bartłomiej Górný**

Chapter D.20

Comarch S.A.  
Krakow, Poland  
[bartlomiej.gorny@comarch.com](mailto:bartlomiej.gorny@comarch.com)



Bartłomiej Górný received his MS degree in Computer Science as well as his PhD degree from AGH University of Science and Technology, Cracow, Poland. He is currently Support Manager in BSS Telco Operations Business Unit in Comarch S.A. (Software House and IT Systems Integrator), Cracow, Poland. His main research interests include artificial intelligence and especially diagnosis of dynamic systems.

**Isar Goyvaerts**

Chapter E.24

Università degli Studi di Torino  
Dipartimento di Matematica "Giuseppe  
Peano"  
Torino, Italy  
[igoyvaer@unito.it](mailto:igoyvaer@unito.it)

Isar Goyvaerts received his PhD from Vrije Universiteit Brussels (Belgium) in 2013. He is currently a Marie Curie Fellow of the Istituto Nazionale di Alta Matematica based at Università degli Studi di Torino (Italy). Besides research in algebra, he has great interest in teaching mathematics and enjoys divulging mathematical topics to a broader audience.

**Teruaki Hayashi**

Chapter I.48

University of Tokyo  
Department of Systems Innovation  
Bunkyo-ku, Tokyo, Japan  
[teru-h.884@nifty.com](mailto:teru-h.884@nifty.com)

Teruaki Hayashi is an Engineering PhD candidate at the University of Tokyo. He is studying Innovators Marketplace and the methods for refining ideas through communications and constraints in action planning. Recently he has been contributing to projects for data driven innovation and decision-making, integrated with communication for creative activities.

**Cyrille Imbert**

Chapter G.34

Université de Lorraine  
Archives Poincaré  
Nancy, France  
[cyrille.imbert@univ-lorraine.fr](mailto:cyrille.imbert@univ-lorraine.fr)



Cyrille Imbert received his PhD in 2008 from Université Paris 1 and has worked ever since as a permanent CNRS Research Fellow at Archives Poincaré (Nancy, France). He works in the field of general philosophy of science, with emphasis on scientific explanation, modeling, and computational issues, and in the (formal) social epistemology of science, with the use of models and computer simulations.

**Hiroyuki Kido**

University of Tokyo  
Department of Systems Innovation  
Bunkyo-ku, Tokyo, Japan  
[kido.hiroyuki@gmail.com](mailto:kido.hiroyuki@gmail.com)



Chapter I.48

Hiroyuki Kido graduated from Niigata University and received his Master's degree from Hokkaido University. He received his PhD from Tokyo Institute of Technology in 2011. Presently, he is Assistant Professor of the School of Engineering at the University of Tokyo. He won the JSAI (the Japanese Society for Artificial Intelligence) Best Paper award in 2012.

**Franziska Klügl**

Örebro University  
School of Science and Technology  
Örebro, Sweden  
[franziska.klugl@oru.se](mailto:franziska.klugl@oru.se)



Chapter G.35

Franziska Klügl received her PhD and her Habilitation in Computer Science from the University of Würzburg (Germany) in 2000 and 2009. Since 2011, she has been working as Professor in Information Technology at Örebro University, Sweden. Her research interests are engineering and application of agent-based simulation and multiagent systems with a special focus on human-computer interaction.

**Peter C.R. Lane**

University of Hertfordshire  
School of Computer Science  
Hatfield, UK  
[p.c.lane@herts.ac.uk](mailto:p.c.lane@herts.ac.uk)



Chapter G.33

Peter Lane has a PhD in Computer Science from the University of Exeter and a BA in Mathematics and Computation from Oxford University. He worked at Nottingham University before taking up his current position at the University of Hertfordshire. His research interests cover aspects of machine learning, including cognitive modeling, data mining, and scientific discovery.

**Antoni Ligęza**

AGH University of Science and Technology  
Applied Computer Science  
Krakow, Poland  
[ligeza@agh.edu.pl](mailto:ligeza@agh.edu.pl)



Chapter D.20

Professor Antoni Ligęza from AGH – University of Science and Technology in Cracow, Poland, works on knowledge engineering (artificial intelligence), knowledge representation and reasoning, rule-based systems, technical diagnostics, constraint programming, logics, and systems science. He has been a visiting researcher at LAAS/Toulouse, the University of Nancy I, the University of the Balearic Islands, the University of Girona, and the University of Caen. He is the author of about 200 research papers.

**Chiara Liscianra**

University of Groningen  
Faculty of Economics and Business,  
Department of Economics, Econometrics  
and Finance  
Groningen, The Netherlands  
[c.liscianra@rug.nl](mailto:c.liscianra@rug.nl)



Chapter D.19

Chiara Liscianra is doing postdoctoral research at the Tint Centre of Excellence in the Philosophy of the Social Sciences, the University of Helsinki. She received her PhD from Tilburg University in 2013. After her stay at Tilburg, Chiara spent 1 year as a Research Fellow at Munich Center for Mathematical Philosophy, Munich University.

**Elisabeth A. Lloyd**

Indiana University  
Department of History and Philosophy of  
Science and Medicine  
Bloomington, USA  
[ealloyd@indiana.edu](mailto:ealloyd@indiana.edu)



Chapter H.42

Elisabeth A. Lloyd's focus is dual: both philosophy of evolution and climate science. She works on the logical structure of evolutionary models and their confirmation, as well as the methodology and testing of biological and climate models. Her book (2005) has led to changes in sex research, including her participation in research on human female sexuality.

**Giuseppe Longo**

Centre Cavallès  
Ecole Normale Sup.  
Paris, France  
[giuseppe.longo@ens.fr](mailto:giuseppe.longo@ens.fr)



Chapter H.38

Giuseppe Longo is a mathematician and Directeur de Recherche, CNRS at Ecole Normale Supérieure, Paris. He is a former Professor of Mathematical Logic, University of Pisa. He spent 3 years in Berkeley, MIT and Carnegie Mellon (USA) as researcher and Visiting Professor. He has worked on applications of mathematics to computer science and recently extended his research interests to theoretical biology.

**Miles MacLeod**

University of Twente  
Department of Philosophy  
Enschede, The Netherlands  
[m.a.j.macleod@utwente.nl](mailto:m.a.j.macleod@utwente.nl)



## Chapter A.5

Miles MacLeod is a philosopher of science with a PhD in History and Philosophy of Science from the University of Vienna. He works and publishes on themes connected to model-building practices in the biological and bioengineering sciences, scientific cognition, and interdisciplinary collaboration. He studies these model-building practices for the most part using qualitative methods.

**Giovanna Magnani**

University of Pavia  
Department of Economics and  
Management  
Pavia, Italy  
[g.magnani@unipv.it](mailto:g.magnani@unipv.it)



## Chapter I.51

Giovanna Magnani is a Post Doc research fellow at the Department of Economics and Management of the University of Pavia. She earned her PhD in Economics and Management in December, 2015. She was visiting PhD candidate at the University of Queensland (Australia) between January and June 2014. She has participated in several international conferences and is co-author of a monograph on International Entrepreneurship.

**Caterina Marchionni**

University of Helsinki  
Department of Political and Economic  
Studies  
Helsinki, Finland  
[caterina.marchionni@helsinki.fi](mailto:caterina.marchionni@helsinki.fi)

## Chapter D.19

Caterina Marchionni obtained her PhD from the Erasmus Institute for Philosophy and Economics, Erasmus University Rotterdam. Currently she is Academy Research Fellow at the Academy of Finland Centre of Excellence in the Philosophy of the Social Sciences (TINT), University of Helsinki. Her research mainly concerns modeling, explanation, and interdisciplinarity.

**Davide Marocco**

Plymouth University, School of Computing  
Electronics and Mathematics  
Plymouth, UK  
[davide.marocco@plymouth.ac.uk](mailto:davide.marocco@plymouth.ac.uk)

## Chapter F.28

Davide Marocco received his PhD in Artificial Intelligence from the University of Calabria, Italy, in 2004. He is Associate Professor (Reader) of Cognitive Robotics and Intelligent Systems at Plymouth University, UK and Coordinator of the CUDA Teaching Centre. His research interests are focused on cognitive robotics and evolutionary robotics models of behaviour, and evolution of communication and language.

**Massimo Marraffa**

University of Rome 'Roma Tre'  
Department of Philosophy,  
Communication and Media Studies  
Rome, Italy  
[massimo.marraffa@uniroma3.it](mailto:massimo.marraffa@uniroma3.it)



## Chapter H.43

Massimo Marraffa is Associate Professor of Philosophy of Science at University Roma Tre (Rome, Italy). His research focuses primarily on issues in the philosophy of the mind and the philosophy of psychology, on which he has published books, articles, and book chapters in Italian and English.

**Mary Ann Metzger**

University of Maryland UMBC  
Department of Psychology  
Baltimore, USA  
[metzger@umbc.edu](mailto:metzger@umbc.edu)



## Chapter F.29

Mary Ann Metzger received her PhD in Psychology from the University of Connecticut, Storrs, CT, USA in 1970 and then did postdoctoral work at the laboratory of William K. Estes, Rockefeller University, New York, NY. She is an Emerita Associate Professor in the Department of Psychology, University of Maryland UMBC, Baltimore, MD. Her work focuses on the development of dynamic models of cognition.

**Gerhard Minnameier**

Goethe University Frankfurt am Main  
Faculty of Economics and Business  
Administration  
Frankfurt am Main, Germany  
[minnameier@econ.uni-frankfurt.de](mailto:minnameier@econ.uni-frankfurt.de)

## Chapter B.8

Gerhard Minnameier received his PhD from Johannes Gutenberg University at Mainz in 1999, where he also habilitated in 2005. He has been Professor of Vocational and Economics Education at RWTH University Aachen, and since 2011 Professor of Business Ethics and Business Education at Goethe University Frankfurt am Main. His key research areas are cognitive structures and learning, especially in the moral domain.

**Maël Montévil**

Chapter H.38

Laboratoire MSC  
 Université Paris 7 Diderot  
 Paris, France  
*mael.montevil@gmail.com*

Dr Maël Montévil works at the crossroad of biology, physics, and philosophy of science in order to contribute to a deeper understanding of biological phenomena. He obtained his PhD from École Normale Supérieure with Giuseppe Longo, was a Postdoctoral Associate in Tufts University School of Medicine with Ana Soto and Carlos Sonnenschein, and in IHPST, Université Paris 1 Panthéon-Sorbonne.

**Eleonora Montuschi**

Chapter I.52

Università Ca'Foscari Venezia  
 Department of Philosophy and Cultural  
 Heritage  
 Venice, Italy  
*eleonora.montuschi@unive.it*



Eleonora Montuschi is Associate Professor in the Department of Philosophy and Cultural Heritage at Ca' Foscari University of Venice and Senior Research Fellow at the London School of Economics and Political Science. Her research interests include philosophy of science and social science, objectivity, the theory and practice of evidence, and methodological issues in the social sciences.

**John Mumma**

Chapter E.23

California State University San  
 Bernardino  
 Department of Philosophy,  
 San Bernardino, USA  
*jmumma@csusb.edu*



John Mumma is an Assistant Professor of Philosophy at California State University San Bernardino. His research interests are in the philosophy of geometry, the philosophy of mathematics, and the philosophy of logic.

**Angel Nepomuceno-Fernández**

Chapter C.13



Universidad de Sevilla  
 Grupo de Investigación en Lógica,  
 Lenguaje e Información  
 Sevilla, Spain  
*nepomuce@us.es*

Angel Nepomuceno-Fernández is a Full Professor of Logic and Philosophy of Science and the principal investigator of the Group of Logic, Language and Information at the University of Seville. His work focuses on nonclassical logics, particularly dynamic epistemic logic and other multimodal logics.

**Nancy J. Nersessian**

Chapter A.5



Harvard University  
 Dept. of Psychology, William James Hall  
 Cambridge, USA  
*nancynersessian@fas.harvard.edu*

Nancy J. Nersessian received her PhD from Case Western Reserve in 1977. She is Regents' Professor (Emerita), Georgia Tech and Research Associate at Harvard University. Her research focuses on the creative modeling practices of scientists and engineers. She is a Fellow of AAAS, the Cognitive Science Society, and a Foreign Member of the Royal Netherlands Academy of Arts and Sciences.

**James Nguyen**

Chapter A.3

London School of Economics and Political  
 Science  
 London, UK  
*j.nguyen1@lse.ac.uk*



James Nguyen is a PhD candidate in the Department of Logic, Philosophy and Scientific Method at the London School of Economics and Political Science. He works on the nature of representation in science.

**Yukio Ohsawa**

Chapter I.48

University of Tokyo  
 Department of Systems Innovation  
 Bunkyo-ku, Tokyo, Japan  
*ohsawa@sys.t.u-tokyo.ac.jp*



Yukio Ohsawa is a professor at the University of Tokyo. His research started from natural language analysis, and, via working in non-linear optics and artificial intelligence, he initiated studies on *chance discovery*, i.e., discovery of events meaningful for decision-making, and extended them to methods for innovation. He has edited and authored several books on the topics of change discovery and innovation.




**Naomi Oreskes**

Chapter H.41

Harvard University  
Department of the History of Science  
Cambridge, USA  
[oreskes@fas.harvard.edu](mailto:oreskes@fas.harvard.edu)

Naomi Oreskes is Professor of the History of Science and Affiliated Professor of Earth and Environmental Sciences at Harvard University.

**Woosuk Park**

Chapter B.9

For biographical profile, please see the section "About the Part Editors".

**Alfredo Paternoster**

Chapter H.43

Università di Bergamo  
Department of Letters, Philosophy,  
Communication  
Bergamo, Italy  
[alfredo.paternoster@unibg.it](mailto:alfredo.paternoster@unibg.it)



Alfredo Paternoster is Associate Professor of Philosophy of Language at University of Bergamo, Italy. His research focuses primarily on issues in philosophy of the mind, philosophy of cognitive sciences, and philosophy of language, on which he has published books, articles, and book chapters in Italian and English.

**Pieter Pauwels**

Chapter I.45

Ghent University  
Department of Architecture and Urban  
Planning  
Ghent, Belgium  
[pipauwel.pauwels@ugent.be](mailto:pipauwel.pauwels@ugent.be)



Pieter Pauwels is a postdoctoral researcher at the Department of Architecture and Urban Planning at Ghent University. He holds a Master's degree (2008, Ghent) and a PhD degree (2012, Ghent) in Engineering Architecture. His research focuses on information system support for architectural design thinking. He works full time on topics affiliated to BIM, design thinking and linked data in architecture and construction.

**Demetris Portides**

Chapter A.2

For biographical profile, please see the section "About the Part Editors".

**Athanasios Raftopoulos**

Chapter F.26

For biographical profile, please see the section "About the Part Editors".

**Ferdinand Rivera**

Chapter E.25

San José University  
Department of Mathematics & Statistics  
San Jose, USA  
[ferdinand.rivera@sjsu.edu](mailto:ferdinand.rivera@sjsu.edu)



Ferdinand Rivera is Professor in the Department of Mathematics and Statistics and Chair of the Department of Elementary Education at San Jose State University. He conducts research in school-based mathematics cognition, focusing on the emergence and development of structural thinking and deep mathematical understanding in children, adolescents, and adults.

**Abilio Rodrigues Filho**

Chapter C.15

Federal University of Minas Gerais  
Department of Philosophy  
Pampulha, Belo Horizonte, Brazil  
[abilio@ufmg.br](mailto:abilio@ufmg.br)



Abilio Rodrigues Filho is Professor of Philosophy at the Federal University of Minas Gerais, Belo Horizonte, Brazil. He received his PhD from the Pontifical Catholic University of Rio de Janeiro, Brazil. His main areas of interest are logic, philosophy of logic, paraconsistency, and intuitionism.


**Federica Russo**

Chapter H.44

University of Amsterdam  
Department of Philosophy  
Amsterdam, The Netherlands  
[f.russo@uva.nl](mailto:f.russo@uva.nl)

Federica Russo received her PhD from the University of Louvain (Belgium) in 2005. Since then, she has held several research and teaching positions in Belgium, Italy, UK, and USA. She is currently Assistant Professor in Philosophy of Science at the University of Amsterdam. Her research interests span the social, biomedical, and policy sciences, with particular attention on questions about causality and modeling.

**Flavia Santoianni**

Chapter I.49

University of Naples Federico II  
Department of Humanities Section of  
Philosophy  
Naples, Italy  
[flavia.santoianni@unina.it](mailto:flavia.santoianni@unina.it)

Flavia Santoianni is Full Professor of Education at University of Naples Federico II. She is Director of the international open access journal *RTH Research Trends in Humanities* and has published 26 books and several articles. Her books have been translated into English and Spanish. Her research interests concern bioeducational sciences, teaching and learning, and design of learning environments.

**Viola Schiaffonati**

Chapter G.36

For biographical profile, please see the section "About the Part Editors".

**Gerhard Schurz**

Chapter B.7

Heinrich Heine University  
Department of Philosophy, DCLPS  
Dusseldorf, Germany  
[schurz@phil.uni-duesseldorf.de](mailto:schurz@phil.uni-duesseldorf.de)



Gerhard Schurz is Professor of Philosophy and Director of the Düsseldorf Center for Logic and Philosophy of Science (DCLPS) at the University of Düsseldorf. Before 2002 he was Associate Professor at the University of Salzburg and Visiting Professor at the University of California at Irvine and at Yale University. His research areas cover philosophy of science, logic, epistemology and cognitive science.

**Nora Alejandrina Schwartz**

Chapter D.17

For biographical profile, please see the section "About the Part Editors".

**Cameron Shelley**

Chapters I.47, I.53

For biographical profile, please see the section "About the Part Editors".

**Fernando Soler-Toscano**

Chapter C.13

Universidad de Sevilla  
Grupo de Investigación en Lógica,  
Lenguaje e Información  
Sevilla, Spain  
[fsoler@us.es](mailto:fsoler@us.es)



Fernando Soler-Toscano is Associate Professor at the University of Seville, Spain. His background is both in philosophy and computer science. His research focuses on formal models of abductive reasoning in dynamic epistemic logic. He also works on algorithmic complexity measures. He is a foreign collaborator of the Lisbon Center for Philosophy of Sciences and a member of the Paris Reasoning group.

