Human error and the problem of causality in analysis of accidents

Rasmussen, Jens

Published in:
Philosophical Transactions of the Royal Society B: Biological Sciences

Publication date:
1990

Document Version
Peer reviewed version

HUMAN ERROR AND THE PROBLEM OF CAUSALITY IN ANALYSIS OF ACCIDENTS[1]

Jens Rasmussen
Risø National Laboratory,
DK 40000, Roskilde, Denmark

Abstract: Present technology is characterised by complexity, rapid change, and growing size of technical systems. This has caused a rapidly increasing concern with the human involvement in system safety. Analyses of the major accidents during recent decades have typically concluded that human errors on part of operators, designers or managers have played a major role. There are, however, several basic problems in analysis of accidents and identification of human error. The paper addresses the nature of causal explanations and the ambiguity of the rules applied for identification of the events to include in analysis and for termination of the search for 'causes.' In addition, the concept of human error is analysed and its intimate relation with human adaptation and learning is discussed. It is concluded that identification of errors as a separate class of behaviour is becoming increasingly difficult in modern work environments. The influence of this change on the control of safety of large scale industrial systems is discussed.

INTRODUCTION

During recent decades the technological evolution has drastically changed the nature of the human factors problems one is faced with in industrial safety. In previous periods, the analysis of industrial systems could be reasonably well decomposed into problems which could be treated separately by different professions, and human factors specialists were primarily involved in attempts match the interface between people and equipment to human characteristics.

Complexity, rapid change, and growing size of technical systems have drastically changed this state of affairs. The hazards involved in operation of large-scale systems lead to reliability and safety requirements which cannot be proven empirically and, consequently, design must be based on models that can be used to predict the effects of technical faults and human errors during operation and to evaluate the ability of the operating organisation to cope with such disturbances. The human factors problems of industrial safety in this situation not only include the classical interface problems, but also problems such as the ability of the designers to predict and supply the means to control the relevant disturbances to an acceptable degree of completeness, the ability of the operating staff to cope with unforeseen and rare disturbances, and the ability of the organisation in charge of operation to maintain an acceptable quality of risk management. The human factors problems of industrial safety have become a true cross-disciplinary issue.

---

[1]Invited presentation at Royal Society meeting on Human Factors in High Risk Situations, London, 28-29 June, 1989. In: Phil. Trans. R. Soc. Lond. B 327, 449-462.

Analyses of the major accidents during recent decades have typically concluded that human errors on part of operators, designers or managers have played a major role. However, to come from this conclusion to suggestion of improvement is no easy matter and this problem appears to raise a couple of very basic issues related to the nature of causal explanations and the concept of human error.

One basic issue to consider is the *changing nature of engineering analysis* which in the past was a fairly well structured and bounded science. Technical designs could be verified and tested by quantitative models and controlled laboratory experiments. Descriptions in terms of mathematical relations among quantitative measurable variables were very effective for verification of design consistency and for optimisation of the internal function, as for instance of the thermodynamic process of a steam locomotive. Description of the less well formed relational structure of an accident such as a train crash was the business of another profession and the complexity of the situation required the use of less stringent causal analyses in terms of chains of events in which the internal functioning of the artefact is of minor importance; a locomotive in an accident is mainly a heavy object, not a thermo-dynamic machine. This separation of the two domains of analysis is no longer acceptable for systems such as chemical plants for which the internal functioning during accidents is determining the external damage. The interaction of relational and causal analysis therefore must be better understood and the integration of the two methods of analysis must be improved.

Another issue is the *notion of human error*. The concept of error is challenged by the rapid technological evolution and the transfer of people from manual manufacturing tasks to supervision and intervention during disturbances in automated systems. In a traditional work setting, the slow pace of change led to the evolution of fairly stable work procedures and it was easy to define human error with reference to normal practice. Consequently, in the traditional technology with slowly changing systems, causes of accidents were rather easy to determine and to attribute to technical faults or human error. Recently, however, some large scale accidents point to the need to explain accidents in terms of structural properties of integrated, large scale systems rather than to isolated links or conditions in a linear causal chain of events.

