About Faults, Errors, and other Dangerous Things

Matthias Rauterberg

Work and Organizational Psychology Unit
Swiss Federal Institute of Technology (ETH)
Nelkenstrasse 11, CH-8092 ZUERICH
+41-1-63-27082, rauterberg@ifap.bepr.ethz.ch

Abstract

In this paper the traditional paradigm for learning and training of operators in complex systems is discussed and criticised. There is a strong influence (the doctrine of 'mental logic') coming from research carried out in artificial intelligence (AI). The most well known arguments against the AI-approach are presented and discussed in relation to expertise, intuition and implicit knowledge. The importance of faults and errors are discussed in the context of a new metaphor for cognitive structures to describe expertise.

Keywords: fault, error, learning, training, cognitive structure, expertise, intuition

1 Introduction

> "I learned more from my defeats than from my victories" (Napoleon, ca. 1800)

Why is this statement sometimes (or always) true? To answer this question we need a new understanding of human errors, inefficient behaviour, and expertise. In this paper we will discuss the importance of learning from unsuccessful behaviour. What percentage of unanticipated events (e.g., accidents) is caused by human error? This is a question that vexed researchers for years in the context of human interaction with complex systems.

The classical understanding of human errors is characterized by a *negative* valuation of erroneous behaviour, something that must be avoided. The Western Culture is constrained by this *taboo*: Not to talk about faults, errors and other dangerous behaviour! This taboo keeps us to present our self as successful as possible. We are–normally–not allowed to discuss in public how and what we could learn from our faults and errors.

Rasmussen defines human errors as follows [20]: "if a system performs less satisfactorily than it normally does–because of a human act–the cause will very likely be identified as a human error". Human errors are the most important cause of accidents or near-miss accidents. Heinrich [9] analysed insurance company records and got the result that approximately 85 percent of accidents are due to human error. Nagel [15] presents the results of his analysis: approximately 70 percent of accidents in aviation operations are classified as human errors. Accidents are categorised as caused by either unsafe acts of persons (e.g., operator error) or by unsafe conditions (cf. [9] and [23]). One consequence of using this dichotomy is often to blame the individual who was injured or who was in charge of the machine that was involved in the accident.

In fact, it is probably meaningless even to ask what proportions of accidents were due to human error. The more important question is what can one learn from his or her errors, and how are these insights and the derived knowledge embedded in the individual cognitive structure.

2 The Artificial Intelligence (AI)-Approaches

The AI-approaches can be distinguished in two different research tracks: (1) the traditional rule-based approach, and (2) the connectionistic approach. In the following we concentrate us on the rule-based approach.

One of the most elaborated modelling approach is Soar [13] [16]. Newell [17] describes Soar as follows: "Soar is … a *symbolic computational system*. … Soar is organised around *problem spaces*, that is, tasks are formulated as search in a space of states by means of operators that produce new states, where operators may be applied repeatedly, to find a desired state that signifies the accomplishment of the task. … Soar is organised entirely as a *production system*, that is, its long-term memory for both program and data

consists of parallel-acting condition-action rules. … Soar incorporates a *goal hierarchy*. … Soar learns continuously from its experience by *chunking*, which constructs new productions (chunks) to capture the new knowledge that Soar developed (in working memory) to resolve its difficulties" ([17] pp. 30-32).

Soar is based on impasse-driven learning. "While Soar is performing a task by using the behaviour model in working memory, it is also learning. It is building chunks every time it impasses from one problem space to another, … These chunks constitute the acquisition of knowledge for doing the task" ([17] pp. 62-62).

The knowledge generated by chunking and stored in the long-term memory represents only successful trials (i.e., solving an impasse). Knowledge of unsuccessful attempts (i.e., not solving an impasse) is not in memory. Learning in Soar means that long-term memory contains evidence only of the sequence of *effective actions*.

But, how would it be if the majority of the knowledge of the long-term memory of humans consists only of *unsuccessful trials?* Soar seems to be a typical representative of a theory driven approach for *error-free skilled behaviour* (cf. for other modelling approaches [4] pp. 80ff and [19] pp. 14ff). Why do we believe that an empirical driven approach – looking to the concrete task solving behaviour of people – is better than a theory driven approach? The answer refers to the following assumption.