**Peter D. Sozou**

Chapter G.33

London School of Economics and  
Political Science  
Centre for Philosophy of Natural and  
Social Science  
London, UK  
[p.sozou@lse.ac.uk](mailto:p.sozou@lse.ac.uk)



Peter Sozou received his PhD from Birkbeck College, University of London, in 1994. He has worked on computer vision, behaviour, theoretical biology, health, reproductive medicine, and computational scientific discovery. He is based at the LSE, and is interested in a range of problems in decision-making. During the writing of this chapter he was a researcher at the University of Liverpool.

**Susan G. Sterrett**

Chapters D.18, H.39



Wichita State University  
Department of Philosophy  
Wichita, USA  
[susangsterrett@gmail.com](mailto:susangsterrett@gmail.com)

S.G. Sterrett is the Curtis D. Gridley Distinguished Professor of the History and Philosophy of Science, Department of Philosophy, Wichita State University in Wichita, Kansas. She has an undergraduate degree in engineering from Cornell University, and degrees in Mathematics (MA) and Philosophy (MA, PhD) from the University of Pittsburgh. She has also taught at Duke University and Carnegie-Mellon University.

**Ryan D. Tweney**

Chapter D.16



Bowling Green State University  
Department of Psychology  
Bowling Green, USA  
[tweney@bgsu.edu](mailto:tweney@bgsu.edu)

Ryan D. Tweney has a PhD in Experimental Psychology. His research has ranged widely over psycholinguistics, the role of confirmation 'bias' in thinking, problem solving, and the history of science and of psychology, to which he has applied cognitive models of thinking. His current research centers on James Clerk Maxwell's use of mathematical representations, and the relation of such representations to model-based reasoning in science.

**Paul L.C. Van Geert**

University of Groningen  
Department of Psychology  
Groningen, The Netherlands  
*p.l.c.van.geert@rug.nl*



Chapter F.30

Paul van Geert is Emeritus Professor at the University of Groningen in the Netherlands, where he held the Chair of Developmental Psychology (1985–2015). He has a pioneering role in the application of dynamic systems theory across the areas of language, cognitive, and social development, and has held visiting professorships at Harvard University and various European Universities.

**Bart Van Kerkhove**

Vrije Universiteit Brussel  
Centre for Logic and Philosophy of  
Science  
Brussels, Belgium  
*bart.van.kerkhove@vub.ac.be*



Chapter E.24

Bart Van Kerkhove has been a member of the Centre for Logic and Philosophy of Science at Vrije Universiteit Brussel since 2000 and received his PhD there in 2005. Since 2008, he has also been a part-time lecturer of the Philosophy Department, teaching various courses, including Logic and Philosophy of Science, Themes from Analytic Philosophy and Philosophy of Mathematics.

**Fernando R. Velázquez-Quesada**

Universidad de Sevilla  
Grupo de Investigación en Lógica,  
Lenguaje e Información  
Sevilla, Spain  
*frvelazquezquesada@us.es*



Chapter C.13

Fernando R. Velazquez-Quesada received his PhD from the Institute for Logic, Language and Computation of the University of Amsterdam. His research focuses on logical representations of individual and collective attitudes towards information (such as knowledge, beliefs, or preferences) as well as the different actions that affect them (as diverse forms of inference, communication, and interaction).

**Harko Verhagen**

Stockholm University  
Department of Computer and Systems  
Sciences  
Kista, Sweden  
*verhagen@dsv.su.se*



Chapter G.35

Harko Verhagen is Associate Professor of Computer and Systems Sciences at Stockholm University, Sweden, where he received his PhD in 2000. His research interests include social simulation, social believability in computer games, IT supported learning, outsourcing, hybrid social systems, normative multiagent systems, and theories of agency. Major application areas are computer games, socio-cognitive technical systems, and social simulations.

**Jonathan Waskan**

University of Illinois at  
Urbana-Champaign  
Department of Philosophy  
Urbana, USA  
*waskan@illinois.edu*



Chapter F.31

Jonathan received his PhD in 1999 through the Philosophy-Neuroscience-Psychology program at Washington University in Saint Louis. He has held teaching and research appointments at William Paterson University, the University of Illinois at Urbana-Champaign (and the Beckman Institute), and the Rotman Institute at the University of Western Ontario. His research largely concerns mental models and their role in scientific explanation.

**John Woods**

University of British Columbia  
Dept. of Philosophy, Vancouver Campus  
Vancouver, Canada  
*john.woods@ubc.ca*



Chapter B.6

John Woods has done pioneering work in the logic of fallacious reasoning, the logic of fiction, argumentation theory, conflict resolution in the abstract sciences, the logic of abduction, Aristotle's earlier logic, naturalized logic, the logic of inconsistency robustness and the logic of legal reasoning. Fellow of the Royal Society of Canada, John Woods is a Life Member of the Association of Fellows, The Netherlands Institute for Advanced Study.

**Alison Wylie**

University of Washington  
Department of Philosophy and  
Anthropology, Savery Hall  
Seattle, USA  
*aw26@uw.edu*



Chapter I.46

Alison Wylie is a philosopher of science who works on issues raised by archaeological practice and feminist research. She has co-edited and authored two books, and contributes to collections on topics.

## Detailed Contents

<b>List of Abbreviations</b> .....	XXXVII
------------------------------------	--------

### Part A Theoretical Issues in Models

#### 1 The Ontology of Models

<i>Axel Gelfert</i> .....	5
1.1 Kinds of Models: Examples from Scientific Practice .....	6
1.2 The Nature and Function of Models .....	8
1.3 Models as Analogies and Metaphors .....	10
1.4 Models Versus the Received View: Sentences and Structures .....	12
1.4.1 Models and the Study of Formal Languages .....	13
1.4.2 The Syntactic View of Theories .....	13
1.4.3 The Semantic View .....	14
1.4.4 Partial Structures .....	15
1.5 The Folk Ontology of Models .....	16
1.6 Models and Fiction .....	18
1.7 Mixed Ontologies: Models as Mediators and Epistemic Artifacts .....	20
1.7.1 Models as Mediators .....	20
1.7.2 Models as Epistemic Artifacts .....	21
1.8 Summary .....	21
<b>References</b> .....	22

#### 2 Models and Theories

<i>Demetris Portides</i> .....	25
2.1 The Received View of Scientific Theories .....	26
2.1.1 The Observation–Theory Distinction .....	27
2.1.2 The Analytic–Synthetic Distinction .....	29
2.1.3 Correspondence Rules .....	30
2.1.4 The Cosmetic Role of Models According to the RV .....	32
2.1.5 Hempel's Provisos Argument .....	33
2.1.6 Theory Consistency and Meaning Invariance .....	34
2.1.7 General Remark on the Received View .....	35
2.2 The Semantic View of Scientific Theories .....	36
2.2.1 On the Notion of Model in the SV .....	38
2.2.2 The Difference Between Various Versions of the SV .....	40
2.2.3 Scientific Representation Does not Reduce to a Mapping of Structures .....	42
2.2.4 A Unitary Account of Models Does not Illuminate Scientific Modeling Practices .....	44
2.2.5 General Remark on the Semantic View .....	46
<b>References</b> .....	47

<b>3</b>	<b>Models and Representation</b>	
	<i>Roman Frigg, James Nguyen</i>	49
3.1	Problems Concerning Model–Representation	51
3.2	General Griceanism and Stipulative Fiat	55
3.3	The Similarity Conception	57
3.3.1	Similarity and ER–Problem	58
3.3.2	Accuracy and Style	62
3.3.3	Problems of Ontology	64
3.4	The Structuralist Conception	66
3.4.1	Structures and the Problem of Ontology	66
3.4.2	Structuralism and the ER–Problem	68
3.4.3	Accuracy, Style and Demarcation	70
3.4.4	The Structure of Target Systems	71
3.5	The Inferential Conception	76
3.5.1	Deflationary Inferentialism	76
3.5.2	Inflating Inferentialism: Interpretation	80
3.5.3	The Denotation, Demonstration, and Interpretation Account	82
3.6	The Fiction View of Models	83
3.6.1	Models and Fiction	84
3.6.2	Direct Representation	86
3.6.3	Parables and Fables	88
3.6.4	Against Fiction	89
3.7	Representation–as	91
3.7.1	Exemplification and Representation–as	91
3.7.2	From Pictures to Models: The Denotation, Exemplification, Keying–up and Imputation Account	93
3.8	Envoi	96
	<b>References</b>	96
<b>4</b>	<b>Models and Explanation</b>	
	<i>Alisa Bokulich</i>	103
4.1	The Explanatory Function of Models	104
4.2	Explanatory Fictions: Can Falsehoods Explain?	108
4.3	Explanatory Models and Noncausal Explanations	112
4.4	How–Possibly versus How–Actually Model Explanations	114
4.5	Tradeoffs in Modeling: Explanation versus Other Functions for Models	115
4.6	Conclusion	116
	<b>References</b>	117
<b>5</b>	<b>Models and Simulations</b>	
	<i>Nancy J. Nersessian, Miles MacLeod</i>	119
5.1	Theory–Based Simulation	119
5.2	Simulation not Driven by Theory	121
5.3	What is Philosophically Novel About Simulation?	124
5.4	Computational Simulation and Human Cognition	127
	<b>References</b>	130

## Part B Theoretical and Cognitive Issues on Abduction and Scientific Inference

<b>6 Reorienting the Logic of Abduction</b>	
<i>John Woods</i> .....	137
6.1 Abduction .....	138
6.1.1 Peirce's Abduction .....	138
6.1.2 Ignorance Problems .....	138
6.1.3 The Gabbay–Woods Schema .....	139
6.1.4 The Yes–But Phenomenon .....	140
6.2 Knowledge .....	141
6.2.1 Epistemology .....	141
6.2.2 Losing the <i>J</i> -Condition .....	142
6.2.3 The Causal Response Model of Knowledge .....	142
6.2.4 Naturalism .....	143
6.2.5 Showing and Knowing .....	143
6.2.6 Explaining the Yes–Buts .....	144
6.2.7 Guessing .....	144
6.2.8 Closed Worlds .....	146
6.3 Logic .....	148
6.3.1 Consequences and Conclusions .....	148
6.3.2 Semantics .....	148
<b>References</b> .....	149
<b>7 Patterns of Abductive Inference</b>	
<i>Gerhard Schurz</i> .....	151
7.1 General Characterization of Abductive Reasoning and Ibe .....	152
7.2 Three Dimensions for Classifying Patterns of Abduction .....	154
7.3 Factual Abduction .....	155
7.3.1 Observable–Fact Abduction .....	155
7.3.2 First-Order Existential Abduction .....	156
7.3.3 Unobservable–Fact Abduction .....	156
7.3.4 Logical and Computational Aspects of Factual Abduction .....	157
7.4 Law Abduction .....	158
7.5 Theoretical–Model Abduction .....	159
7.6 Second–Order Existential Abduction .....	161
7.6.1 Micro–Part Abduction .....	161
7.6.2 Analogical Abduction .....	161
7.6.3 Hypothetical (Common) Cause Abduction .....	162
7.7 Hypothetical (Common) Cause Abduction Continued .....	162
7.7.1 Speculative Abduction Versus Causal Unification: A Demarcation Criterion .....	163
7.7.2 Strict Common–Cause Abduction from Correlated Dispositions and the Discovery of New Natural Kinds .....	164
7.7.3 Probabilistic Common–Cause Abduction and Statistical Factor Analysis .....	167
7.7.4 Epistemological Abduction to Reality .....	168
7.8 Further Applications of Abductive Inference .....	169
7.8.1 Abductive Belief Revision .....	169
7.8.2 Instrumental Abduction and Technological Reasoning .....	170
<b>References</b> .....	171

<b>8</b>	<b>Forms of Abduction and an Inferential Taxonomy</b>	
	<i>Gerhard Minnameier</i> .....	175
8.1	Abduction in the Overall Inferential Context .....	177
8.1.1	Disentangling Abduction and IBE .....	177
8.1.2	The Dynamical Interaction of Abduction, Deduction, and Induction .....	179
8.1.3	Abduction and Abstraction .....	180
8.2	The Logicity of Abduction, Deduction, and Induction .....	183
8.2.1	Inferential Subprocesses and Abduction as Inferential Reasoning .....	183
8.2.2	The Validity of Abduction, Deduction, and Induction .....	184
8.3	Inverse Inferences .....	185
8.3.1	Theorematic Deduction as Inverse Deduction .....	185
8.3.2	An Example for Theorematic Deduction .....	187
8.3.3	Inverse Abduction and Inverse Induction .....	188
8.4	Discussion of Two Important Distinctions Between Types of Abduction .....	189
8.4.1	Creative Versus Selective Abduction .....	189
8.4.2	Factual Versus Theoretical Abduction .....	190
8.4.3	Explanatory Versus Nonexplanatory Abduction .....	192
8.5	Conclusion .....	193
	<b>References</b> .....	193
<b>9</b>	<b>Magnani's Manipulative Abduction</b>	
	<i>Woosuk Park</i> .....	197
9.1	Magnani's Distinction Between Theoretical and Manipula- tive Abduction .....	197
9.2	Manipulative Abduction in Diagrammatic Reasoning .....	198
9.2.1	Abductive and Manipulative Aspects of Diagrammatic Reasoning .....	198
9.2.2	Magnani on Manipulative Abduction in Diagrammatic Reasoning .....	201
9.3	When Does Manipulative Abduction Take Place? .....	203
9.4	Manipulative Abduction as a Form of Practical Reasoning .....	204
9.5	The Ubiquity of Manipulative Abduction .....	206
9.5.1	Manipulative Abduction in Fallacies .....	206
9.5.2	Manipulative Abduction in Animals .....	207
9.6	Concluding Remarks .....	212
	<b>References</b> .....	212
<b>Part C The Logic of Hypothetical Reasoning, Abduction, and Models</b>		
<b>10</b>	<b>The Logic of Abduction: An Introduction</b>	
	<i>Atocha Aliseda</i> .....	219
10.1	Some History .....	219
10.1.1	Induction and Abduction .....	219
10.1.2	The Founding Father: C.S. Peirce .....	220
10.1.3	The Cognitive Sciences .....	221

10.1.4	Some Examples .....	221
10.2	Logical Abduction .....	222
10.2.1	Argument.....	222
10.2.2	Inference to the Best Explanation .....	223
10.2.3	A Taxonomy .....	223
10.3	Three Characterizations .....	225
10.3.1	Inferential.....	225
10.3.2	Computational .....	226
10.3.3	Epistemic Change .....	227
10.4	Conclusions .....	228
	<b>References</b> .....	229
<b>11</b>	<b>Qualitative Inductive Generalization and Confirmation</b>	
	<i>Mathieu Beirlaen</i> .....	231
11.1	Adaptive Logics for Inductive Generalization .....	231
11.2	A First Logic for Inductive Generalization .....	232
11.2.1	General Characterization of the Standard Format .....	232
11.2.2	Proof Theory .....	233
11.2.3	Minimal Abnormality.....	236
11.3	More Adaptive Logics for Inductive Generalization .....	237
11.4	Qualitative Inductive Generalization and Confirmation.....	240
11.4.1	I-Confirmation and Hempel's Adequacy Conditions .....	240
11.4.2	I-Confirmation and the Hypothetico-Deductive Model .....	242
11.4.3	Interdependent Abnormalities and Heuristic Guidance... ..	243
11.5	Conclusions .....	245
11.A	Appendix: Blocking the Raven Paradox? .....	246
	<b>References</b> .....	247
<b>12</b>	<b>Modeling Hypothetical Reasoning by Formal Logics</b>	
	<i>Tjerk Gauderis</i> .....	249
12.1	The Feasibility of the Project .....	249
12.2	Advantages and Drawbacks .....	251
12.3	Four Patterns of Hypothetical Reasoning.....	252
12.4	Abductive Reasoning and Adaptive Logics .....	255
12.5	The Problem of Multiple Explanatory Hypotheses.....	256
12.6	The Standard Format of Adaptive Logics.....	256
12.6.1	Dynamic Proof Theory .....	257
12.7	$LA_s^t$ : A Logic for Practical Singular Fact Abduction.....	258
12.7.1	Lower Limit Logic .....	258
12.7.2	Set of Abnormalities $\Omega$ .....	258
12.7.3	Reliability Strategy.....	259
12.7.4	Practical Abduction .....	260
12.7.5	Avoiding Random Hypotheses.....	260
12.8	$MLA_s^t$ : A Logic for Theoretical Singular Fact Abduction.....	261
12.8.1	Formal Language Schema.....	261
12.8.2	Lower Limit Logic .....	261
12.8.3	Intended Interpretation of the Modal Operators.....	261
12.8.4	Set of Abnormalities.....	261
12.8.5	First Proposal $\Omega_1$ .....	262
12.8.6	Simple Strategy .....	262