CAUSAL ANALYSIS OF ACCIDENTS

When analysing accidents after the fact, we are following a chain of events up-stream in order to understand, *why* it happened; to find somebody to blame, *who* done it; or find out *how* to improve the system. We are trying to describe a particular course of events and to identify the causes of the particular accident. It is, however, important to consider the implicit frame of reference of a causal analysis (Rasmussen, 1988a).

Causal Explanation.

A classical engineering analysis is based on mathematical equations relating physical, measurable variables. The generalisation depends on a selection of relationships which are 'practically isolated' (Russell, 1912). This is possible

when they are isolated by nature (e.g., being found in the planetary system) or because a system is designed so as to isolate the relationship of interest (e.g., in scientific experiment or a machine supporting a physical process in a controlled way). In this representation, material objects are only implicitly present in the parameter sets of the mathematical equations. The representation is particularly well suited for analysis of the optimal conditions and theoretical limits of physical processes in a technical system which, by its very design, carefully separates physical processes from the complexity of the outside world.

A causal representation is expressed in terms of regular causal connections of events. Russell (1912) discusses the ambiguity of the terms used to define causality: the necessary connection of events in time sequences. The concept of an 'event,' for instance, is elusive: the more accurate the definition of an event, the less is the probability that it is ever repeated. Completeness removes regularity. The solution is, however, not to give up causal explanations. Representation of the behaviour of the physical world in causal terms is very effective for describing accidents because the objects of the real world are explicit in the model and changes such as technical faults are easily modelled. This is not the case in a model based on relations among quantitative variables in which properties of an object are embedded in several parameters and equations. On the other hand, rather than to seek objective definitions it must be realised that regularity in terms of causal relations is found between kinds of events, *types*, not between particular, individually defined events, *tokens*.

The behaviour of the complex, real world is a continuous, dynamic flow which can only be explained in causal terms after decomposition into discrete events. The concept of a causal interaction of events and objects depends on a categorisation of human observations and experiences. Perception of occurrences as events in causal connection does not depend on categories which are defined by lists of objective attributes but on categories which are identified by typical examples, prototypes (as defined by Rosch, 1975). This is the case for objects as well as for events. Everybody knows perfectly well what 'a cup' is. To define it objectively by a list of attributes that separates cups from jars, vases and bowls is no trivial problem and it has been met in many attempts to design computer programs for picture analysis. The problem is, that the property to be 'a cup' is not a feature of an isolated object but depends on the context of human needs and experience. The identification of events in the same way depends on the relationship in which they appear in a causal statement. An objective definition, therefore, will be circular.

A classical example is "the short-circuit caused the fire in the house" (Mackie, 1965). This statement in fact only interrelates the two prototypes: the kind of short-circuit that can cause a fire in that kind of house. The explanation that the short-circuit caused a fire may be immediately accepted by an audience from a region where open wiring and wooden houses are commonplace, but not in a region where brick houses are the more usual kind. If not accepted, a search for more information is necessary. Short-circuits normally blow fuses, therefore further analysis of the conditions present in the electric circuit is necessary, together with more information on the path of the fire from the wiring to the house. A path of unusually in-

flammable material was probably present. In addition, an explanation of the short-circuit - its cause - may be needed. The explanation depends on a decomposition and search for unusual conditions and events. The normal and usual conditions will betaken for granted, i.e., implicit in the intuitive frame of reference. Therefore, in causal explanations, the level of decomposition needed to make it understood and accepted depends entirely on the intuitive background of the intended audience. If a causal statement is not accepted, formal logical analysis and deduction will not help, it will be easy to give counter-examples which can not easily be falsified. Instead, further search and decomposition are necessary until a level is found where the prototypes and relations match intuition. (The reason that nuclear power opponents do not accept risk analysis may be that they have an intuition very different from the risk analyst's intuition, rather than a lack of understanding of risk and probability).

Accident Analysis.