2.1 Implicit assumption

Most of the known modelling approaches is based on the implicit assumption that the "mental model maps completely to the relevant part of the conceptual model, e.g. the user virtual machine. Unexpected effects and errors point to inconsistency between the mental model and the conceptual model" ([27] p. 258). This one-to-one mapping between the mental model and the conceptual model of the interactive system implies a *positive* correlation between the complexity of the observable behaviour and the complexity of the assumed mental model. But this assumption seems to be – in this generality – wrong.

Based on the empirical result in [21] (see chapter 3.2), that the complexity of the observable behaviour of novices is larger than the complexity of experts' behaviour, we must conclude that the behavioural complexity is *negatively* correlated with the complexity of the mental model. If the cognitive structure is too simple, then the concrete task solving process must be filled up with many heuristics or trial and error behaviour [21].

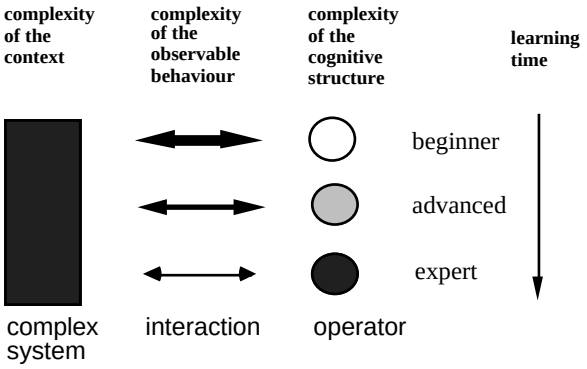

Fig. 1. The relationship between the complexity of the human interaction and of the cognitive structure.

Learning how to solve a specific task with a given system means that the behavioural complexity decreases and the cognitive complexity increases (cf. Fig. 1). Now, one of the central question is: What kind of knowledge is stored in the cognitive structure? Before we are able to give a preliminary answer to this question, we have to discuss the consequences of the traditional AI-paradigm.

2.2 Undesirable consequences

The famous classification of Rasmussen [20] in skill-based, rule-based, and knowledge-based behavior is one consequence of taking the AI-approach seriously. Human behaviour can be therefor classified (1) as error-free skilled behaviour (cf. [13]), (2) as inefficient behaviour (cf. [29]), and (3) as erroneous behaviour (cf. [28]).

It is very important to notice that in a rule-based expert system all unsuccessful trials during a task solving process (e.g., heuristic search) are thrown away after finding the correct solution path. A traditional expert system has no power of recollection of all these unsuccessful trials. Only correct solution paths are stored and could be retrieved later. This is an undesirable consequence of the classical AI-approach.

Newell [16] describes the chunking in Soar as to be not sensitive to success and failure, only to whatever resolves an impasse. "An impasse can be resolved as much by failure–the decision to leave a space as unprofitable–as by success" ([16], p. 313). It would be interesting to see

whether this type of failure is the only one in Soar. Although Rumelhart and Norman [24] claimed to investigate and to model specific errors of highly skilled typists, they exclude explicitly all "mechanism involved in learning".

Soar shows the direction of an alternative modelling approach for human learning processes. Each human has at least one recollection of an unsuccessful problem solving strategy in his or her individual learning history. This argument has a strong empirical evidence (cf. also [21]).

A second consequence of the traditional AI-approach is the fact that all inferences of a heuristic problem or task solving process must have a 'mental logic'. The most glaring problem is that people make mistakes. They draw invalid conclusions, which should not occur if deduction is guided by a 'mental logic'.