12.8.7	Contradictory Hypotheses .....	262
12.8.8	Predictions and Evidence .....	262
12.8.9	Contradictions .....	263
12.8.10	Tautologies .....	263
12.8.11	Second Proposal $\Omega_2$ .....	263
12.8.12	Most Parsimonious Explanantia .....	263
12.8.13	Notation .....	264
12.8.14	Final Proposal $\Omega$ .....	264
12.9	Conclusions .....	265
12.A	Appendix: Formal Presentations of the Logics $LA_5^f$ and $MLA_5^s$ .....	265
12.A.1	Proof Theory .....	265
12.A.2	Semantics .....	266
	<b>References</b> .....	267
<b>13</b>	<b>Abductive Reasoning in Dynamic Epistemic Logic</b>	
	<i>Angel Nepomuceno–Fernández, Fernando Soler–Toscano,</i>	
	<i>Fernando R. Velázquez–Quesada</i> .....	269
13.1	Classical Abduction .....	270
13.2	A Dynamic Epistemic Perspective .....	272
13.2.1	What Is an Abductive Problem? .....	272
13.2.2	What Is an Abductive Solution? .....	273
13.2.3	How is the Best Explanation Selected? .....	273
13.2.4	How is the Best Explanation Incorporated Into the Agent's Information? .....	274
13.2.5	Abduction in a Picture .....	274
13.3	Representing Knowledge and Beliefs .....	275
13.3.1	Language and Models .....	275
13.3.2	Operations on Models .....	276
13.4	Abductive Problem and Solution .....	278
13.4.1	Abductive Problem .....	278
13.4.2	Classifying Problems .....	279
13.4.3	Abductive Solutions .....	279
13.4.4	Classifying Solutions .....	280
13.5	Selecting the Best Explanation .....	281
13.5.1	Ordering Explanations .....	282
13.6	Integrating the Best Solution .....	284
13.6.1	Abduction in a Picture, Once Again .....	285
13.6.2	Further Classification .....	285
13.6.3	Properties in a Picture .....	287
13.7	Working with the Explanations .....	287
13.7.1	A Modality .....	288
13.8	A Brief Exploration to Nonideal Agents .....	289
13.8.1	Considering Inference .....	290
13.8.2	Different Reasoning Abilities .....	290
13.9	Conclusions .....	290
	<b>References</b> .....	292
<b>14</b>	<b>Argumentation and Abduction in Dialogical Logic</b>	
	<i>Cristina Barés Gómez, Matthieu Fontaine</i> .....	295
14.1	Reasoning as a Human Activity .....	295
14.2	Logic and Argumentation: The Divorce .....	297
14.3	Logic and Argumentation: A Reconciliation .....	299

14.3.1	What is Dialogical Logic? .....	299
14.3.2	Particle Rules .....	300
14.3.3	Structural Rules .....	301
14.3.4	Winning Strategy and Validity .....	302
14.4	Beyond Deductive Inference: Abduction .....	303
14.4.1	The GW Model of Abduction .....	303
14.5	Abduction in Dialogical Logic.....	306
14.5.1	Triggering.....	306
14.5.2	Guessing .....	308
14.5.3	Committing.....	309
14.6	Hypothesis: What Kind of Speech Act?.....	310
14.7	Conclusions .....	312
	<b>References</b> .....	312
<b>15</b>	<b>Formal (In)consistency, Abduction and Modalities</b>	
	<i>Juliana Bueno-Soler, Walter Carnielli, Marcelo E. Coniglio,</i>	
	<i>Abilio Rodrigues Filho</i> .....	315
15.1	Paraconsistency.....	315
15.2	Logics of Formal Inconsistency .....	316
15.2.1	mbC: A Minimal LFI .....	318
15.2.2	A Logic of Evidence and Truth .....	321
15.3	Abduction .....	322
15.3.1	mbC-Tableaux.....	324
15.3.2	Quantification .....	326
15.4	Modality.....	327
15.4.1	The Anodic System $K^\diamond$ .....	328
15.4.2	The Logic $mbC^\square$ .....	329
15.4.3	Extensions of $mbC^\square$ .....	330
15.5	On Alternative Semantics for $mbC$ .....	331
15.6	Conclusions .....	333
	<b>References</b> .....	334
<b>Part D Model-Based Reasoning in Science and the History of Science</b>		
<b>16</b>	<b>Metaphor and Model-Based Reasoning in Mathematical Physics</b>	
	<i>Ryan D. Tweney</i> .....	341
16.1	Cognitive Tools for Interpretive Understanding.....	343
16.1.1	Model-Based Reasoning.....	343
16.1.2	Metaphoric Processes.....	344
16.1.3	Long Term Working Memory .....	344
16.2	Maxwell's Use of Mathematical Representation .....	345
16.2.1	From Faraday to Maxwell .....	345
16.2.2	Faraday: Magnetic Lines Within a Magnet.....	346
16.2.3	Maxwell: Magnetic Lines Within a Magnet .....	347
16.3	Unpacking the Model-Based Reasoning .....	348
16.4	Cognition and Metaphor in Mathematical Physics .....	350
16.5	Conclusions .....	351
	<b>References</b> .....	352

<b>17 Nancy Nersessian's Cognitive–Historical Approach</b>	
<i>Nora Alejandrina Schwartz</i> .....	355
17.1 Questions About the Creation of Scientific Concepts .....	356
17.1.1 The Problem of Conceptual Change .....	356
17.1.2 The Naturalistic Approach to Science: Revision of the Problem .....	357
17.1.3 The Naturalistic Recasting .....	357
17.2 The Epistemic Virtues of Cognitive Historical Analysis .....	359
17.2.1 The Cognitive–Historical Approach .....	359
17.2.2 Epistemic Virtues and Dimensions of this Approach .....	360
17.2.3 Cognitive Methods to Investigate Conceptual Innovation .....	362
17.3 Hypothesis About the Creation of Scientific Concepts .....	363
17.3.1 Dynamic Hypothesis .....	364
17.3.2 The Power of Model–Based Reasoning .....	367
17.4 Conclusions .....	373
<b>References</b> .....	373
<b>18 Physically Similar Systems – A History of the Concept</b>	
<i>Susan G. Sterrett</i> .....	377
18.1 Similar Systems, the Twentieth Century Concept .....	379
18.2 Newton and Galileo .....	380
18.2.1 Newton on Similar Systems .....	380
18.2.2 Galileo .....	381
18.3 Late Nineteenth and Early Twentieth Century .....	383
18.3.1 Engineering and Similarity <i>Laws</i> .....	384
18.3.2 Similar Systems in Theoretical Physics: Lorentz, Boltzmann, van der Waals, and Onnes .....	386
18.3.3 Similar Systems in Theoretical Physics .....	391
18.4 1914: The Year of <i>Physically Similar Systems</i> .....	397
18.4.1 Overview of Relevant Events of the Year 1914 .....	398
18.4.2 Stanton and Pannell .....	398
18.4.3 Buckingham and Tolman .....	399
18.4.4 Precursors of the <i>Pi-Theorem</i> in Buckingham's 1914 Papers .....	406
18.5 Physically Similar Systems: The Path in Retrospect .....	408
<b>References</b> .....	409
<b>19 Hypothetical Models in Social Science</b>	
<i>Alessandra Basso, Chiara Lisciandra, Caterina Marchionni</i> .....	413
19.1 Hypothetical Modeling as a Style of Reasoning .....	413
19.2 Models Versus Experiments: Representation, Isolation and Resemblance .....	416
19.3 Models and Simulations: Complexity, Tractability and Transparency	420
19.4 Epistemology of Models .....	423
19.4.1 Instrumentalism and Predictive Ability .....	424
19.4.2 Isolation of Causal Mechanisms or Capacities .....	424
19.4.3 Learning About Possibilities .....	425
19.4.4 Inferential Aids .....	426
19.4.5 Models as Blueprints for the Design of Socio–Economic Mechanisms .....	427

19.4.6	Where Do We Go From Here? .....	428
19.5	Conclusions .....	428
19.A	Appendix: J.H. von Thünen's Model of Agricultural Land Use in the Isolated State .....	429
19.B	Appendix: T. Schelling's Agent-Based Model of Segregation in Metropolitan Areas .....	430
	<b>References</b> .....	431
<b>20</b>	<b>Model-Based Diagnosis</b>	
	<i>Antoni Ligeza, Bartłomiej Górny</i> .....	435
20.1	A Basic Model for Diagnosis .....	437
20.2	A Review and Taxonomy of Knowledge Engineering Methods for Diagnosis .....	438
20.2.1	Knowledge Engineering .....	438
20.2.2	Expert Methods .....	439
20.2.3	Model-Based Methods .....	439
20.3	Model-Based Diagnostic Reasoning .....	440
20.4	A Motivation Example .....	440
20.5	Theory of Model-Based Diagnosis .....	442
20.6	Causal Graphs .....	444
20.7	Potential Conflict Structures .....	446
20.8	Example Revisited. A Complete Diagnostic Procedure .....	448
20.9	Refinement: Qualitative Diagnoses .....	450
20.9.1	Qualitative Evaluation of Faults .....	450
20.9.2	Elimination of Spurious Diagnoses .....	450
20.9.3	Deduction for Enhanced Diagnoses .....	452
20.9.4	Analysis of Diagnoses .....	452
20.10	Dynamic Systems Diagnosis: The Three-Tank Case .....	454
20.11	Incremental Diagnosis .....	456
20.12	Practical Example and Tools .....	458
20.13	Concluding Remarks .....	459
	<b>References</b> .....	460
<b>21</b>	<b>Thought Experiments in Model-Based Reasoning</b>	
	<i>Margherita Arcangeli</i> .....	463
21.1	Overview .....	464
21.1.1	Galileo on Falling Bodies .....	464
21.1.2	Stevin's Chain Thought Experiment .....	465
21.1.3	Newton's Bucket Thought Experiment .....	465
21.1.4	Gettier's Thought Experiment .....	466
21.1.5	Twin Earth .....	466
21.1.6	Mary the Super-Scientist .....	467
21.2	Historical Background .....	467
21.2.1	The Rise of the Term .....	467
21.2.2	The Classical Phase .....	468
21.2.3	The Contemporary Phase .....	469
21.3	What Is a Thought Experiment? .....	469
21.3.1	Thought Experiments and the Experimental Realm .....	470
21.3.2	Thought Experiments and the Theoretical Realm .....	472
21.3.3	Thought Experiments and Their Features .....	473

21.4	What Is the Function of Thought Experiments? .....	475
21.4.1	Sorting Thought Experiments .....	476
21.4.2	Thought Experiments and Kinds of Knowledge .....	478
21.4.3	The Epistemological Status of Thought Experiments .....	480
21.5	How Do Thought Experiments Achieve Their Function? .....	484
21.5.1	A Cognitive Approach to Thought Experimentation .....	484
21.5.2	Imagination and Thought Experimentation .....	485
21.5.3	The Narrative Dimension of Thought Experimentation ....	486
	<b>References</b> .....	487

## Part E Models in Mathematics

### 22 Diagrammatic Reasoning in Mathematics

	<i>Valeria Giardino</i> .....	499
22.1	Diagrams as Cognitive Tools .....	499
22.2	Diagrams and (the Philosophy of) Mathematical Practice .....	501
22.3	The Euclidean Diagram .....	503
22.3.1	The (Greek) Lettered Diagram .....	504
22.3.2	Exact and Co-Exact Properties .....	505
22.3.3	Reasoning in the Diagram .....	506
22.3.4	Concrete Diagrams and Quasi-Concrete Geometrical Objects .....	508
22.4	The Productive Ambiguity of Diagrams .....	509
22.5	Diagrams in Contemporary Mathematics .....	510
22.5.1	Analysis .....	511
22.5.2	Algebra .....	513
22.5.3	Topology .....	514
22.6	Computational Approaches .....	515
22.6.1	(Manders') Euclid Reloaded .....	516
22.6.2	Theorem Provers .....	517
22.7	Mathematical Thinking: Beyond Binary Classifications .....	518
22.8	Conclusions .....	520
	<b>References</b> .....	521

### 23 Deduction, Diagrams and Model-Based Reasoning

	<i>John Mumma</i> .....	523
23.1	Euclid's Systematic Use of Geometric Diagrams .....	524
23.2	Formalizing Euclid's Diagrammatic Proof Method .....	525
23.2.1	The Formal System <b>FG</b> .....	526
23.2.2	The Formal System <b>Eu</b> .....	528
23.3	Formal Geometric Diagrams as Models .....	532
	<b>References</b> .....	534

### 24 Model-Based Reasoning in Mathematical Practice

	<i>Joachim Frans, Isar Goyvaerts, Bart Van Kerkhove</i> .....	537
24.1	Preliminaries .....	537
24.2	Model-Based Reasoning: Examples .....	538
24.2.1	First Example: From Euclidean Geometry .....	538
24.2.2	Second Example: From Approximation Theory .....	539
24.2.3	Third Example: From Category Theory .....	540

24.3	The Power of Heuristics and Plausible Reasoning .....	540
24.4	Mathematical Fruits of Model-Based Reasoning .....	542
24.5	Conclusion .....	546
24.A	Appendix.....	546
	<b>References</b> .....	548

## 25 Abduction and the Emergence of Necessary Mathematical Knowledge

	<i>Ferdinand Rivera</i> .....	551
25.1	An Example from the Classroom .....	551
25.2	Inference Types .....	555
	25.2.1 Abduction .....	557
	25.2.2 Induction.....	558
	25.2.3 Deduction and Deductive Closure .....	559
25.3	Abduction in Math and Science Education.....	561
	25.3.1 Different Kinds of Abduction .....	561
	25.3.2 Abduction in Mathematical Relationships .....	562
25.4	Enacting Abductive Action in Mathematical Contexts .....	564
	25.4.1 Cultivate Abductively-Infused Guesses with Deduction ...	564
	25.4.2 Support Logically-Good Abductive Reasoning .....	565
	25.4.3 Foster the Development of Strategic Rules in Abductive Processing.....	565
	25.4.4 Encourage an Abductive Knowledge-Seeking Disposition	565
	<b>References</b> .....	566

## Part F Model-Based Reasoning in Cognitive Science

### 26 Vision, Thinking, and Model-Based Inferences

	<i>Athanasios Raftopoulos</i> .....	573
26.1	Inference and Its Modes .....	576
26.2	Theories of Vision .....	577
	26.2.1 Constructivism .....	577
	26.2.2 Theory of Direct Vision or Ecological Theory of Visual Perception .....	580
	26.2.3 Predictive Visual Brain: Vision and Action.....	581
26.3	Stages of Visual Processing .....	585
	26.3.1 Early Vision.....	585
	26.3.2 Late Vision .....	586
26.4	Cognitive Penetrability of Perception and the Relation Between Early Vision and Thinking.....	588
	26.4.1 The Operational Constraints in Visual Processing .....	589
	26.4.2 Perceptual Learning .....	590
26.5	Late Vision, Inferences, and Thinking .....	591
	26.5.1 Late Vision, Hypothesis Testing, and Inference .....	593
	26.5.2 Late Vision and Discursive Understanding.....	594
26.6	Concluding Discussion .....	596
26.A	Appendix: Forms of Inferences.....	597
	26.A.1 Deduction .....	597
	26.A.2 Induction.....	597
	26.A.3 Abduction or Inference to the Best Explanation .....	597

26.A.4	Differences Between the Modes of Inference .....	598
26.B	Appendix: Constructivism .....	598
26.C	Appendix: Bayes' Theorem and Some of Its Epistemological Aspects .....	600
26.D	Appendix: Modal and Amodal Completion or Perception.....	600
26.E	Appendix: Operational Constraints in Visual Processing .....	601
	<b>References</b> .....	602
<b>27</b>	<b>Diagrammatic Reasoning</b>	
	<i>William Bechtel</i> .....	605
27.1	Cognitive Affordances of Diagrams and Visual Images .....	606
27.2	Reasoning with Data Graphs .....	608
	27.2.1 Data Graphs in Circadian Biology .....	608
	27.2.2 Cognitive Science Research Relevant to Reasoning with Graphs.....	611
27.3	Reasoning with Mechanism Diagrams .....	613
	27.3.1 Mechanism Diagrams in Circadian Biology .....	613
	27.3.2 Cognitive Science Research Relevant to Reasoning with Mechanism Diagrams.....	615
27.4	Conclusions and Future Tasks .....	616
	<b>References</b> .....	617
<b>28</b>	<b>Embodied Mental Imagery in Cognitive Robots</b>	
	<i>Alessandro Di Nuovo, Davide Marocco, Santo Di Nuovo, Angelo Cangelosi</i> . .....	619
28.1	Mental Imagery Research Background .....	620
28.2	Models and Approaches Based on Mental Imagery in Cognitive Systems and Robotics .....	622
28.3	Experiments .....	624
	28.3.1 The Humanoid Robotic Platform: The iCub .....	624
	28.3.2 First Experimental Study: Motor Imagery for Performance Improvement .....	624
	28.3.3 Second Experimental Study: Mental Training Evoked by Language .....	628
	28.3.4 Third Experimental Study: Spatial Imagery.....	630
28.4	Conclusion .....	635
	<b>References</b> .....	635
<b>29</b>	<b>Dynamical Models of Cognition</b>	
	<i>Mary Ann Metzger</i> .....	639
29.1	Dynamics .....	639
	29.1.1 Time and Complexity .....	640
	29.1.2 Cognition and Action .....	641
29.2	Data-Oriented Models .....	641
	29.2.1 Methods .....	641
	29.2.2 Example: Motor Coordination .....	642
	29.2.3 Example: Decision Under Risk .....	643
	29.2.4 Summary .....	644
29.3	Cognition and Action Distinct .....	644
	29.3.1 Recognition Memory Model.....	644
	29.3.2 Adaptive Control of Thought – Rational .....	645
	29.3.3 Artificial Neural Networks Methods .....	646
	29.3.4 Adaptive Resonance Theory .....	647