The very nature of causal explanations shapes the analysis of accidents. Decomposition of the dynamic flow of changes will normally terminate when a sequence is found including events which match the prototypes familiar to the analyst. The resulting explanation will take for granted his frame of reference and in general, only what he finds to be unusual will be included: the less familiar the context, the more detailed the decomposition. By means of the analysis, a causal path is found up-stream from the accidental effect. This path will be prepared by resident conditions which are latent effects of earlier events or acts. Also these resident pathogens (Reason, 1989) can be explained by causal back-tracking and in this case branches in the path are found. To explain the accident, these branches are also traced backward until all conditions are explained by abnormal, but familiar events or acts. The point is: how do the degree of decomposition of the causal explanation and selection of the side-branches depend on the circumstances of the *analysis*? Another question is: What is the stop-rule applied for termination of the search for causes? Ambiguous and implicit stop rules will make the results of analyses very sensitive to the topics discussed in the professional community at any given time. There is a tendency to see what you expect to find; during one period, technical faults were in focus as causes of accidents, then human errors predominated while in the future focus will probably move up-stream to designers and managers. This points to the question whether system break-down is related to higher level functional structures and feedback mechanisms rather than to the local conditions of events. In that case, traditional causal attributions turn out to be fighting symptoms rather than the structural origin of break-down.

The adoption of stop-rules is very important in the control of causal explanations. Every college student knows the relief felt when finding a list of solutions to math problems. Not that it gave the path to solution to any great extent, but it gave a clear stop-rule for the search for possible mistakes, overseen preconditions, and calculation errors. The result: hours saved and peace of mind. A more professional example to the same point is given by Kuhn (1976). He mentions the fact that chemical research was only able to come up with whole-number relations between elements of chemical substances after the acceptance of John Dalton's chemical atom theory.

There had been no stop rule for the efforts in refinement of the experimental technique until the acceptance of this theory.

Stop-rules are not usually formulated explicitly. The search will typically be terminated pragmatically in one of the following ways: (a) An event will be accepted as a cause and the search terminated if the causal path can no longer be followed because information is missing; (b) A familiar, abnormal event is found to be a reasonable explanation; or (c) A cure is available. The dependence of the stop rule upon familiarity and the availability of a cure makes the judgement very dependent upon the role in which a judge finds himself. An operator, a supervisor, a designer, and a legal judge will reach different conclusions.

To summarise: identification of accident causes is controlled by pragmatic, subjective stop-rules. These rules depend on the aim of the analysis, i.e., whether the aim is to explain the course of events, to allocate responsibility and blame, or to identify possible system improvements in order to avoid future accidents.

Analysis for Explanation.

In an analysis to explain an accident, the backtracking will be continued until a cause is found *which is familiar* to the analysts. If a technical component fails, a component fault will only be accepted as the prime cause if the failure of the particular type of component appears to be 'as usual.' Further search will probably be made, if the consequences of the fault make the designer's choice of component quality unreasonable, or if a reasonable operator could have terminated the effect, had he been more alert or been better trained. In such a case, a design or manufacturing error, respectively an operator error will be accepted for explanation.

In most recent reviews of larger industrial accidents, it has been found that human errors are playing an important role in the course of events. Very frequently, *errors are attributed to operators involved in the dynamic flow of events*. This can be an effect of the very nature of the causal explanation. Human error is, particularly at present, familiar to an analyst: to err is human, and the high skill of professional people normally depend on departure from normative procedures as we will see in a subsequent section. To work according to rules has been an effective replacement for formal strikes among civil servants.

Analysis for Allocation of Responsibility.