2.3 Critical statements

The doctrine of 'mental logic' can certainly be formulated in a way that meets the methodological criterion of *effectiveness*. The trouble with mental logic is thus empirical. Johnson-Laird [12] describes six main problems: (1) People make fallacious inferences. (2) There is no definitive answer to the question: Which logic, or logic's, are to be found in the mind? (3) How is logic formulated in the mind? (4) How does a system of logic arise in the mind? (5) What evidence there is about the psychology of reasoning suggests that deductions are not immune to the content of the premises. (6) People follow extralogical heuristics when they make spontaneous inferences. Why does cognitive psychology constrain the modern research to the doctrine of 'mental logic'? To come up with an answer, we have to look on the discussion and review coming from the Dreyfuses.

To Dreyfus [6], the world of the subjective is more important than that of the objective; reality is defined from within–in terms of the individual and his power to perceive and act, to know truths that are unutterable. Dreyfus concludes that some of the things' people do are intrinsically human and cannot be mechanised. To Dreyfus they are *intuition, insight,* and *comprehension*– the ability to immediately grasp complex situations, resolving ambiguities, weeding the relevant from the irrelevant (cf. the "exformation" described by Nørretranders [18]).

According to Dreyfus, the conviction that we can formalise reality, explaining everything with rules, began–as far back as the days of ancient Greece–and has become so dominant in the twentieth century that few people question it. This is one explanation for the doctrine of 'mental logic'.

Mary Henle [10] declares: "I have never found errors which could unambiguously be attributed to faulty reasoning." She suggests that mistakes arise because people misunderstand or forget premises, and because they import additional and unwarranted factual assumptions into their reasoning.

The Dreyfuses [5] argument that only novices use facts and rules. But as we become expert, we forget the rules and act intuitively, automatically adjusting our behaviour to the perceived constraints. Most scientists assume that these kinds of abilities are based on the unconscious and simultaneous processing of signals coming from the eyes, the ears, and the hands. But the Dreyfuses [5] believe that intuition defies rational powers of description, that it can't be computerised. Like judgement and wisdom it is one of the atomic element of our world (i.e. *irreducible*).

We share the critique of the Dreyfuses, but we do not follow their conclusions. To overcome the deadlock and 'mystical' situation following the Dreyfuses we need a new understanding of knowledge that gives an expert the ability to act intuitively.

3 Empirical Studies of 'Erroneous' Behaviour

> "There are no perfect humans, there are only perfect intentions" (Buddha)

Our basic assumption is that human behaviour cannot be erroneous. Of course, human decisions and the behavioural consequences of these decisions can be classified as erroneous and faulty, but from a pure introspective standpoint – from the internal *psycho*-logic of the subject – each decision is the best solution fulfilling all actual constrains and restrictions: lack of information and/or motivation, lack of knowledge and/or qualification, over or under estimation of the task and/or context complexity etc. In this sense we share the position of the Dreyfuses.

3.1 The 'law of requisite variety'

Humans need variety to behave and to adapt. A total static environment is insufferable. Ashby ([2] p. 90) summarises his analysis of regulation

and adaptation of biological systems as follows: "The concept of regulation is applicable when there is a set D of disturbances, to which the organism has a set R of responses, of which on any occasion it produces some one, r_j say. The physico-chemical or other nature of the whole system then determines the outcome. This will have some value for the organism, either Good or Bad say. If the organism is well adapted, or has the know-how, its response r_j as a variable, will be such a function of the disturbance d_i that the outcome will always lie in the subset marked Good. The law of requisite variety then says that such regulation cannot be achieved unless the regulator R, as a channel of communication, has more than a certain capacity. Thus, if D threatens to introduce a variety of 10 bits into the outcomes, and if survival demands that the outcomes be restricted to 2 bits, then at each action R must provide variety of at least 8 bits."

If we try to translate this 'law of requisite variety' to normal human behaviour then we can describe it as follows: All human behaviour is characterized by a specific extent of variety (see Fig. 2). If the system – in which the human has to behave – constrains this normal variety then we can observe 'errors'. In this sense an error is the necessary violation of system's restrictions.