29.3.5	Summary .....	648
29.4	Cognition and Action Intrinsically Linked .....	648
29.4.1	Methods .....	648
29.4.2	Embodied Cognition .....	650
29.4.3	Motor Theory .....	651
29.4.4	Simulation Theory .....	651
29.4.5	Free Energy Theory .....	652
29.4.6	Evolution of Cognitive Search .....	653
29.4.7	Summary .....	653
29.5	Conclusion .....	653
	<b>References</b> .....	655
<b>30</b>	<b>Complex versus Complicated Models of Cognition</b>	
	<i>Ruud J.R. Den Hartigh, Ralf F.A. Cox, Paul L.C. Van Geert</i> .....	657
30.1	Current Views on Cognition .....	658
30.1.1	Central Control versus Self-Organization .....	658
30.1.2	Static versus Dynamic Models .....	659
30.2	Explaining Cognition .....	660
30.2.1	Research Strategies and Complicated Models .....	660
30.2.2	Research Strategies and Complex Models .....	660
30.2.3	Analyses to Untangle Cognition Based on Complicated Models .....	661
30.2.4	Analyses to Capture the Complexity of Cognition .....	661
30.3	Is Cognition Best Explained by a Complicated or Complex Model? ..	662
30.3.1	Explaining Real-Time Cognitive Performance .....	662
30.3.2	Explaining Long-Term Cognitive Development .....	663
30.4	Conclusion .....	666
	<b>References</b> .....	666
<b>31</b>	<b>From Neural Circuitry to Mechanistic Model-Based Reasoning</b>	
	<i>Jonathan Waskan</i> .....	671
31.1	Mechanistic Reasoning in Science .....	672
31.2	The Psychology of Model-Based Reasoning .....	673
31.3	Mental Models in the Brain: Attempts at Psycho-Neural Reduction .....	675
31.3.1	From Structural to Functional Isomorphism .....	676
31.3.2	Distinctive Features of Scale Models .....	678
31.3.3	Does Computational Realization Entail Sentential Representation? .....	681
31.3.4	What About POPI? .....	682
31.3.5	Bridging the Divide .....	684
31.3.6	Bottom-Up Approaches .....	685
31.4	Realization Story Applied .....	686
31.4.1	AI and Psychology: Towards an Intuitive Physics Engine ..	686
31.4.2	Exduction .....	687
31.5	Mechanistic Explanation Revisited .....	687
31.5.1	The Prediction and Ceteris Paribus Problems .....	688
31.5.2	Beyond Mental Models .....	689
31.6	Conclusion .....	690
	<b>References</b> .....	690



## Part G Modelling and Computational Issues

<b>32 Computational Aspects of Model-Based Reasoning</b>	
<i>Gordana Dodig-Crnkovic, Antonio Cicchetti</i> .....	695
32.1 Computational Turn Seen from Different Perspectives .....	695
32.2 Models of Computation .....	697
32.2.1 Turing Model of Computation and Its Scope .....	698
32.2.2 Computation as Information Processing.....	698
32.3 Computation Versus Information.....	700
32.4 The Difference Between Mathematical and Computational (Executable) Models .....	702
32.5 Computation in the Wild .....	703
32.5.1 Physical Computation – Computing Nature as Info-Computation .....	703
32.5.2 New Computationalism. Nonsymbolic versus Symbolic Computation .....	704
32.6 Cognition: Knowledge Generation by Computation of New Information .....	706
32.6.1 Distributed Cognition and Model-Based Reasoning.....	707
32.6.2 Computational Aspects of Model-Based Reasoning in Science .....	708
32.7 Model-Based Reasoning and Computational Automation of Reasoning.....	709
32.8 Model Transformations and Semantics: Separation Between Semantics and Ontology .....	712
<b>References</b> .....	715
<b>33 Computational Scientific Discovery</b>	
<i>Peter D. Sozou, Peter C.R. Lane, Mark Addis, Fernand Gobet</i> .....	719
33.1 The Roots of Human Scientific Discovery .....	720
33.2 The Nature of Scientific Discovery .....	721
33.3 The Psychology of Human Scientific Discovery .....	722
33.4 Computational Discovery in Mathematics .....	723
33.4.1 Logic Theorist .....	723
33.4.2 AM and EURISKO .....	724
33.4.3 GRAFFITI .....	724
33.5 Methods and Applications in Computational Scientific Discovery ...	725
33.5.1 Massive Systematic Search Within a Defined Space .....	726
33.5.2 Rule-Based Reasoning Systems .....	726
33.5.3 Classification, Machine Vision, and Related Techniques ..	727
33.5.4 Data Mining.....	727
33.5.5 Finding Networks.....	727
33.5.6 Evolutionary Computation .....	728
33.5.7 Automation of Scientific Experiments .....	729
33.6 Discussion.....	730
<b>References</b> .....	731
<b>34 Computer Simulations and Computational Models in Science</b>	
<i>Cyrille Imbert</i> .....	735
34.1 Computer Simulations in Perspective .....	736

34.1.1	The Recent Philosophy of Scientific Models and Computer Simulations .....	736
34.1.2	Numerical Methods and Computational Science: An Old Tradition .....	737
34.1.3	A More or Less Recent Adoption Across Scientific Fields ...	738
34.1.4	Methodological Caveat .....	738
34.2	The Variety of Computer Simulations and Computational Models...	739
34.2.1	Working Characterization .....	739
34.2.2	Analog Simulations and Their Specificities .....	740
34.2.3	Digital Machines, Numerical Physics, and Types of Equivalence .....	741
34.2.4	Non-Numerical Digital Models .....	741
34.2.5	Nondeterministic Simulations .....	742
34.2.6	Other Types of Computer Simulations .....	742
34.3	Epistemology of Computational Models and Computer Simulations	743
34.3.1	Computer Simulations and Their Scientific Roles .....	743
34.3.2	Aspects of the Epistemological Analysis of Computer Simulations.....	744
34.3.3	Selecting Computational Models and Practices .....	746
34.3.4	The Production of 'New' Knowledge: In What Sense? .....	748
34.4	Computer Simulations, Explanation, and Understanding .....	750
34.4.1	Traditional Accounts of Explanation .....	751
34.4.2	Computer Simulations: Intrinsically Unexplanatory? .....	751
34.4.3	Computer Simulations: More Frequently Unexplanatory? ..	752
34.4.4	Too Replete to Be Explanatory? The Era of Lurking Suspicion	754
34.4.5	Bypassing the Opacity of Simulations.....	757
34.4.6	Understanding and Disciplinary Norms.....	758
34.5	Comparing: Computer Simulations, Experiments and Thought Experiments .....	758
34.5.1	Computational Mathematics and the Experimental Stance	759
34.5.2	Common Basal Features.....	759
34.5.3	Are Computer Simulations Experiments? .....	762
34.5.4	Knowledge Production, Superiority Claims, and Empiricism	765
34.5.5	The Epistemological Challenge of Hybrid Methods .....	767
34.6	The Definition of Computational Models and Simulations.....	767
34.6.1	Existing Definitions of Simulations .....	768
34.6.2	Pending Issues .....	770
34.6.3	When Epistemology Cross-Cuts Ontology .....	773
34.7	Conclusion: Human-Centered, but no Longer Human-Tailored Science .....	773
34.7.1	The Partial Mutation of Scientific Practices .....	774
34.7.2	The New Place of Humans in Science .....	774
34.7.3	Analyzing Computational Practices for Their Own Sake....	774
34.7.4	The Epistemological Treatment of New Issues.....	775
	<b>References</b> .....	775
<b>35</b>	<b>Simulation of Complex Systems</b>	
	<i>Paul Davidsson, Franziska Klügl, Harko Verhagen</i> .....	783
35.1	Complex Systems .....	783
35.1.1	Features Associated with Complex Systems .....	784

35.1.2	Summing Up .....	785
35.2	Modeling Complex Systems .....	785
35.2.1	Macro-Level Versus Micro-Level Simulation .....	786
35.2.2	Purpose of Modeling Complex Systems .....	788
35.3	Agent-Based Simulation of Complex Systems .....	789
35.3.1	Elements of Agent-Based Simulation Models .....	790
35.3.2	Engineering Agent-Based Simulations .....	792
35.4	Summing Up and Future Trends .....	795
	<b>References</b> .....	796
<b>36</b>	<b>Models and Experiments in Robotics</b>	
	<i>Francesco Amigoni, Viola Schiaffonati</i> .....	799
36.1	A Conceptual Premise .....	799
36.2	Experimental Issues in Robotics .....	801
36.3	From Experimental Computer Science to Good Experimental Methodologies in Autonomous Robotics .....	802
36.4	Simulation .....	804
36.5	Benchmarking and Standards .....	807
36.6	Competitions and Challenges .....	809
36.7	Conclusions .....	812
	<b>References</b> .....	812
<b>37</b>	<b>Biorobotics</b>	
	<i>Edoardo Datteri</i> .....	817
37.1	Robots as Models of Living Systems .....	817
37.1.1	Data-Oriented and Model-Oriented Simulations .....	817
37.1.2	The Structure of Biorobotic Methodology .....	819
37.2	A Short History of Biorobotics .....	825
37.2.1	Cybernetic and Artificial Intelligence .....	825
37.2.2	Contemporary Invertebrate and Vertebrate Simula- tion Studies .....	826
37.3	Methodological Issues .....	826
37.3.1	The Epistemic Requirements of <i>Good</i> Biorobots .....	826
37.3.2	On the Meaning of Behavior .....	830
37.3.3	Robots and Their Environment: Robotic versus Computer Simulations .....	832
37.4	Conclusions .....	833
	<b>References</b> .....	834
 <b>Part H Models in Physics, Chemistry and Life Sciences</b>		
<b>38</b>	<b>Comparing Symmetries in Models and Simulations</b>	
	<i>Giuseppe Longo, Maël Montévil</i> .....	843
38.1	Approximation .....	844
38.2	What Do Equations and Computations Do? .....	845
38.2.1	Equations .....	845
38.2.2	From Equations to Computations .....	845
38.2.3	Computations .....	847
38.3	Randomness in Biology .....	848
38.4	Symmetries and Information in Physics and Biology .....	849

38.4.1	Turing, Discrete State Machines and Continuous Dynamics	849
38.4.2	Classifying Information .....	851
38.5	Theoretical Symmetries and Randomness .....	852
	<b>References</b> .....	854
<b>39</b>	<b>Experimentation on Analogue Models</b>	
	<i>Susan G. Sterrett</i> .....	857
39.1	Analogue Models: Terminology and Role .....	858
39.1.1	Analogue Models and Scale Models .....	858
39.1.2	The Role of Analogue Models in Philosophy of Science ...	860
39.1.3	Analogue Models in History of Science .....	861
39.2	Analogue Models in Physics .....	868
39.2.1	Lessons from the Nineteenth Century .....	868
39.2.2	Sound as an Analogue of Light: The Power of Experimentation on Analogues .....	868
39.2.3	Water as an Analogue of Electricity: Limitations of Generalizing from Analogues .....	869
39.2.4	Some Recent Results Using Analogue Models .....	870
39.3	Comparing Fundamental Bases for Physical Analogue Models .....	873
39.3.1	Three Kinds of Bases for Physical Analogue Models .....	875
39.4	Conclusion .....	876
	<b>References</b> .....	877
<b>40</b>	<b>Models of Chemical Structure</b>	
	<i>William Goodwin</i> .....	879
40.1	Models, Theory, and Explanations in Structural Organic Chemistry .	881
40.2	Structures in the Applications of Chemistry .....	883
40.3	The Dynamics of Structure .....	885
40.3.1	Recognizing the Importance of Conformations .....	886
40.3.2	Using Conformations in Organic Chemistry .....	887
40.4	Conclusion .....	889
	<b>References</b> .....	889
<b>41</b>	<b>Models in Geosciences</b>	
	<i>Alisa Bokulich, Naomi Oreskes</i> .....	891
41.1	What Are Geosciences? .....	891
41.2	Conceptual Models in the Geosciences .....	892
41.3	Physical Models in the Geosciences .....	893
41.4	Numerical Models in the Geosciences .....	895
41.5	Bringing the Social Sciences Into Geoscience Modeling .....	897
41.6	Testing Models: From Calibration to Validation .....	898
41.6.1	Data and Models .....	898
41.6.2	Parametrization, Calibration, and Validation .....	899
41.6.3	Sensitivity Analysis and Other Model Tests .....	901
41.7	Inverse Problem Modeling .....	902
41.8	Uncertainty in Geoscience Modeling .....	903
41.9	Multimodel Approaches in Geosciences .....	907
41.10	Conclusions .....	908
	<b>References</b> .....	908

<b>42 Models in the Biological Sciences</b>	
<i>Elisabeth A. Lloyd</i> .....	913
42.1 Evolutionary Theory .....	913
42.1.1 The Structure of Darwinian Evolutionary Models .....	913
42.1.2 The Structure of Population Genetic Evolutionary Models .....	914
42.1.3 Representational Adequacy of Models.....	918
42.1.4 Expansions and Alternative Views of the Structure of Evolutionary Theory .....	920
42.2 Confirmation in Evolutionary Biology .....	922
42.2.1 Confirming and Testing Models .....	922
42.3 Models in Behavioral Evolution and Ecology.....	925
42.3.1 The Phenotypic Gambit .....	925
42.3.2 Evolutionary Stable Strategies, Animal Signalling .....	925
42.3.3 Physiological/Evolution Models, Cognitive Ethology .....	926
42.3.4 Optimality Models Including Agent Models .....	927
<b>References</b> .....	927
<b>43 Models and Mechanisms in Cognitive Science</b>	
<i>Massimo Marraffa, Alfredo Paternoster</i> .....	929
43.1 What is a Model in Cognitive Science?.....	929
43.1.1 Computational Models .....	929
43.1.2 An Example of Computational Model .....	932
43.1.3 Function and Functional Explanation .....	934
43.1.4 Computational Models and Mechanistic Explanations ....	936
43.1.5 Dynamical Systems .....	939
43.2 Open Problems in Computational Modeling .....	940
43.2.1 Computationalism and Central Cognition .....	940
43.2.2 The Dynamicist Challenge: Is Integration Possible?.....	944
43.3 Conclusions .....	948
<b>References</b> .....	949
<b>44 Model-Based Reasoning in the Social Sciences</b>	
<i>Federica Russo</i> .....	953
44.1 Modeling Practices in the Social Sciences .....	954
44.1.1 Social-Scientific Objects.....	954
44.1.2 Quantitative Modeling .....	955
44.1.3 Qualitative Modeling .....	957
44.1.4 Experimental and Quasi-Experimental Modeling .....	957
44.2 Concepts of Model .....	958
44.2.1 Models as Representations .....	958
44.2.2 Models as Objects.....	960
44.3 Models and Reality .....	962
44.3.1 Mediators .....	962
44.3.2 Isolations.....	962
44.3.3 Maps.....	963
44.4 Models and Neighboring Concepts.....	963
44.4.1 Simulations .....	963
44.4.2 Causation and Explanation .....	964
44.4.3 Truth and Validity .....	965
44.5 Conclusion .....	967
<b>References</b> .....	968

## Part I Models in Engineering, Architecture, and Economical and Human Sciences

<b>45 Models in Architectural Design</b>	
<i>Pieter Pauwels</i> .....	975
45.1 Architectural Design Thinking .....	976
45.1.1 The Architectural Designer as a Practitioner .....	976
45.1.2 Where Are the Models in all This? .....	976
45.1.3 Abstraction, Sense-Making, and Framing into Mental Models .....	978
45.1.4 Accessing Background Knowledge Through Analogical Reasoning? .....	979
45.1.5 Abstraction from Representation Model to Mental Model .....	980
45.2 BIM Models and Parametric Models .....	981
45.2.1 New Technological Media in Design Thinking .....	981
45.2.2 BIM Models and Parametric Models .....	982
45.2.3 Features and Issues in the Usage of the New Modeling Applications .....	982
45.3 Implementing and Using ICT for Design and Construction .....	984
45.3.1 Pragmatic Usage of Semantic Modeling Applications .....	984
45.3.2 The Usage of Design Agents or Assistants .....	985
<b>References</b> .....	987
<b>46 Representational and Experimental Modeling in Archaeology</b>	
<i>Alison Wylie</i> .....	989
46.1 Philosophical Resources and Archaeological Parallels .....	990
46.2 The Challenges of Archaeological Modeling .....	991
46.3 A Taxonomy of Archaeological Models .....	992
46.3.1 Phenomenological Models of Archaeological Subject and Source Data .....	992
46.3.2 Scaffolding Models: Measurement Tools and Guides to Interpretation .....	995
46.3.3 Reconstructive and Explanatory Models .....	996
46.4 Conclusions .....	1000
<b>References</b> .....	1000
<b>47 Models and Ideology in Design</b>	
<i>Cameron Shelley</i> .....	1003
47.1 Design and Ideology .....	1003
47.2 Models and Ideology .....	1004
47.3 Revivalism: Looking to the Past .....	1005
47.4 Modernism: Transcending History .....	1006
47.5 Industrial Design: The Shape of Things to Come .....	1009
47.6 Biomimicry .....	1011
47.7 Conclusion .....	1013
<b>References</b> .....	1013
<b>48 Restructuring Incomplete Models in Innovators Marketplace on Data Jackets</b>	
<i>Yukio Ohsawa, Teruaki Hayashi, Hiroyuki Kido</i> .....	1015
48.1 Chance Discovery as a Trigger to Innovation .....	1016

48.2	Chance Discovery from Data and Communication .....	1016
48.2.1	Chance Discovery as a Problematic Child of Data Mining..	1016
48.3	IM for Externalizing and Connecting Requirements and Solutions .	1020
48.4	Innovators Marketplace on Data Jackets .....	1022
48.4.1	Marketplaces of Data .....	1022
48.4.2	The Procedure of IMDJ .....	1022
48.5	IMDJ as Place for Reasoning on Incomplete Models .....	1023
48.5.1	Grounding Incompletely Defined Models Into Well-Defined Models .....	1023
48.5.2	Abductive Reasoning for Thoughts and Communications in IMDJ .....	1025
48.6	Conclusions .....	1029
	<b>References</b> .....	1029
<b>49</b>	<b>Models in Pedagogy and Education</b>	
	<i>Flavia Santoianni</i> .....	1033
49.1	Pluralism .....	1034
49.1.1	Theoretical Plurality .....	1034
49.1.2	Multidisciplinary Plurality .....	1035
49.1.3	Disciplinary Multiplicity .....	1036
49.2	Dialecticity .....	1039
49.2.1	Science and Philosophy .....	1039
49.2.2	Theory and Practice .....	1040
49.3	Applied Models .....	1042
49.3.1	Traditional Models .....	1042
49.3.2	Actual Models .....	1044
49.3.3	Experimental Models .....	1046
49.4	Conclusions .....	1048
	<b>References</b> .....	1048
<b>50</b>	<b>Model-Based Reasoning in Crime Prevention</b>	
	<i>Charlotte Gerritsen, Tibor Bosse</i> .....	1051
50.1	Ambient Intelligence .....	1053
50.2	Methodology .....	1054
50.3	Domain Model .....	1055
50.3.1	Crime Displacement .....	1055
50.3.2	Formalization .....	1056
50.4	Analysis Model .....	1058
50.5	Support Model .....	1060
50.6	Results .....	1060
50.7	Discussion .....	1062
	<b>References</b> .....	1062
<b>51</b>	<b>Modeling in the Macroeconomics of Financial Markets</b>	
	<i>Giovanna Magnani</i> .....	1065
51.1	The Intrinsic Instability of Financial Markets .....	1066
51.1.1	The Interpretation of the General Theory .....	1066
51.1.2	The Nature of the Capitalist System .....	1068
51.1.3	Cash Flows Analysis and Classification of Financial Postures .....	1069
51.2	The Financial Theory of Investment .....	1071
51.2.1	Aggregate Profit Determination .....	1071

51.2.2	The Two-Price Model and the Determination of Investment .....	1072
51.3	The Financial Instability Hypothesis Versus the Efficient Markets Hypothesis .....	1074
51.4	Irving Fisher's Debt-Deflation Model .....	1074
51.4.1	Debt Deflation as a Cycle Theory .....	1075
51.4.2	How Debt Deflation Model Fits the Great Depression .....	1077
51.4.3	How Debt-Deflation Model Fits Current Economic Conditions .....	1077
51.5	Policy Implications and the Shareholder Maximization Value Model .....	1079
51.5.1	Stability is Destabilizing .....	1079
51.5.2	From the Debt Deflation Model to Policy Proposals .....	1080
51.5.3	Financialization, Neoliberalization, and the 2008 Crisis..	1081
51.6	Integrating the Minskyian Model with New Marxists and Social Structure of Accumulation (SSA) Theories .....	1085
51.7	Risk and Uncertainty .....	1086
51.7.1	Models of Risk and Uncertainty in Economics and Business Studies .....	1086
51.7.2	Models of Risk .....	1092
51.7.3	Models of Uncertainty .....	1093
	<b>References</b> .....	1098
<b>52</b>	<b>Application of Models from Social Science to Social Policy</b>	
	<i>Eleonora Montuschi</i> .....	1103
52.1	Unrealistic Assumptions .....	1105
52.1.1	Quality of Life .....	1105
52.1.2	QALY: What Are We Measuring? .....	1106
52.1.3	Unrealistic Utility Assumptions .....	1107
52.2	Real Experiments, Not Models Please! .....	1110
52.2.1	Causal Models: How Can They Come to the Rescue .....	1111
52.2.2	Class Size Reduction Programmes .....	1111
52.2.3	TINP and BINP .....	1112
52.2.4	Children Mortality and Inflicted Death .....	1114
52.3	Conclusions .....	1115
	<b>References</b> .....	1116
<b>53</b>	<b>Models and Moral Deliberation</b>	
	<i>Cameron Shelley</i> .....	1117
53.1	Rules .....	1118
53.2	Mental Models .....	1119
53.3	Schemata .....	1121
53.4	Analogy .....	1122
53.5	Empathy .....	1124
53.6	Role Models .....	1125
53.7	Discussion .....	1126
	<b>References</b> .....	1127
	<b>About the Authors</b> .....	1129
	<b>Detailed Contents</b> .....	1141
	<b>Subject Index</b> .....	1163



## Subject Index

2-D CAD model 981  
3-D BIM model 981

### A

- abduction 138, 152, 175, 179,  
219–221, 228, 296, 303–306,  
308–311, 322, 343, 436, 448, 551,  
555, 575–577, 590, 594, 596–598
- argument 219, 221–223, 228
  - automatic explanations 326
  - conceptual 254
  - creative 189, 191, 561
  - existential 253
  - explanatory 192
  - factual 182, 190
  - first-order existential 191
  - generalization 253
  - guessing 557
  - hypothetical (common) cause 192
  - iconic 191
  - inference to the best explanation 219–222, 228
  - instrumental 193
  - inverse 188
  - logic program 323
  - logical approach 222
  - logical characterization 225
  - logical formulation 221
  - manipulative 182, 191, 561
  - mbC-tableaux 324
  - nonexplanatory 192
  - observable-fact 191
  - of generalization 258
  - overcoded 561
  - pattern 220, 265
  - selective 179, 189, 191
  - singular fact 253
  - tableaux 323
  - taxonomy 223
  - theoretical 180, 190, 561
  - theoretical model 182, 191
  - to reality 168
  - trans-paradigmatic 191
  - undercoded 561
  - unobservable-fact 191
  - visual 191
- abductive 574
- action 554
  - anomaly 224, 228
  - belief expansion 169
  - belief revision 169
  - dialogue 306, 309, 312
  - expansion 227
  - explanation 224
  - framework 555
  - generalization 558
  - guessing 199
  - inference 459, 576–578, 586,  
592, 594–597
  - instinct 153, 210
  - logic programming (ALP) 226,  
986
  - novelty 224, 228
  - outcome 224
  - parameter 224
  - processing 555, 565
  - reasoning 249, 250, 255, 565,  
979, 1025
  - revision 227
  - searching 199
  - trigger 224
- abductive problem 224–227, 271,  
273, 278, 307–309
- anomalous 279
  - expected 279
  - novel 279
- abductive solution 226, 271, 273,  
279
- adequate 286
  - consistent 280
  - explanatory 281
  - neutral 286
  - strong 286
  - successful 286
  - trivial 281
  - weakly explanatory 286
- abductively-infused guesses 564
- abnormal behavior (AB) 442
- abstract
- concept 188
  - explanation 820
  - structure 959
- abstraction 40, 180, 538, 820, 831,  
962
- precise 182
  - subjectal 182
- acting on conclusions 256
- action at a distance 342
- action-based 205
- action-practical component 472
- active externalism 945
- actogram 609
- actual model
- constructivist 1046
  - contextualist 1045
  - culturalist 1045
- adaptive
- model 1047
  - strategy 232, 256
- Adaptive Control of  
Thought-Rational (ACT-R) 645,  
791
- adaptive logic (AL) 158, 232, 251,  
255, 265
- program 252
  - standard format 256, 265
- adding premises 259, 262
- affine transformation 381
- affinity 381
- affinity diagram 1021
- affirming the consequent 261
- affordance 210, 934
- age of computer simulation 805
- agency 789
- agent-based
- model (ABM) 121, 421, 790, 898
  - simulation (ABS) 783, 789, 794,  
1052
- Alchourrón, Gärdenfors and  
Makinson (AGM) 169, 291
- algorithm execution 699
- altruism 1122
- ambient
- agent model 1054
  - intelligence (AmI) 1052
- ambiguity 499, 508–510, 515, 520,  
1091
- amoeba-based computing (ABC)  
858

- ampliative logic 251
  - analog
    - computability 740
    - computer 738–740
    - experiment 866
    - gravity 868, 875
    - model 866, 872, 875
    - simulation 740
  - analogical
    - abduction 161, 254
    - argument 996
    - reasoning 766, 979
  - analogous measurement 403
  - analogue
    - gravity 857
    - model 857, 860, 865, 871
    - space-time 858, 868
  - analogy 10, 177, 343, 349, 386, 857, 868–872, 875, 1020, 1125
    - emotional coherence 1123
    - structural coherence 1123
  - analysis
    - model 1052
    - of qualitative diagnoses 452
  - analytical
    - argument 563
    - sociology 964
  - analytic–synthetic distinction 29
  - AND/OR causal graph 449
  - animal 209
    - abduction 197, 209
    - artifactual mediator 210
  - animat approach 832
  - A-not-B error 650
  - anthropocentric predicament 773
  - anthropology 954, 957
  - anticipation 651
  - antirealism 962
  - applicability of mathematics
    - condition 54
  - apposite bridge 979
  - appositional reasoning 980
  - approximation 124
  - approximation theory 539
  - archaeological model 990
  - Archimedes 499
  - architectural design 975
  - architecture 1005, 1012
  - architecture, engineering and construction (AEC) 982
  - argument 469
  - artifact 211
  - artificial
    - apparatus 205
    - intelligence (AI) 155, 220, 322, 436, 593, 671, 679–681, 684, 689, 726, 803, 825, 929, 1052
    - intelligence (AI) strong 730
    - lesion 823
    - life 471
    - neural network (ANN) 619, 902, 929
    - translation 828
  - as-built model 975
  - asset 1072
  - associational model 964
  - asymmetry between prediction and explanation 968
  - atmospheric science 898
  - atomic hypothesis 161
  - attractiveness 1055
  - attractor 640, 661
  - attribute 1018
  - Auguste Daubrée 893
  - Augustus Welby Northmore Pugin 1005
  - authenticity 1005
  - autogenic behavior 906
  - Automated Mathematician (AM) 724
  - autonomous
    - agents and multiagent systems (AAMAS) 803
    - mobile robot 806
    - robotics 799, 802
  - autonomy 789, 962
  - avoiding random hypothesis 260
  - awareness 1018
  - axiom 959
  - axiomatization 15, 545
- 
- B**
- background knowledge 252, 1111
  - backpropagation through time (BPTT) 632
  - backward-chaining 157
  - BACON 725
  - balance of force 892
  - bank 1068
  - bar graph 611
  - basic card 1020
  - basic property of determinedness (bd) 321
  - Bayesian 648
    - analysis 153
    - confirmation theory 166
  - network (BN) 728
  - behavior 830, 1096
  - behavioral
    - comparison 831
    - explananda 819
    - knowledge 759
    - mechanism 819
    - regularity 819
    - semantics 711
    - uncertainty 1087, 1096
  - behavior-based architecture 825
  - behaviorist model 1042
  - belief revision 227, 277
  - belief, desire, intention (BDI) 791
  - beliefs in model 648
  - beliefs, obligations, intentions and desires (BOID) 791
  - benchmark 803, 809
  - benefit 1108
  - best explanation 223
  - best solution 273, 281
  - best-first search 157
  - bifurcation point 644
  - bimodal modeling 126
  - biodiversity 899
  - bioeducational science 1046
  - biological mimicry 826
  - biology 471
    - optimality model 112
  - biomechanical model 900
  - biomolecular simulation 817
  - birobotic
    - discovery 832
    - experiment 822
  - bisimulation 280
  - black hole 870–873, 876
  - black-box model 895
  - Boltzmann 377, 386, 390, 391, 397, 408
  - borrower 1073
  - bottom particle 321
  - boundary condition 899, 904
  - bounded rationality 420
  - Boyle's Law 388
  - braided
    - monoidal categories (BMN) 540
    - river 116
  - brain state 652
  - Buckingham 377, 379, 380, 384, 398–408
  - building information model (BIM) 975, 982

## C

- calculus 342  
 calibration 899  
 canonical framework 122  
 capacity 424, 819  
 capital asset pricing model (CAPM) 1092  
 Carnot 510, 511  
 Carter 511–513, 515  
 Cartwright 14, 105  
 case 556  
 case-based reasoning (CBR) 436, 978  
 cash  
 – flow 1069  
 – inflow (CIF) 1067  
 – outflow (COF) 1069  
 categorical property 166  
 category theory 540, 543  
 causal  
 – explanation 105, 107, 111–113  
 – graph (CG) 439, 444, 446  
 – hypothesis 427  
 – influence 441  
 – mechanism 424  
 – model 964, 1104  
 – path model 1112  
 – pie 1111  
 – relation 437, 1111  
 – response (CR) 142  
 – unification 162  
 causal connection (CC) 164  
 – principle 164  
 causal relationship 444  
 – conjunctive 445  
 – disjunctive 445  
 – potential 445  
 – strong 445  
 causality 964, 968, 1022  
 causal-mechanical account 103, 112  
 causal-mechanical model 113  
 cause 652  
 Cayley graph 513  
 cellular automata (CA) 742, 786, 791, 1052  
 – model 116, 752  
 cellular automaton evolutionary slope and river (CAESAR) 896  
 central  
 – cognition 940  
 – controller 658  
 – processing 942  
 Cerenkov radiation 857  
 CERN 736, 959  
 ceteris paribus 688, 690  
 chains of reasoning connections (COR) 370  
 chance discovery 1015–1017  
 channel-hillslope integrated landscape development (CHILD) 896  
 chaotic system 784  
 chemical structure 880, 885  
 chemical synthesis 884  
 Chemla 510  
 children malnutrition  
 – BINP 1113  
 – in developing countries 1113  
 – TINP 1113  
 choice set 236  
 chronological model 993  
 cicada 113  
 circadian biology 605  
 classical  
 – abduction 270  
 – computationalism 695  
 – economics 960  
 – logic (CL) 233, 250, 317  
 – mechanics 109  
 – negation 319  
 – particle mechanic (CPM) 36  
 – physics 159  
 – positive propositional logic (CPL+) 318  
 – propositional logic (CPL) 319  
 classification 725, 727  
 classifying abduction 206  
 class-size reduction 1112  
 climate science 891, 896  
 clock-controlled genes (CCG) 614  
 closure uncertainty 903  
 coalition enforcement 207  
 co-evolution of scenarios 1019  
 co-exact property 505–507, 512, 516  
 cognition 119, 123, 127, 341, 619  
 cognition and action 640  
 cognitive  
 – affordance 606  
 – and physical search 653  
 – computation 709  
 – development 643  
 – history 504  
 – labor division 116  
 – manipulating 201, 202  
 – mechanism 154  
 – mediator 210  
 – model 793, 1043  
 – niche 210  
 – penetrability (CP) 573, 575  
 – process 483, 616  
 – rationality 207  
 – robot 619, 635  
 – robotic 635  
 – science 112, 606  
 – studies of science 358  
 – system 619, 622, 635  
 – tool 500  
 – underpinning 484  
 – unity thesis 555  
 cognitive-historical approach 355  
 cognitively impenetrable (CI) 575  
 coherence 480  
 coherent system of units 407, 408  
 collaboration 123  
 colligation 183, 188  
 color vision 963  
 combination of abductions 255  
 commitment 301, 306, 309–312  
 common  
 – cause abduction 164  
 – cause principle 956  
 – features account 113  
 – pool resources and multi-agent systems (CORMAS) 795  
 – sense entity 962  
 communication 1016  
 communication with presenting users'/inventors' conditions 1020  
 comparative claim 766  
 comparison 803  
 – experiment 812  
 – or combination of data 1025  
 compensation 441, 442  
 completely defined 1024, 1026  
 complex  
 – adaptive system 784  
 – cognitive system 660  
 – dynamic systems (CDS) 570, 657  
 – model 662, 1033  
 – paramorphic model 996  
 – system 421, 783, 964  
 – system explanation 753  
 complexity 123, 126, 128, 180, 738, 783, 906  
 – hierarchical 183, 191  
 – science 641  
 complicated model of cognition 660  
 composition theorem 457

- computation  
 – concept network 700  
 – inquiry 772  
 – non-symbolic 704  
 – symbolic 704  
 computational  
 – (executable) model 695  
 – abduction 157  
 – and representational theory of mind (CRTM) 940  
 – approach 696  
 – description 930  
 – explanation 748, 757, 929–938  
 – instrument 761  
 – landscape 747  
 – language 752  
 – mathematics 758  
 – matrix representation (CMR) 682  
 – method 738  
 – neuroscience 934  
 – studies 773  
 – theory 932  
 – theory of the mind 736  
 – thinking 697  
 – tool 695  
 – turn 695, 738  
 computational model 799, 930, 932  
 – epistemology 743  
 computational modeling 699, 704, 931  
 – open problem 940  
 computationally  
 – equivalent 741  
 computer 905  
 – metaphor 658  
 – modeling 473  
 – science 802  
 computer simulation 114, 420, 472, 663, 742, 804, 825, 832, 843, 847, 861, 865, 895, 964  
 – agent-based 742  
 – coupled 742  
 – definition 767  
 – demonstration 769  
 – epistemological analysis 744  
 – explanatory relevance 754  
 – hybrid method 767  
 – identity 770  
 – inferential immediacy 754  
 – multiscale 742  
 – physical 770  
 computer-aided design (CAD) 981  
 computer-implemented method 768  
 conceivability 482  
 concept of similarity 383  
 conceptual  
 – abstraction 162  
 – emergence 749  
 – model 892, 903  
 concession problem 296  
 condition of similarity 394  
 conditional inference rule 257  
 conditional rule (RC) 233  
 confidence 1088  
 confirmation 220, 652, 922–924  
 conflict set 441, 443  
 – minimal 443  
 conformation 885  
 conformity 1008  
 confounding factor 1105  
 confounding variable 956  
 conservatism 1005  
 consistency 315–317  
 consistency connective 317  
 consistency-based reasoning 459  
 constraint 121–126, 371, 1024  
 constraint programming 459  
 construal 204  
 construction 554, 975  
 constructivism 966  
 constructivist model 1046  
 consumer 1020  
 consumerism 1009  
 content element 166  
 context of discovery 470  
 context of justification 470  
 contextualist model 1045  
 contraction 169, 227  
 contradiction 263, 315–317, 320  
 contradictory hypothesis 262  
 control parameter 642  
 controlled experiment 807, 812  
 convergence 905  
 co-occurrences 1018  
 corollarial and theorematic reasoning 198, 201  
 Corollarial deduction 198  
 correctness theory of truth 967  
 correlated dispositions 164  
 correlation 956  
 correspondence 966  
 – rule 13, 27, 30, 959  
 corresponding  
 – motion 390  
 – quantities of heat 390  
 – rule 959  
 – states 387–389  
 – times 390  
 corroboration 263  
 counterfactual account of explanation 106  
 counterfactual obligation 1120  
 coupled  
 – cognitive system 128  
 – model 896, 908  
 – model intercomparison project (CMIP) 896  
 – oscillator 642  
 covering law model 672  
 Craver 105, 114  
 creative  
 – abduction 151, 158, 198, 254  
 – communication 1019  
 – fact abduction 255  
 – hypothesis formation 252  
 – leap 979  
 – reasoning 555  
 – thinking 975  
 creativity excludes logic 249  
 cretaceous mass extinction 906  
 criminal 1055  
 criminology 1051  
 criterion of demarcation 959  
 criterion of falsification 959  
 critical velocity 866  
 criticism 1023  
 crowdsourcing 127  
 culturalist model 1045  
 current model 1046  
 cusp catastrophe 644  
 cybernetic intelligence 825  
 cycle 1075  
 cycle of erosion 892
- 
- D**
- 
- dark information 688  
 Darwin 210, 913  
 – robot 827  
 Darwinian evolution 913  
 data 508, 761, 955, 960, 967  
 – analysis 1015  
 – collection 966  
 – driven innovation 1020, 1022  
 – driven science 121  
 – graph 606  
 – jacket (DJ) 1015, 1022  
 – mining 727, 1016  
 – model 899, 902, 915  
 – set 806  
 Davis 537, 542