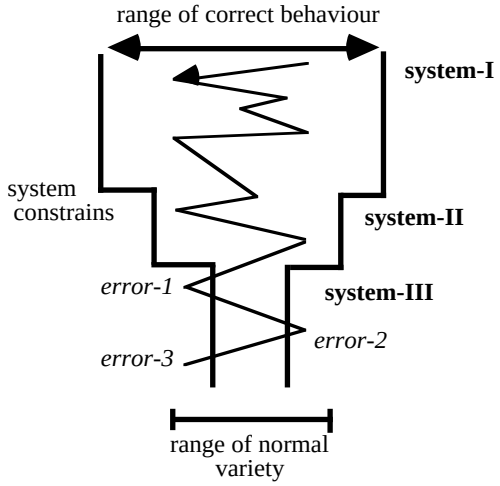

Fig. 2. The correlation between 'erroneous' behaviour and constraining behavioural variety.

If a system constrains human behaviour to only one possible 'correct solution path' then we can observe a maximum of violations, say errors. Husseiny and Sabri [11] counted 644 "critical incidents" in a representative study analysing complex systems (this is equivalent to an error rate of 16%); they noted that in "non nuclear complex systems" the rate of slips lies only between 3% and 7%. Most complex systems are designed to constrain the operator's behaviour to a minimum of variety. Ulich [25] arguments against this 'one best way' doctrine of system design because users differ inter- and intra-individually. A system must have a minimum of flexibility to give all users the opportunity to behave in an error-free way.

To investigate the relationship between behavioural and cognitive complexity we try to observe individual behaviour in its 'natural sense'. All deviations of the correct solution path are only interpreted as exploratory behaviour caused by the need for variety.

3.2 About the relationship between behavioural and cognitive complexity

In one of our experiments we compared the task solving behaviour of novices (subjects without experiences of electronic data processing EDP) with the behaviour of experts (subjects with a lot of EDP experience). We could show, that the complexity of the observable task solving process (the 'behavioural complexity') of novices is significantly larger than the complexity of the observable behaviour of experts (see Fig. 3). All twelve novices and all twelve experts solved exactly the same four different tasks (see [21]).

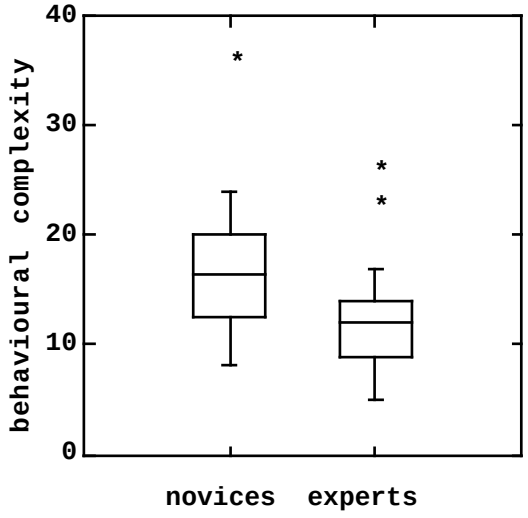

Fig. 3. The Box-and-Whisker Plot shows the 'behavioural complexity' for 'novices-experts' (N=24).

This important result seems to be – at the first glance – trivial. But what does it really mean? If we assume – and this is highly plausible – that

the complexity of the mental model of the experts is significantly larger than the cognitive complexity of the novices, then we must conclude that the correlation between behavioural and cognitive complexity is negative! And, this result is not trivial. Note in Fig. 3, that the minimal task complexity of all four tasks was only reached by one expert (task one: behavioural complexity = 6, the 'one best way'). We do not argument that this minimum cannot be reached, but to constrain human behaviour only to this minimum leads directly to the paradox of 'errors' of high skilled and over-trained experts.

3.3 What do we learn from errors?

Arnold and Roe assume ([1] p. 205), "that errors may have great functionality for the user, especially during learning. When the user is able to find out what has caused the error and how to correct it, errors may be highly informative. This implies that one should not try to prevent all errors." This hypothesis was tested later in an empirical investigation by Frese et al [7].