- DDI account 9, 82  
 debt 1067  
 – deflation 1074, 1077, 1080  
 decision 1017  
 – making 1018, 1020  
 – under risk 643  
 deduction 175, 179, 436, 532, 533, 551, 555, 559, 575, 598, 675, 687  
 – corollarial 186  
 – inverse 185  
 – theorematic 185, 187  
 deductive 574, 597  
 – closure 551, 555, 559, 564  
 – logics 251  
 – nomological (DN) 103, 751  
 deep knowledge 436, 439  
 deep-time 899  
 defeasibility 1119  
 defeasible reasoning 250, 255, 258, 298  
 definitory rules 565  
 deflationary inferentialism 76  
 degree of freedom (DoF) 624  
 DEKI account 93  
 delineating phenomena 608  
 demarcation criterion (DC) 163, 959  
 demographic projection 968  
 demography 954  
 demonstration 82  
 demonstrative reasoning 556  
 DENDRAL 725  
 – heuristic 726  
 denotation 82  
 denotation, demonstration, interpretation (DDI) 9, 82, 767  
 denotation, exemplification, keying-up and imputation (DEKI) 95  
 deontic logic 327, 1119  
 deoxyribonucleic acid (DNA) 720, 849  
 dependence 962  
 depression 1076  
 derivability adjustment theorem (DAT) 318  
 description logics (DL) 710  
 design 1003  
 detection bias 899  
 deterministic 960  
 diagnosis 443  
 diagnostic function 437  
 diagram 605, 671, 689  
 – discipline 506, 516  
 diagrammatic reasoning 197, 200–202  
 diagrammatic reasoning and deduction (DIAMOND) 517  
 diagrammatic representation 607  
 diagrammatical reasoning 203  
 dialecticity 1033, 1039  
 dialogical logic 299–302, 306, 310  
 difference-making relation 964  
 differential equation 858, 868  
 digital machine 741  
 dimensional  
 – analysis 402, 407, 859–862, 865, 875  
 – equation 402  
 – homogeneity 401  
 dimensionless parameter 396, 397, 399, 402, 407, 859, 869, 875  
 dimensionless quantity 741  
 dinosaur 906  
 direct  
 – observational discovery 720  
 – perception 210  
 – representation 86  
 direction issue 1034  
 directly action-guiding experiment 800  
 discontinuity 661  
 discovery 557  
 Discrete Event System Specification (DEVS) 788  
 discrete state machine (DSM) 850  
 discursive inference 574–576, 584, 588, 592, 596  
 disjunction of abnormalities 234  
 disjunctive conceptual fault (DCF) 441, 443, 448  
 displacement of crime 1051  
 disposition-abductive 565  
 dispositional property 166  
 distorted model 860  
 distorted scale model 860, 894  
 distributed cognition 128, 707, 963  
 distributed, asynchronous, heterogeneous, and concurrent networks 695  
 distributedness 784  
 diversity constraint 62  
 domain model 1052  
 domain specific language (DSL) 710  
 Doppler effect 857, 868, 872  
 double description 553  
 dual recurrent neural network (DRNN) 625  
 dynamic  
 – epistemic logic (DEL) 270  
 – frames approach 371  
 – hypothesis 364  
 – mechanistic explanation 947  
 – micro-simulation 787  
 – model 659, 768, 804  
 – process 659  
 – proof theory 257  
 – similarity 395, 396  
 – sufficiency 917  
 dynamical  
 – analysis 939  
 – similarity 388, 394–397  
 – system 939  
 – versus computational 641  
 dynamicism 939, 944  
 dysfunctional behavior 821
- 
- E**
- 
- early vision 575, 585, 591–595  
 Earth system model (ESM) 896  
 earthquake 866–868, 897  
 ease of learning (E.O.L.) 1044  
 EAUI-conception 207  
 eco-cognitive 206  
 – model of abduction 197  
 – perspective 207  
 ecological approach 658  
 econometric model 964  
 economic  
 – engineering 427  
 – entity 962  
 – theory 422  
 economics 178, 425, 478, 954  
 educability 1041  
 education 1033  
 educational  
 – development 1034  
 – developmental science 957, 1036  
 efficiency 905  
 efficient market 1074  
 Einstein–Podolsky–Rosen (EPR) 188, 474  
 electrical circuit 858, 875  
 electron 959  
 electronic numerical integrator and computer (ENIAC) 736  
 element 502, 507

- elementary diagnose 437, 444
  - set of 437
- elementary geometry 532, 533
- element-creative abduction 156
- embodied cognition 619, 623
- embodiment 472
- emotion 1123, 1126
- emotional contagion 1124
- empathy 1124
- empirical
  - control 966
  - data 898
  - model 955, 960, 963
  - structural argument 560
- empiricism
  - anthropocentric 760
- energetics 401
- energy-balance model 896
- enriched model 1046
- ensemble concept 390
- environment 832
- environmental approach 359
- environmental uncertainty 1087
- epidemic spread 784
- epidemiology 955
- episodic memory 650
- epistemic
  - action 503, 514
  - agent 953, 961, 962
  - artifact 22
  - change 227
  - experiment 800
  - logic (EL) 270
  - opacity 126, 746
  - representation (ER) 52
  - representation problem 52
  - requirement 828
  - theory 228
  - tool 21, 822, 962
- epistemological
  - abduction to reality 168
  - contradiction 316
  - role 744
- epistemologically substitutable 760
- epistemology 141
  - complex system 745
  - multilayered 744
- equation 383, 956
  - based model 421
  - of motion 393
  - of state 388
- equifinality 907
- equilibrium
  - analysis 422
  - explanation 107
- equivalence computational 741
- ergodicity 1091, 1094
- Ermentrout–Kopell model 113
- erroneous generalization problem 559
- error
  - of prediction 648
  - of reasoning 207
  - term 960
- ethnography 363, 957
- Euclid 499, 503, 506
- Euclidean diagram 503–508, 510, 512–515, 519
- Euclidean geometry 499–504, 506–508, 516–520, 538, 543
- Euclid’s Elements 524–526, 531, 532
- euRathlon 810
- EURISKO 724
- European Organization for Nuclear Research (CERN) 736, 959
- euthanasia 1108
- event-related potentials (ERP) 587
- evidence 262, 317, 320, 744, 915, 964
  - based medicine 964
  - extraction and link discovery (EELD) 1016
  - of production 964
- evolution 913–915, 917–920, 925–927
- evolution of form 920
- evolutionary
  - abduction 160
  - computation 728
  - computation (genetic programming) 725
  - computational 730
  - model 922
  - stable strategy (ESS) 925
  - synthesis 921
  - theory 913, 916, 920–922, 925
- exact property 505, 508, 516
- exception 820
- excluded middle 320
- execution 804
- exemplification 91
- exogeneity 956
- expectation 648, 1073
- expected utility (EU) 1090
- expected utility theory (EUT) 643
- experiment 744, 760, 762, 799, 957
  - epistemic motivation 765
  - epistemic privilege 766
  - experimental proof 758
  - experimental stance 758
  - on diagrams 200
  - social experiment 1104
  - target system 765
  - theoretical 752
- experimental
  - computer science 802
  - diagram 199
  - economics 766
  - manipulation 203, 205, 998
  - methodology 802
  - realm 469
  - water tank 385
- experimental model 391, 955, 957, 1046
  - adaptive 1047
  - basin 385
  - enriched 1046
  - organismic 1047
- experimentation 125, 416, 958
  - on analog 861
  - on analog model 861, 865, 873
- experimenter’s regress 762
- expert
  - knowledge 438
  - system 438, 459
- expertise 344, 351
- explained variance 168
- explanation 425, 673, 680, 687–690, 751, 803, 821, 830, 895, 922, 968
  - and prediction 881
  - causal account 755
  - causal model 751
  - complex system 753
  - computational 747
  - explanatory value 756
  - in abduction 557
  - likeliest 152
  - manipulationist account 756
  - mathematical 747
  - paradox 108
  - statistical-relevance model 672
  - unificationist model 751
- explanatory 224
  - and predictive power of model 964
  - depth 112
  - fiction 109

- power 964, 968
- relevance 754
- explicit rules in patterning 554
- exploitation 1085
- explorative cooperation 123
- explorative experiment 805
- explosive 316
- external
  - artifactual model 205
  - model 198
  - reality hypothesis (ERH) 75
  - validity 958
- extrapolative abduction 161
- extrinsic representation 680, 682–684, 688

## F

- fable 88, 1105
- fact 966
- factual abduction 155, 253
- failure 771
- fairness 1108
- fallacy 197, 206, 212
- fallibility 474
- false assumptions 424, 425
- falsification 263, 961
- falsity of the assumption 965
- families of probability distribution 967
- family resemblance 701
- Faraday 342, 345, 351
- fast enabling link (FEL) 111
- fault
  - detection 440
  - negative 450
  - positive 450
- feasibility experiment 812
- feature-based modeling (FBM) 982
- feedforward neural network (FFNN) 622
- feedforward sweep (FFS) 585
- feeling of knowing (F.O.K.) 1044
- Ferreiros 503, 508
- Fibonacci 517, 520
- fiction 18, 83, 103, 107, 110–112, 116, 121, 486
- fictional entity 824, 961
- fictionalization 105, 111
- field experiment 812
- final derivability 235
- financial 1068

- financial posture 1069
  - Ponzi 1070
- financialization 1081–1084
- finite element model (FEM) 683
- first-order
  - abduction 254
  - existential abduction 156
  - logic (FOL) 1026
- Fisher's Sex Ratio model 112
- fitness function 926
- five-dimensional (5-D) 982
- folk ontology 16
- force 345
- forecasting 895
- formal
  - diagrammatic proof system 524, 528–532, 534
  - language schema 261
  - model 265, 990
- formality condition 940
- formalization 523, 532
- format of representation 758
- fossil 899
- four-dimensional (4-D) 982
- fractal dimension (FD) 663
- fracture 867
- frame problem 679–682, 684–686, 690
- framing 978
- free probability theory 511
- front eye fields (FEF) 591
- Froude 377, 385
  - number 385
  - similarity 385
- Fumerton's reduction 155
- function 383, 408
- functional
  - abstraction 111
  - explanation 929, 934
  - kind 111
  - magnetic resonance imaging (fMRI) 752
  - realization 936
  - substitutability 760
- functionalism 1008
  - disaggregation 1007
  - efficiency 1011
  - universalism 1008
- functionalist model 111
- functionality benchmark 811
- functionally substitutable 760
- fundamental theorem (FT) 540

## G

- Gabbay–Woods (GW) 270
- Galileian thought experiment 1105
- Galileo Galilei 380, 383, 387, 392, 397, 408, 464, 509, 520
- game of make-believe 18
- game theory 420, 1095
- Gang of Eighteen 207
- Gaussian random matrix (GRM) 511
- general
  - circulation model (GCM) 6, 896
  - definition of information (GDI) 966
  - equilibrium theory 1096
  - griceanism 55
  - purpose language (GPL) 710
- general griceanism 55
- generalization 820
- generalized abduction 171
- generalized likelihood uncertainty estimation (GLUE) 907
- generative model 975
- generativity 681, 684–686
- genetic programming 728
- genic selection 919
- Gentner 344, 350
- geographical information system (GIS) 789, 993, 1052
- geology 891
- geometric
  - diagram 523
  - group theory 513
  - similarity 382
- geometrical construction 203
- geometrical similarity 381, 395
- geometrically similar 381
  - mass 392
  - system 394
- geomorphic transport function (GTF) 895
- geomorphic-orogenic landscape evolution model (GOLEM) 907
- geomorphology 116, 896
- geophysics 858, 861–868, 891
- geophysics, experimentation on analog model 876
- Gestetner mimeograph 1009
- Giaquinto 502, 518–520
- Giere 19, 115, 899
- glacio-eustasy 905
- glaciology 891, 896
- GLAUBER 725

global  
 – broadcast 944  
 – model 1023  
 – recurrent processing (GRP) 586  
 glocal model 1016, 1023  
 glyphs 613  
 GOLEM 725, 727  
 good experimental methodology (GEM) 803  
 Gothic Revival 1005  
 GRAFFITI 724  
 GRAM 725  
 gravitational force 385  
 Great Lake Basin 905  
 Greenhill 381  
 greenhouse warming 898  
 Grice 507, 510  
 grid cell 685  
 Grosholz 509, 512  
 ground truth 804, 808  
 grounding relation 52  
 guardian 1055  
 GW schema 303–305, 310

---

## H

---

Hawking  
 – emission 872  
 – radiation 872, 875  
 heat map 610  
 hedge 1070  
 Helmholtz 377, 391–394, 402, 408, 862, 869, 875  
 Hepburn 383, 408  
 Hersh 537, 542  
 heuristic function 1103  
 hierarchical model 908  
 Higgs mechanism 959  
 high spatial frequency (HSF) 586  
 Hilbert 502, 505, 508  
 hippocampus 685  
 historical-fact abduction 156  
 history 114  
 hitting set 442, 443  
 homalite 866, 867  
 home blindness 957  
 homeomorphic model 990  
 homologous 382, 404  
 – acceleration 381  
 – force 385  
 – linear dimension 384  
 – motion 385  
 – path 381

– place 384  
 – property 383  
 – quantity 383  
 – time 381  
 – velocity 381  
 homologue 383  
 honesty 1008  
 honeybee 113  
 Hopf monoids (HM) 540  
 Horn clause 1025  
 Horton law 895  
 hot spot 1051  
 how-actually 103, 114–116  
 – explanation 426  
 how-possibly 103, 114–116  
 – explanation 426  
 Hubbert 861–865  
 human  
 – aware 1054  
 – computer 737  
 – error 1114  
 – offloading 128  
 – reasoning pattern 250  
 – reasoning process 252  
 – scientific discovery 722  
 Humpty Dumpty problem 56  
 hybrid simulation 826  
 hydrodynamics 391  
 hydrological model 897  
 hydrology 891, 904  
 hypothesis 220, 223, 252  
 – formation 251, 255  
 – generation 209  
 hypothetical  
 – cause abduction 162  
 – reasoning 249, 265  
 – reasoning pattern 250  
 – theory 1028  
 hypothetico-deduction 563  
 hypothetico-deductive  
 – account of confirmation 1000  
 – method 966  
 – model of confirmation (HD) 242  
 hypothetico-structural (HS) 104  
 – model explanation 104, 105

---

## I

---

IBE 220  
 I-confirmation 240  
 iconic  
 – based inference 557  
 – model 990  
 iconicity 507, 510  
 iCub 619, 623, 624, 628–632, 635  
 ideal  
 – mechanism 821  
 – system 820  
 idealization 40, 124, 480, 821  
 ideation 975  
 ideology 1004  
 – biomimicry 1011  
 – industrial design 1009  
 – modernism 1006  
 – revivalism 1005  
 idle cash balance (ICB) 1069  
 ignorance problem 296, 303  
 imaginary experiment 470  
 imagination 472  
 – nonsensory forms of imagination 486  
 – sensory imagination 473  
 imagined concrete thing 6  
 imitation 767  
 Immanuel Kant 504, 518, 893, 1118  
 implementation 475, 821, 829  
 in principle/in practice distinction 739  
 in vivo observation 363  
 inattentive blindness 958  
 incomplete model 1015, 1027  
 incremental  
 – belief revision 170  
 – diagnosis 456  
 indebtedness 1078  
 independence challenge 62  
 independent testability 167  
 individual  
 – based modeling 785  
 – based simulation 791  
 – difference 644  
 induction 152, 175, 179, 209, 219–221, 551, 558, 575, 598  
 – abductive 179  
 – enumerative 219, 223  
 – inverse 188  
 inductive 574, 597  
 – generalization 231  
 – inference 577  
 – leap 559  
 – statistical inference 155  
 industrial design 1009  
 inference 175, 177, 183  
 – analogical 192  
 – inverse 185



- to the best available explanation (IBAE) 152
- to the best explanation (IBE) 151, 176, 209, 219, 252, 557, 575, 586, 598
- toward a law 557
- inferentialist view of models 426
- inferotemporal cortex in the brain (IT) 583
- inflation 1078
- info-computation 699
- information
  - and communication technologies (ICT) 982
  - concept network 700
  - dynamics 699
  - processing 698, 705
- informational view 9, 20
- initial condition 899, 904
- innovation network 792
- innovator 1015, 1021
- instability 1066, 1074
- instance 238
  - negative instance 238
  - positive instance 238
- instantial view 8
- institutional design 427
- instruction 1034
- instrumental
  - abduction 170, 206
  - apparatus 475
- instrumentalism 424
- instrumentalist
  - account of models 961
  - interpretation 165
  - position 962
- integrative systems biology (ISB) 122
- intellectual seemings 483
- intelligence quotient (IQ) 659
- intention 1019
- intercomparison project 896
- interface 774
- intermediate
  - attractor 644
  - information acquisition 158
- intermodel comparison 896, 907
- internal comparison 823
- interpretation 80, 160
- interpretation of result 966
- interpretative issue 1034
- intervention 957
- intrinsic representation 681–685
- intuition 480

- intuitionistic logic 250, 320
- intuitive physics engine (IPE) 686
- invariance 957
- inventor 1020
- inverse
  - problem 721, 722
  - problem modeling 902
  - square law 342
- investigating conclusions 256
- investment 1071, 1072
- invisible gorilla 958
- Ising model 6
- isolation 416, 962, 967
- isomorphic mapping 162
- isomorphism 37, 676–678, 684–686
- isostatic uplift 905

## J

- judgment 178, 181, 188, 1111
  - abductive 184, 190
  - of learning (J.O.L.) 1044
- justification 554, 803
  - empirical 745
  - theoretical 745
- justificatory function 153

## K

- kairctic account 108
- KEKADA 725
- kinematics of conceptual change 367
- kinetic theory of gases 388
- knot
  - diagram 514, 520
  - theory 514, 517, 520
- knowledge 467, 743, 748, 1053, 1086–1088
  - a posteriori 479
  - a priori 479
  - ability knowledge 479
  - base (KB) 437
  - behavioral 759
  - compilation 450, 459
  - counterfactual 479
  - engineering (KE) 438
  - generation 695
  - modal 479
  - propositional 479

## L

- laboratory
  - effect 895
  - experiment 363
- Lakatos 350, 541
- Lake Superior distorted model 861
- landscape evolution model (LEM) 895
- late vision 575, 585, 587, 591–597
- latent variable 960, 1018
- lateral occipital complex (LOC) 591
- lattice 741
  - gas automaton model 112
  - gas model 752
  - method 747
  - network 742
- law 556, 915, 919
  - abduction 158, 253
  - of correspondence 382
  - of corresponding states 387, 389, 390
  - of dynamical similarity 396
  - of gravitation 406
  - of similarity 384, 398
- learning 425, 1035
- lender 1073
- lesion studies 823
- lettered diagram 504, 505
- Levi identity 170
- liability 1067
- liberalization 1083
- life size replica model 894
- likeliest explanation 152
- line graph 611
- linear association 661
- lines of force 345
- linguistic agent 206
- liquidation 1075
- liquidity 1077
- literary fiction 85
- living system behavior 830
- local
  - factor 1111
  - model 1023
  - or global minimum 642
  - recurrent processing (LRP) 585
- localism 1006
- logic 403, 406
  - dialogical 299–302, 306, 310
  - first-order (FOL) 1026
  - for abduction 256, 258
  - theorist (LT) 723

logical  
 – abnormality 255, 257  
 – content 749  
 – explosion 256  
 – goodness 557  
 – law 474  
 logicity 176, 183  
 logics as model 250  
 logics of deontic (in)consistency (LDI) 328  
 logics of formal inconsistency (LFI) 315–317  
 – anodic system 327  
 – cathodic system 327  
 – first-order 326  
 – modal logics 327  
 – QmbC 326  
 – QmbC-tableaux 326  
 long term  
 – memory (LTM) 344, 592  
 – potentiation (LTP) 937  
 – working memory (LTWM) 343  
 Lorentz 377, 386, 387  
 Lotka–Volterra model 7  
 loveliest explanation 152  
 low spatial frequency (LSF) 586  
 lower limit logic (LLL) 232, 256, 261  
 LTWM 350

## M

Macbeth 506–508, 519  
 Mach 380, 381, 868, 871, 875  
 machine  
 – functionalism 935  
 – vision 727  
 macro-level simulation 786  
 Magnani's discovery of manipulative abduction 212  
 magnetic  
 – field 345  
 – force 347  
 – induction 347  
 magnetism 346, 347  
 make-believe 18, 486  
 Mäki 417, 425  
 Manders 506–508, 512–514, 524–526, 530  
 Manhattan project 738, 742  
 manifestation 437  
 manipulation 200, 418, 473, 957  
 – function of language 207  
 – on diagrams 200  
 manipulative abduction 197, 201–205, 211, 708  
 – fallacy 206  
 – form of practical reasoning 204  
 – in animals 207, 211  
 – ubiquity 206  
 many-body model 6, 21  
 map 963, 967, 1018  
 mapping 66  
 marking  
 – for minimal abnormality 236  
 – for reliability 234  
 masculinity 1012  
 mass-balance model 896  
 massive modularity hypothesis (MMH) 941  
 material implication 251  
 material substrate 822  
 materiality 418, 765  
 mathematical  
 – discovery 541  
 – experiment 543  
 – explanation 202, 544  
 – inductive proof 560  
 – inductive proof for the sum of the interior angles in an  $n$ -sided convex polygon 560  
 – knowledge 551  
 – model 103, 106, 112–114, 537, 546, 893, 990  
 – reasoning 200  
 – representation 961  
 – structure 959  
 mathematically equivalent 741  
 mathematics 468, 723  
 – Euclidean 468  
 matrix functional 320  
 mature mathematical formalism 21  
 maximum likelihood method 728  
 Maxwell 110, 342, 347, 351, 377, 390, 397  
 – equations 350  
 – theory of electromagnetism 10  
 – Treatise 346  
 MAYA principle 1009  
 mbC 318  
 – bivalued semantics 320  
 – non-matrix-functional semantics 320  
 – semantics 319  
 – soundness and completeness 320  
 mean squared error (MSE) 632  
 meaning invariance 34  
 meaning schema 367  
 meaningfulness 967  
 measurement error 960  
 mechanical  
 – model 11  
 – similarity 383, 388, 398, 403  
 – similitude 386  
 mechanicism 929, 944  
 mechanism 819, 964, 1114  
 – causal 1114  
 – diagram 606, 613  
 – governing (MG) 829  
 – implemented (MI) 829  
 – sketch 830  
 mechanism-model-mapping constraint (3M) 106  
 mechanistic  
 – computational explanation 947  
 – explanation 127, 936  
 – model 111–113  
 – theory 829  
 MEACHEM 725  
 mediating instrument 962  
 mediator 967, 995  
 mental  
 – animation 615, 672  
 – imagery 607, 621–624, 629, 650  
 – map 671  
 – model 17, 129, 485, 671–678, 684–690, 1119  
 – model co-evolution 977  
 – rotation 674, 681  
 – simulation 621–624, 635, 943  
 metamodel 907  
 metaphor 12, 343, 349  
 metaphoric  
 – appreciation 980  
 – bases of mathematics 343  
 – understanding 348  
 metareflective model 1044  
 meteorology 898  
 method  
 – language 1034  
 – of dimensions 405  
 – of discovery 252  
 – of multiple working hypotheses 905  
 – of variation 474  
 – research 1034  
 methodological pluralism 123  
 micro-level simulation 783, 786  
 micro-macro link 783, 788, 793

- micro-part abduction 161
  - military intelligence 207
  - military nature of language 207
  - mind-dependent 962
  - mind-reading 1124
  - miniature universe 403, 405
  - minimal
    - diagnose 437, 442
    - disjunction of abnormalities 260
    - inferred Dab-formula 234
    - model 112, 113
    - representationalism 945
  - minimization of free energy 652
  - mirror diagram 202
  - mirror role 202
  - misbehavior 440
  - mixed ontology 6, 20
  - mobile robot 805
  - modal logic for abduction 258
  - modal operator
    - interpretation 261
  - mode of representation 510
  - model 25, 38, 39, 586, 596, 601, 804, 879–881, 885–887, 913, 918, 968, 990, 1003
    - as epistemic object 961
    - as imagined physical system 961
    - as mediator 20
    - assumption 923
    - chemical structure 879
    - citizen 1004
    - computational 702
    - experiment 401, 407
    - explanation 106
    - generative 582, 588
    - mathematical 702
    - of complexity 660
    - of computation 697
    - of data 36, 898, 963
    - of education 1037–1042
    - of interaction 806
    - ontology 1103
    - organic chemistry 881
    - organism 7
    - parameter 923
    - robustness 925, 998
    - solution 1004
    - type 914
  - model of cognition
    - Adaptive Control of Thought – Rational (ACT-R) 645
    - Adaptive Resonance Theory (ART) 647
    - Artificial Neural Networks (ANN) 646
    - Embodied Cognition 650
    - Free Energy Theory 652
    - Motor Theories 651
    - Multiprocess Models 648
    - Simulation Theory 651
  - model-based 573, 576, 596
    - abduction 198, 201
    - approach 439
    - diagnosis 435, 439, 442, 444
    - diagnostic reasoning 440
    - explanation 1020
    - reasoning (MBR) 129, 221, 343, 366, 435, 459, 523, 532, 533, 707–710, 1016, 1022
    - science 414
  - model-building 121, 128
  - modeling
    - building and testing 966
    - complex system 785
    - from above 121
    - practice 967
  - model-oriented computer simulation 825
  - modernism 1009
  - modus ponens 1120
  - modus tollens 1120
  - molecular motions 389
  - money 1067, 1077, 1081
  - money-manager capitalism 1082
  - monotonic 687
  - Monte Carlo
    - experiment 742
    - method 742, 762
    - simulation 747
  - moral
    - accounting 1121
    - analogy 1122
    - decision-making (MoralDM) 1123
    - dilemma 327
    - justice 1121
    - model 1126
    - problem 1117
    - schemata 1121
    - theorizing 1117
  - moral deliberation 1117
    - externalism 1126
    - internalism 1126
  - morphism 68
  - morphology 832
  - most advanced yet acceptable (MAYA) 1009
  - motion of a pendulum 959
  - motion of particles 959
  - motor
    - coordination 642
    - imagery 619, 626, 635
  - multi-agent simulator of neighborhoods (MASON) 794
  - multi-agent-based simulation (MABS) 786
  - multidisciplinarity 1036
  - multi-level model 956
  - multimodal abduction 210
  - multiple
    - explanatory hypothesis 256
    - realizability 107, 112, 748
- 
- ## N
- 
- narrative 486
  - natural
    - experiment 958
    - kind 164
    - law 481
    - selection 913, 917, 920, 925, 1011
    - versus formal language 252
  - naturalization of logic 206
  - Navier–Stokes equation 116, 895
  - necessary mathematical knowledge 554
  - negation 316
  - negative analogy 12
  - Nelsen 503, 517
  - neoliberal paradox 1083
  - neoliberalism 1083, 1085
  - neoliberalization 1081
  - Nersessian 17, 343, 355
  - net cash inflow (NCF) 1069
  - network 664, 727, 792
  - Netz 504, 515
  - neural implementation 824
  - neural system 111, 113
  - neuroscience 111, 113
  - neutral analogy 11
  - new computationalism 695
  - new observation 968
  - new riddle of induction 241
  - Newton 341, 377, 380–394, 408, 465
  - Newtonian abduction 160
  - Newton’s law 13
  - nomological-deductive model 935
  - non numerical digital model 741

non-bonding interactions 887  
 noncausal account of explanation 111  
 noncausal explanation 112  
 nonclassical logics 317  
 nonconcrete model 64  
 nonempirical virtue 480  
 non-Euclidean geometry 202  
 nonexplanatory abduction 206  
 nonfinancial corporation (NFC) 1083  
 nonhuman animal 208, 210  
 nonlinear dynamic systems theory 639  
 nonlinearity 784  
 nonmonotonic logic 250  
 nonomniscient agents 289  
 nontheory  
   – based simulation 121  
   – driven simulation 119, 123  
 norm 756  
 normal science 159  
 Norman 504, 519  
 novelty 749  
 numerical  
   – analysis 737, 741  
   – approximation 741  
   – approximation algorithm 904  
   – experiment 472  
   – experimentation 473  
   – method 737  
   – model 858, 866, 895  
   – science 741

## O

object  
   – management group (OMG) 710  
   – oriented programming (OOP) 847  
   – permanence 650  
 objectivity 963, 1110  
 observable  
   – entity 959  
   – fact abduction 155  
   – phenomenon 73  
 observation 183, 187, 473, 960  
 observation report 240  
 observational model 955, 957  
 observational statement 959  
 observationally equivalent 403  
 observations (OBS) 442  
 observation–theory distinction 27

oceanography 891  
 off-line cognition 650  
 Onnes 377, 386, 387, 389, 390, 397, 403, 408  
 ontology 713  
 opacity 754  
 open formula 427  
 optical role 202  
 optimal foraging model 926  
 optimality model 105, 107, 926  
 optimization 642  
 Oreskes 861, 865, 896  
 organic chemistry 879–881, 883–885, 889  
 organism 211  
 organismic model 1047  
 orientational way of knowing 553  
 Outdoor Streamlab 894  
 over-constrained (of model) 1105  
 over-constraining assumption 419, 425  
 over-indebtedness 1077  
 overparametrized 901  
 ozone hole 903

## P

paideia 1036  
 paleontology 891, 899  
 pancomputationalism 772  
 Pannell 381, 398, 404  
 Panza 505, 508  
 parable 88  
 paracompleteness 321  
 paraconsistency 315  
 paraconsistent logic 315  
   – semantics 331  
 paraconsistent modal logics 315, 327  
 parallel distributed processing (PDP) 707  
 parallel life 499  
 parameter 899, 916  
   – incommensurability 904  
   – uncertainty 903  
 parameter-fixing 122, 126  
 parametric model 975, 982  
 parametric sufficiency 917  
 parametrization 899  
 paramorphic model 990  
 parsimony 263  
 partial  
   – implementation 824  
   – information 876  
   – interpretation 30  
   – isomorphism 42  
   – knowledge 875  
   – relation 67  
   – structure 42, 67  
   – structure approach 15  
 particle filter 649  
 particle mechanic 15  
 parts of models 923  
 passers-by 1055  
 path diagram 1111  
 pattern 575, 582, 587, 593, 600, 642  
   – detection 607  
   – explicating 251  
   – justification 553  
   – matching 576, 594, 596  
   – of abduction 252  
   – of abductive inference 151  
 patterning task 553  
 patterns thinking 553  
 PC algorithm 725  
 Peano arithmetic 532  
 pedagogical antinomy 1035  
 pedagogy 1033–1040  
 Peirce 198, 322, 504, 509, 512  
   – abduction 138  
 pendulum 382, 385, 392  
 perceiving affordance 210  
 perception as abduction 211  
 perceptual-like clues 557  
 personal identity 478  
 perspectival approach 963  
 Petri net 787  
 phenomenological model 44, 104, 106, 114, 992  
 phenomenology 642  
 phenotypic gambit 925  
 philosophy 464  
   – epistemology 479  
   – Kantian philosophy 468  
   – of information 966  
   – of mathematical practice 502  
   – of science 124  
   – Western philosophy 464  
 phronomical affinity 381  
 photoelastic image 867  
 physical  
   – agency 713  
   – computation 706  
   – intuition 483  
   – model 390, 893  
   – similarity 405, 862, 865

- physical, natural, embodied  
   computation 695  
 physically similar system 377, 379,  
   380, 389, 394, 400, 404–406, 408,  
   409, 859–862, 875  
 physics 464  
   – absolute and relative motions 465  
   – Aristotelian physics 464  
   – law of the inclined plane 465  
   – Newtonian theory 466  
   – relativist theories of motion 477  
   – second law of thermodynamics  
   477  
 physics-based simulation 119  
 pink noise 663  
 place cell 685  
 planetary geomorphology 892  
 plate tectonic 891, 906  
 Plato 500, 518  
 plausibility model 275  
 plausible reasoning 541, 556  
 pluralism 1012, 1033  
 policy 892, 958  
   – making 788  
 politics 478  
 population genetic 913, 916,  
   919–923, 925  
 population genetics 918  
 Port House architectural design  
   project 985  
 portfolio 1069  
 Positive classical propositional logic  
   (CPL+) 318  
 possibility 482  
   – of misrepresentation 54  
 possible explanation 151  
 possible-translations semantics (PTS)  
   315, 331  
 post facto explanation 167  
 postcognitivist model 1044  
 post-constraint 1019  
 posterior 649  
 potential  
   – conflict structure (PCS) 446  
   – function 642  
   – outcome model 958  
 practical  
   – abduction 255, 260  
   – fact abduction 265  
   – reasoning 197, 204–206, 212  
   – singular fact abduction 258  
 practitioner 975  
 pragmatic view 20  
 pragmatism 182  
 Prandtl 377, 395  
 preconstraint 1019  
 predator–prey  
   – behavior 927  
   – population 785  
 prediction 262, 542, 968  
   – error 649  
   – of total recall (P.T.R.) 1044  
 predictive ability 424  
 prelinguistic agent 206  
 prenex conjunctive normal (PCN)  
   264  
 preparing activities in abduction  
   555  
 pretend play 19  
 price 1072  
 prima facie duties 1118  
 principle  
   – of compositionality 320  
   – of corresponding states 388–391,  
   397  
   – of dimensional homogeneity 402,  
   404, 406  
   – of dynamical similarity 396, 399,  
   404  
   – of explosion 315–317, 320  
   – of mechanical similitude 408  
   – of noncontradiction 316  
   – of property independence (POPI)  
   682  
   – of similarity 388, 397  
   – of similitude 403–405, 407  
   – of synthetic intelligence (PSI)  
   790  
   – of the relativity of size 404  
 probabilistic structure 965  
 probability 1086, 1088–1090,  
   1092, 1095  
   – density function (pdf) 649  
   – distribution 959  
   – theory 955  
 problem  
   – design exploration model 976  
   – of cause 905  
   – of external validity 895  
   – of scale 905  
   – solving process 364  
 process 771  
   – based theory 728  
   – uncertainty 903  
 processes of induction (PI) 221  
 proof construction 259  
 propensity score 958  
 prospect theory (PT) 643  
 provisos 33  
 pseudo-confirmation 167  
 psychology 957  
   – cognitive model 111  
 purpose in abduction 552  
 Pythagoras' theorem 500, 539
- 
- ## Q
- 
- qualitative  
   – causal rule 451  
   – diagnosis 450, 459  
   – model 955, 957  
 quality adjusted life year (QALY)  
   1105  
 quality of life 1106  
 quantitative  
   – analysis 955  
   – genetic 917  
   – model 955–957, 967  
   – outcome 475  
 quantity 383, 385, 394, 402, 404,  
   407  
 quantum mechanics (QM) 474, 848  
 quasi-experiment 958  
 quasi-money 1068  
 queuing system 787
- 
- ## R
- 
- random hypothesis 263  
 randomized controlled trial (RCT)  
   958, 1110  
 rare event 1017  
 raster plot 609, 610  
 rational expectation 1097  
 rational value theory (RVT) 643  
 rationality 422, 1086, 1091  
 raven paradox 241  
 ray (or geometrical) theory of optics  
   109  
 Rayleigh 377, 394, 396–401, 403,  
   407  
   – plot 610  
 real experiment (RE) 468  
 realism 962  
   – –anti-realism debate 959  
 realist interpretation 165  
 reality 957  
   – gap 807  
 reasoning 1015  
   – abductive 484  
   – ampliative 220