Frese et al [7] describe the following four reasons for the positive role of errors in training: (1) "the mental model of a system is enhanced when a person makes an error … (2) mental models are better when they also encompass potential pitfalls and error prone problem areas … (3) when error-free training is aspired, the trainer will restrict the kind of strategies used by the trainees, because unrestricted strategies increase the chance to error … (4) errors not only appear in training but also in the actual work situation." They compared two groups: one group with an error training (N=15), and a second group with an error-avoidant training (N=8). In a speed test the error-training subjects produced significant fewer errors than the error-avoidant group.

Gürtler ([8] p. 95) got the same results in the context of sport: "there, where more accidents were counted in the training phase, appeared less – above all of less grave consequences – accidents during the match. Few accidents during the training correlate with accidents of grave consequences during the match."

Van der Schaaf ([26] p. 85) concludes, that "every time an operator, manager, procedure, or piece of equipment 'behaves' in an unexpected way and thereby prevents a likely breakdown of the production system … or restores the required levels of safety and reliability, these *positive deviations* could be detected, reported and analysed in order to improve the qualitative insight into system functioning on the whole." This conclusion is not only valid for the global 'accident driven' design process "on the whole", this statement is also valid on the individual level of operating a complex system.

4 An Alternative View on the Knowledge of Mental Models

First, let us shortly summarise the traditional approach for learning based on training. To avoid unnecessary knowledge about unsafe acts beyond stable system's reaction operators are only trained on key emergency procedures. The beneficial effects of *extensive* training of these key emergency procedures are that they become the dominant and easily retrieved habits from long-term memory when stress imposes that bias. Sometimes emergency procedures are inconsistent with normal operations. To minimise the uncertainty coming from these inconsistencies Wickens demands the following design: "Clearly, where possible, systems should be designed so that procedures followed under emergencies are as consistent as possible with those followed under normal operations" ([28] p. 422).

We try to argument against this position. But, what is wrong with this traditional position? Nothing, of course not! Except the assumption that "knowledge about 'unsafe acts beyond stable system reactions' is *unnecessary* or *dangerous"*. If our experimental results (the *negative* correlation between behavioural and cognitive complexity, see [21]) are correct (and there is no evidence that they are not correct), then we must conclude that the cognitive structure of experts contains knowledge about unsuccessful trials (see [22]). What does this result mean for the cognitive structure of mental models about complex systems? Our conclusion is that humans need for effective and correct behaviour in critical situations a huge amount of knowledge 'about unsafe acts beyond stable system reactions'.

4.1 A metaphor for traditional learning and knowledge acquisition

To describe the traditional training procedure and the intended effects on the cognitive structure of operators, we introduce the following metaphor: The cognitive structure is a 'landscape' (cf. 'Tabula rasa'). This landscape–without knowledge–has a flat structure. Learning and

training of correct behaviour means 'to run a ditch' (cf. Fig. 4). The 'flow' of the actual behaviour can be described with 'a rolling ball' on this landscape. The 'course of the ball' describes exactly the observable behaviour.

Intensive training results in 'deepening the ditch' (cf. Fig. 5) to make sure that the operator behaves correctly especially in critical situations. But what happens if there is no ditch for the actual and–in the worst case–dangerous situation?

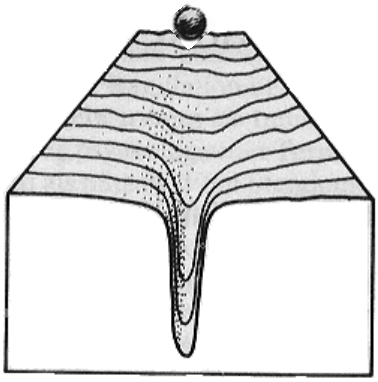

Fig. 4. The 'landscape'-metaphor for the cognitive structure with a 'rolling ball' to symbolise the 'flow' of actions at the *beginning* of a training.

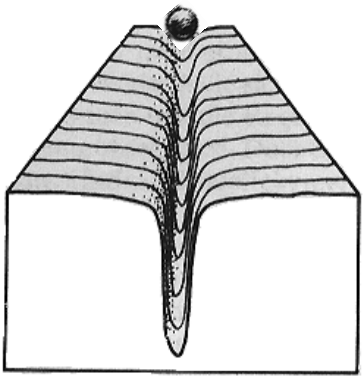

Fig. 5. The 'landscape'-metaphor for the cognitive structure with a 'rolling ball' to symbolise the 'flow' of actions at the *end* of a training.

Bainbridge [3] describes very clearly the problems arising when an operator has to take over a complex process during a monitoring task. To take-over process control is especially problematic when the system runs into an unknown state.

In this situation there is no 'ditch' to guide the 'ball'. Training in a simulator is one possible consequence, better is permanent on-line control in the real process. Operators should have on-line control over the real process (cf. [3]). High skilled operators tend to lose the potential to be aware of the whole process. They need a special qualification to get *open minded* (to increase their perceptual range; cf. [14]).

4.2 An alternative metaphor for learning and knowledge acquisition

What would happen, if operators are trained in simulators to get experience 'about unsafe acts beyond stable system reactions'? Following our metaphor the cognitive structure could be described as filled up with knowledge about unsuccessful behaviour: the 'hills' and 'mountains' (see Fig. 6).

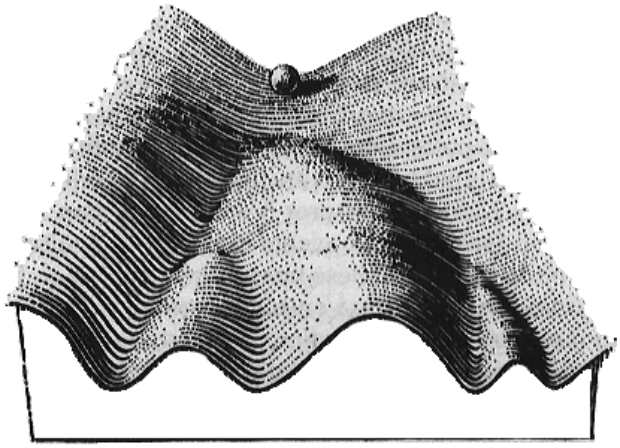

Fig. 6. The 'landscape'-metaphor for the cognitive structure with a 'rolling ball' to symbolise the 'flow' of actions and 'hills' as knowledge about unsuccessful behaviour.

The decisions for the actual behaviour are carried out to avoid *known* errors and faults! The 'ball' is guide by the 'valleys' and 'dales'. To minimise the effort is the basic principle for each actual decision. The implicit knowledge about effort comes from previous faults, errors and other dangerous behaviour. This is a possible explanation of *real* expertise.

Of course, this alternative metaphor for learning and knowledge acquisition can not replace the traditional view. This alternative is a completion, and probably a valid and very helpful one.

References

[1] Arnold, B. & Roe, R. (1987) User errors in Human-Computer Interaction. In: M. Frese, E. Ulich & W. Dzida (Eds.) Human Computer Interaction in the Work Place. Amsterdam: Elsevier, pp. 203-220.

[2] Ashby, W. R. (1958) Requisite variety and its implications for the control of complex systems. *Cybernetica* 1(2):83-99.

[3] Bainbridge, L. (1982) Ironies of Automation. In Analysis, Design, and Evaluation of Man-Machine Systems (G. Johannsen & J.E. Rijnsdorp, Eds.) Düsseldorf: VDI/ VDE, pp. 151-157.

[4] Booth, P. A. (1991) Errors and theory in human–computer interaction. *Acta Psychologica* 78: 69-96.

[5] Dreyfus, H. & Dreyfus, S.(1979) Mind over machine–the power of human intuition and expertise in the era of the computer. New York: The Free Press.

[6] Dreyfus, H. (1979) What computers can't do–the limits of artificial intelligence. New York: The Free Press.

[7] Frese, M., Brodbeck, F., Heinbokel, T., Mooser, C., Schleiffenbaum, E. & Thiemann, P. (1991) Errors in training computer skills: on the positive function of errors. *Human-Computer Interaction* 6:77-93.