- analogical 193
- corollarial 185
- counterfactual 484
- deductive 484
- defeasible 220
- explanatory 220
- heuristic 484
- hypothetical 190, 219, 484
- inductive 484
- nonmonotonic 220, 224
- nonpropositional 484
- propositional 484
- simulative model-based 484
- spatial reasoning 473
- statistical 220
- theorematic 185, 187
- with combinations and analogy 1020
- reasoning pattern 250, 265
  - hypothetical 220, 224
- received view (RV) 25
- reciprocation 1121
- recognition
  - density 652
  - memory 644
- recollection 1019
- recommendation 1061
- reconstructive explanatory model 996
- recurrent
  - model 746
  - neural network (RNN) 622
  - processes (RP) 585
- recursive
  - decomposition 940
  - relationship 659
- reduced
  - complexity model 116, 896
  - pressure 388
  - relation equation 379, 384, 408
  - temperature 388
  - volume 388
- reductionism 1012
- reductionist 1012
  - approach 657
- Reech 377, 385, 386
- reference
  - benchmark controller (RBC) 627
  - strongly referential 750
  - weakly referential 750
- regional climate model (RCM) 896
- reification 111
- reinforcement learning (RL) 112

- relational
  - dimension 553
  - structure 162
- relationship between models and reality 962
- reliability strategy 259, 265
- repeatability 803
- representation 49, 341, 351, 605, 933, 958
  - external 605, 606
  - functional conception 77
- representational
  - demarcation problem 52
  - model 995
  - power 1103
  - practice 367
  - view 8
- representation-as 91
- reproducibility 762, 803
- reputation 1055
- requirement 1023
  - of directionality 54, 79
- resemblance 417, 425
- residual claimants 1084
- restitution 1122
- result 556
- retinotopy 677
- revivalism 1005
- Reynolds 377, 394, 396, 404
- Riabouchinsky 377, 407
- right-wrong thesis 141
- ripple 894
- risk 1071, 1092
- RoboCup 810
- robot 725
  - challenge 809
  - competition 809
  - scientist 725, 729
  - system 799, 801, 807
- robotic 619, 622, 624
  - behavior 830
  - competition 808
  - simulation 822
  - simulation model 805
- robotics 635, 799, 801
- robustness 422
  - analyse 116, 901
  - derivational robustness analysis 419
- RoCKIn 811
- role 760
- role model 1125
- Rousseaux 858, 872, 875
- routine activity theory 1056

- rule 1118
  - abduction 253
  - of evolution 640
  - of inference 959
- rule-based induction 726
- Russell 377, 385, 406–408, 502, 503

---

## S

---

- sample size 957
- sampling bias 899
- San Francisco Bay model 769, 894
- satisfaction criterion 241
- scaffolding model 995
- scale
  - effect 894
  - invariant 642
  - model 385, 398, 671–676, 681–684, 689, 858–862, 894
- scaling 408, 893
- scenario 1017
- scenarization 1019
- schemata 1121
- science
  - computational 735
  - of education 1039
  - technology, engineering, and mathematics (STEM) 377
- scientific
  - discovery 719
  - discovery system 725
  - knowledge 959
  - model-based reasoning 365
  - object 955, 961
  - practice 954, 963
  - representation 42, 417
  - rigour 957
  - statement 967
  - styles of reasoning 413
  - talent development 663
  - theory 959
  - understanding 963
- screening off 956
- search space 153
- search strategy 154
- second-order
  - abduction 254
  - existential abduction 161
- securitization 1083
- seeing as 980
- segregation 430
- selective abduction 151, 158, 253
- self organization 659, 783

- semantic
- constraint 932
  - information 967
  - modeling pragmatic usage 984
  - tableau 532, 533
  - value 320
- semantic view (SV) 14, 25, 36
- of theories 120, 920
- semantics 13, 712
- behavioural 711
  - denotational model 702
  - operational model 703
  - structural 711
- semiclassical model 108–110
- semiotic control 562
- sense-making 978
- sensing external events 1019
- sensitivity analyse 901
- sensorimotor system 128
- sentential model 990
- sentential representation 607, 679
- set
- of abnormalities 232, 258, 261
  - of abnormality 256, 261
  - theoretic or mathematical model 959
  - theoretic structure 959, 967
  - theory 920
- seven bridges of Königsberg 113
- shallow knowledge 436, 459
- Shell for Simulated Agent Systems (SeSAM) 795
- short term memory (STM) 344, 729
- sign activity 209
- similar
- machines 390
  - motion 381, 396
  - structure 384
  - system 377, 380–387, 390–397, 401–406, 408, 409
- similarity 9, 42, 58, 381, 385, 394, 1026
- conception 57
  - dynamic 385
  - kinematic 385
  - law 384
  - of motion 389, 398, 404
  - of states 388
  - principle 384
- similitude 386, 408
- simple abduction 253
- simple strategy 262, 265
- simplicity 480
- simplification 120
- simulacrum account of explanation 105
- simulate mechanisms 615
- simulated robot 806
- simulating 769
- simulation 129, 671–674, 681, 686, 804, 929–931, 963, 968, 991
- accuracy 828
  - as experiment 818
  - data-oriented 817, 827
  - inaccuracy 830
  - model-oriented 817, 827
  - opacity 757
- simulation studies
- invertebrate 826
  - vertebrate 826
- simulationist's regress 762
- simulator 806
- singular fact abduction 255, 265
- singularity 906
- sketching 616
- social
- behavior 786, 791
  - experiment 1104
  - pedagogy 1036
  - science 897, 955
  - sciences 967
  - structure of accumulation (SSA) 1085
- socioeconomic
- mechanism 427, 956
  - status (SES) 659
- sociology 954, 957
- of scientific knowledge (SSK) 362
- sociophysics 786
- sociotechnical system 787
- solution 1023
- intensive viewpoints 1019
  - uncertainty 903
- sonic analog 871, 876
- sonic analogue 857, 871
- spatial
- array 675, 678
  - distribution model 993
  - reasoning 605
- special needs education 1038
- special theory of relativity 317
- specialist modelling group (SMG) 985
- speculative abduction 162
- speech act 310–312
- square array pattern 553
- stability 1079
- stakeholder 1020, 1084
- standard 807
- format (SF) 232, 256
  - logic 319
  - of accuracy 53
- Stanton 381, 398, 404
- Starikova 513, 520
- state
- of affairs 966
  - of motion 389
  - space 916, 917
- static model 659
- statistical
- control 956
  - factor analysis 167
  - law 919
  - model 955, 967
- statistics 955
- steroid 888
- sticky information 1019
- stipulative fiat 55
- stochastic representation of reality 960
- Stokes 377, 391–394, 396–398
- strategic
- function 153
  - rationality 207
  - rule 565
- strategy 1058
- goal-oriented 976
  - minimal abnormality 258
  - problem-focused 976
  - reliability 258
  - solution-focused 976
- streamlining 1010
- strengthening the antecedent 264
- structural
- argument 560
  - discontinuity 560
  - formula 879–883, 885–887, 889
  - generalization 555
  - uncertainty 903
- structural model 956, 965
- explanation 107
  - uncertainty 903
- structuralism 68
- structure 36, 960, 967, 1114
- of scientific theory 26
- subjective expected utility (SEU) 1090
- substantial argument 563
- success 771
- Sugarscape 786
- superiority claim 765

- supersonic flow 870  
 Suppes 14, 898  
 support model 1052  
 supposition 486  
 suprachiasmatic nucleus (SCN) 609  
 surface of separation 869, 875  
 surface wave 860, 868, 871  
 surplus-meaning 688  
 surprise 642, 750  
 surrogative reasoning condition 51, 79  
 surveillance policy 1062  
 swap structures 331  
 syllogism 158  
 symbol 645  
   – manipulation 695  
 sympathy 1124  
 syntactic view 13  
   – of theories 120  
 syntax 13  
 system 383, 386, 394  
   – approach 1114  
   – definition of 442  
   – description (SD) 442  
 systematic search 719, 725  
   – massive 726  
 systems biology 121–124, 127
- T**
- 
- target system 963  
 task benchmark 811  
 tautology 263  
 taxonomy of computation 697  
 teaching model 1042  
 technical artifact 799  
 techniques for data analysis 957, 967  
 technological  
   – abduction 171  
   – reasoning 170  
   – theory 192  
 technology 961  
 testbed 803  
 theorem of Menelaus 510  
 theorematic deduction 198, 201  
 theorematic reasoning 197  
 theoretic control 562  
 theoretical 256  
   – abduction 197, 203, 255  
   – concept 164  
   – isolation 963  
   – model 7, 10, 474, 818, 955, 960, 963  
   – model abduction 159, 254  
   – plurality 1034  
   – realm 469  
   – singular fact abduction 261  
   – underdetermination 474  
 theories in social sciences 960  
 theory 316, 820, 957  
   – articulation 888  
   – based simulation 119–121  
   – consistency 34  
   – induced system 821  
   – laden 956  
   – ladenness of observation 28  
   – of education 1040  
   – of mind mechanism (ToMM) 942  
   – of numbers 499, 515, 520  
   – of similitude 386  
   – structure 916, 921  
 thermodynamically corresponding operation 390  
 thermodynamics 401  
 thinking 573–576, 582, 588, 592, 596  
 third variable problem 956  
 thought experiment (TE) 128–130, 347, 464, 748, 758, 1118  
   – brain in the vat 464  
   – Burge's arthritis 464  
   – Chinese room 464  
   – clock-in-the-box thought experiment 473  
   – Condillac's statue 464  
   – constructive thought experiment 476  
   – counter thought experiment 476  
   – Darwin on the evolution of the eye 464  
   – destructive thought experiment 476  
   – Einstein's lift 464  
   – Einstein's light beam thought experiment 481  
   – EPR 474  
   – functional thought experiment 478  
   – Galileo's 464  
   – genealogical thought experiment 478  
   – Gettier's thought experiment 466  
   – inverted spectrum 464  
   – Johnston's thought experiment 478  
   – Kant on handedness 464  
   – Keith Lehrer's Mr. Truetemp thought experiment 482  
   – Kripke's thought experiment on Gödel 482  
   – Lucretius' thought experiment 476  
   – Mary the Super-Scientist 467  
   – Maxwell's demon thought experiment 477  
   – Newton's bucket 465  
   – Parfit's amoeba 469  
   – philosophical thought experimentation 467  
   – platonic thought experiment 477  
   – Poincaré's disk world 464  
   –  $\gamma$ -ray microscope 464  
   – ring of Gyges 464  
   – Schrödinger's cat thought experiment 476  
   – scientific thought experiments 468  
   – Stevin's chain thought experiment 465  
   – the cow in the field problem 466  
   – the knowledge argument 467  
   – Thomson's violinist 469  
   – thought experiment of the ship of Theseus 478  
   – Twin Earth 466  
 threshold 906  
 Thurston 505, 515  
 tiger bush 115  
 time  
   – and complexity 639  
   – of vibration 382  
   – series 640  
 Tolman 398, 403–405  
 Toulmin Model 297, 299  
 tractability 422, 738  
 tractable model 769  
 tradeoff 103, 115  
 traditional model 1042  
   – behaviorist 1042  
   – cognitive 1043  
   – metareflective 1044  
 trajectory 640  
 transparency 422  
 trial experiment 812  
 trigger 224  
 trivial 316  
 true explanation in abduction 552



truth 321, 963, 965  
 – and falsity 965  
 – functional 320  
 truthfulness 967  
 truthmaker 966  
 tsugo 1019  
 tsugology 1019  
 Turing model 115  
 – of computation 698  
 two route, prefrontal instruction,  
 competition of affordances,  
 language simulation (TRoPICAL)  
 623

## U

ultra-speculative 1066  
 uncertain model 1024  
 uncertainty 1088  
 unconditional inference rule 257  
 unconditional rule (RU) 233  
 underdetermination 753, 902, 907  
 understanding 544, 755, 758, 961  
 undeterminedness 321  
 unfolding 744, 759, 768, 772  
 unification 769, 823  
 unified modeling language (UML)  
 710, 792  
 unitary account of models 44  
 universal constant 899  
 universality 107, 113  
 unobservable entity 959

unobservable-fact abduction 156  
 unrealistic assumptions 1104  
 Unruh 870–872, 876  
 unveiling role 202  
 update observation 276  
 upgrade operation 277  
 use  
 – explanation 746, 747  
 – justification 746  
 – of simulation autonomous robotics  
 805  
 usefulness of the model 965  
 utility 1090, 1097  
 – model 1107

## V

validation 120, 125, 827, 899  
 validity 184, 955, 965, 968  
 – abductive 178  
 – deductive 189  
 – empirical 179  
 – external 764, 1112  
 – internal 764, 1112  
 value at risk (VaR) 1093  
 van der Waals 377, 386–389, 397,  
 408  
 variable 899, 956, 960, 1016, 1018  
 variety of evidence 924  
 Vaschy 377, 406–408  
 velcro 1013  
 veridicality thesis 966

verification 125, 827  
 – criterion 959  
 virtual time 794  
 virtuous circularity 363  
 viscosity 394–397  
 Visser 858, 872  
 visual  
 – abduction 254  
 – image 606  
 – mental imagery 485  
 – system 606  
 visualization 474, 758  
 Voisin plan 1007  
 volcano 858, 863–865, 868  
 von Thünen 416, 429  
 vortex model 17

## W

wage-squeeze 1085  
 Warmr system 728  
 Wason selection task 1119  
 wave optics 109  
 weighted feature matching account of  
 model world-relation 62  
 well-formed formula 959  
 white hole 871–873  
 whole system model 991  
 working characterization 740  
 working memory (WM) 344, 587,  
 650, 674, 689, 944  
 worldly system 961

---

## Recently Published Springer Handbooks

**Springer Handbook of Model-Based Science** (2017)

ed. by Magnani, Bertolotti, 1179 p., 978-3-319-30525-7

**Springer Handbook of Odor** (2017)

ed. by Buettner, 1151 p., 978-3-319-26930-6

**Springer Handbook of Electrochemical Energy** (2017)

ed. by Breitkopf, Swider-Lyons, 1016 p., 978-3-662-46656-8

**Springer Handbook of Robotics (2nd)** (2016)

ed. by Siciliano, Khatib, 2227p., 978-3-319-32550-7

**Springer Handbook of Ocean Engineering** (2016)

ed. by Dhanak, Xiros, 1345 p., 978-3-319-32550-7

**Springer Handbook of Computational Intelligence** (2015)

ed. by Kacprzyk, Pedrycz, 1633 p., 978-3-662-43505-2

**Springer Handbook of Marine Biotechnology** (2015)

ed. by Kim, 1512 p., 978-3-642-53970-1

**Springer Handbook of Acoustics (2nd)** (2015)

ed. by Rossing, 1286 p., 978-1-4939-0754-0

**Springer Handbook of Spacetime** (2014)

ed. by Ashtekar, Petkov, 887 p., 978-3-642-41991-1

**Springer Handbook of Bio-/Neuro-Informatics** (2014)

ed. by Kasabov, 1230 p., 978-3-642-30573-3

**Springer Handbook of Nanomaterials** (2013)

ed. by Vajtai, 1222 p., 978-3-642-20594-1

**Springer Handbook of Lasers and Optics (2nd)** (2012)

ed. by Träger, 1694 p., 978-3-642-19408-5

**Springer Handbook of Geographic Information** (2012)

ed. by Kresse, Danko, 1120 p., 978-3-540-72678-4

---

**Springer Handbook of Medical Technology** (2011)

ed. by Kramme, Hoffmann, Pozos, 1500 p., 978-3-540-74657-7

**Springer Handbook of Metrology and Testing (2nd)** (2011)

ed. by Czichos, Saito, Smith, 1229 p., 978-3-642-16640-2

**Springer Handbook of Crystal Growth** (2010)

ed. by Dhanaraj, Byrappa, Prasad, Dudley, 1816 p., 978-3-540-74182-4

**Springer Handbook of Nanotechnology (3rd)** (2010)

ed. by Bhushan, 1961 p., 978-3-642-02524-2

**Springer Handbook of Automation** (2009)

ed. by Nof, 1812 p., 978-3-540-78830-0

**Springer Handbook of Mechanical Engineering** (2009)

ed. by Grote, Antonsson, 1576 p., 978-3-540-49131-6

**Springer Handbook of Experimental Solid Mechanics** (2008)

ed. by Sharpe, 1096 p., 978-0-387-26883-5

**Springer Handbook of Speech Processing** (2007)

ed. by Benesty, Sondhi, Huang, 1176 p., 978-3-540-49125-5

**Springer Handbook of Experimental Fluid Mechanics** (2007)

ed. by Tropea, Yarin, Foss, 1557 p., 978-3-540-25141-5

**Springer Handbook of Electronic and Photonic Materials** (2006)

ed. by Kasap, Capper, 1406 p., 978-0-387-26059-4

**Springer Handbook of Engineering Statistics** (2006)

ed. by Pham, 1120 p., 978-1-85233-806-0

**Springer Handbook of Atomic, Molecular, and Optical Physics (2nd)** (2005)

ed. by Drake, 1506 p., 978-0-387-20802-2

**Springer Handbook of Condensed Matter and Materials Data** (2005)

ed. by Martienssen, Warlimont, 1120 p., 978-3-540-44376-6