[8] Gürtler, H. (1988) Unfallschwerpunktanalyse des Sportspiels. In: E. Rümmele (Ed.) Sicherheit im Sport – eine Herausforderung für die Sportwissenschaft. Köln: Strauss, pp. 91-100.

[9] Heinrich, H. (1959) Industrial accident prevention (4th edition). New York: McGraw-Hill.

[10] Henle, M. (1978) Foreword to R. Revlin & R. E. Mayer (eds.) Human Reasoning. Washington: Winston.

[11] Husseiny, A. & Sabri, Z. (1980) Analysis of human factor in operation of nuclear plants. *Atomenergie, Kerntechnik* Vol. 2.

[12] Johnson-Laird, P. (1983) Mental models. Cambridge (UK): Cambridge University Press.

[13] Laird, J., Newell, A. & Rosenbloom, P. (1987) SOAR: An architecture for general intelligence. *Artificial Intelligence* 33:1-64.

[14] Moray, N. & Rotenberg, I. (1989) Fault management in process control: eye movements and action. *Ergonomics* 32:1319-1342.

[15] Nagel, D. (1988) Human error in aviation operations. In: E. Wiener and D. Nagel (Eds.) Human Factors in Aviation. San Diego: Academic Press, pp. 263-303.

[16] Newell, A. (1990) Unified Theories of Cognition. Cambridge: Harvard University Press.

[17] Newell, A. (1992) Unified theories of cognition and the role of SOAR. In: J. Michon and A. Akyürek (Eds.) SOAR: A Cognitive Architecture in Perspective. New York: Kluwer, pp. 25-79.

[18] Nørretranders, T. (1991) *Mærk verden*. Kopenhagen: Gyldendal.

[19] Oostendorp, H. van & Walbeehm, B. J. (1995) Towards modelling exploratory learning in the context of direct manipulation interfaces. *Interacting with Computers* 7(1): 3-24.

[20] Rasmussen, J. (1986) Information Processing and Human-Machine Interaction. (System Science and Engineering Vol 12, A. Sage, ed.) New York: North-Holland.

[21] Rauterberg, M. (1993) AMME: an automatic mental model evaluation to analyze user behaviour traced in a finite, discrete state space. *Ergonomics* 36: 1369-1380.

[22] Rauterberg, M. (1995, in press) From Novice to Expert Decision Behaviour: a Qualitative Modelling Approach with Petri Nets. In: Proceedings of 6th International Conference on 'Human-Computer Interaction', Tokyo/Yokohama, July 9-14, 1995.

[23] Reason, J. (1990) Human Error. New York: Cambridge University Press.

[24] Rumelhart, D. & Norman, D. (1982) Simulating a skilled typist: a study of skilled cognitive-motor performance. *Cognitive Science* 6:1-36.

[25] Ulich, E. (1994, 3rd edition) Arbeitspsychologie. Stuttgart: Poeschel.

[26] Van der Schaaf, T. (1992) Near miss reporting in the chemical process industry. Proefschrift, TU Eindhoven.

[27] Van der Veer, G., Guest, S., Haselager, P., Innocent, P., McDaid, E., Oesterreicher, L., Tauber, M., Vos, U. & Waern, Y. (1990) Desiging for the mental model: an interdisziplinary approach to the definition of a user interface for electronic mail systems. In: D. Ackermann and M. Tauber (Eds.) Mental Models and Human-Computer Interaction 1. Amsterdam: North-Holland, pp. 253-288.

[28] Wickens, C. (1992) Engineering Psychology and Human Performance (2nd edition). New York: HarperCollins.

[29] Zapf, D., Brodbeck, F., Frese, M., Peters, H. & Prümpers, J. (1992) Errors in working with office computers: a first validation of a taxonomy for observed errors in a field setting. *International Journal of Human-Computer Interaction* 4:311-339.

published in:
H. Stassen & P. Wieringa (1995, eds.)
Proceedings of the XIV European Annual Conference on Human Decision Making and Manual Control.
Delft: Delft University of Technology
Faculty of Mechanical Engineering and Marine Technology
ISBN 90-370-0132-7