In order to allocate responsibility, the stop-rule of the backward tracing of events will be to identify a person who made an error and at the same time, '*was in power of control*' of his acts. The very nature of the causal explanation will focus attention on people directly and dynamically involved in the flow of abnormal events. This is unfortunate because they can very well be in a situation where they do not have the 'power of control.' Traditionally, a person is not considered in power of control when physically forced by another person or when subject to disorders such as e.g., epileptic attacks. In such cases, acts are involuntary (Fitzgerald, 1961; Feinberg, 1965), from a judgement based on physical or physiological factors. It is, however, a question as to whether cognitive, psychological factors should be taken more into account when judging 'power of control.' Inadequate response of operators to

unfamiliar events depends very much on the conditioning taking place during normal work. This problem also raises the question of the nature of human error. The behaviour of operators is conditioned by the conscious decisions made by work planners or managers. They will very likely be more 'in power of control' than an operator in the dynamic flow of events. However, their decisions may not be considered during a causal analysis after an accident because they are 'normal events' which are not usually represented in an accident analysis. Furthermore, they can be missed in analysis because they are to be found in a conditioning side branch of the causal tree, not in the path involved in the dynamic flow.

Present technological development toward high hazard systems requires a very careful consideration by designers of the effects of 'human errors' which are commonplace in normal, daily activities, but unacceptable in large-scale systems. There is considerable danger that systematic traps can be arranged for people in the dynamic course of events. The present concept of 'power of control' should be reconsidered from a cognitive point of view, as should the ambiguity of stop-rules in causal analysis to avoid unfair causal attribution to the people involved in the dynamic chain of events.

Analysis for System Improvements.

Analysis for therapeutic purpose, i.e., for system improvement, will require a different focus with respect to selection of the causal network and of the stop-rule. The stop-rule will now be related to the question of whether an effective *cure is known*. Frequently, cure will be associated with events perceived to be 'root causes'. In general, however, the effects of accidental courses of events can be avoided by breaking or blocking any link in the causal tree or its conditioning side branches. Explanatory descriptions of accidents are, as mentioned, focused on the unusual events. However, the path can also be broken by changing normal events and functions involved. The decomposition of the flow of events, therefore, should not focus on unusual events, but also include normal activities.

The aim is to find conditions sensitive to improvements. Improvements imply that some person in the system makes decisions differently in the future. How do we systematically identify persons and decisions in a (normal) situation when it would be psychologically feasible to ask for a change in behaviour as long as reports from accidents focus only on the flow of unusual events? An approach to such an analysis for improving work safety has been discussed elsewhere (Leplat and Rasmussen, 1984).

Another basic difficulty is that this kind of analysis for improvement presupposes a stable causal structure of the system, it does not take into account closed loops of interaction among events and conditions at a higher level of individual and organisational adaptation. A new approach to generalisation from analysis of the particular tokens of causal connections found in accident reports is necessary. The causal tree found by an accident analysis is only a record of one past case, not a model of the involved relational structure.

HUMAN ERROR, A STABLE CATEGORY?

A number of problems are met when attempts are made to improve safety of socio-technical systems from analyses tied to particular paths of accidental events. This is due to the fact that each path is a particular token shaped by higher order relational structures. If changes are introduced to remove the conditions of a particular link in the chain, odds are that this particular situation will never occur again. We should be fighting types, not individual tokens (Reason & Wagenaar, 1989). Human behaviour is constrained in a way that makes the chain of events reasonably predictable only in the immediate interface to the technical systems. The longer away from the technical core we are, the more degrees of freedom agents have in their mode of behaviour. Consequently, the less certain is also the reference in terms of normal or proper behaviour for judging 'errors'. In this situation, improvements of safety features of a socio-technical system depend on a *global, structural* analysis: No longer can we assume the particular traces of human behaviour to be predictable. Tasks will be formed for the occasion, and design for improvements must be based on attempts to find means of control at higher levels than the level of particular task procedures. If, for instance, socio-technical systems have features of adaptation and self-organisation, changes to improve safety at the individual task level can very well be compared to attempts to control the temperature in a room with a thermostat-controlled heater by opening the window. In other words, it is not sensible to try to change performance of a feedback system by changes inside the loop, you have to identify mechanisms that are sensitive, i.e., related to the control reference itself.

In traditional, stable systems human errors are related to features such as 1. conflicts among cognitive control structures and 2. stochastic variability, both of which can be studied separately under laboratory conditions. In modern, flexible and rapidly changing work conditions and socio-technical systems other features are equally important, such as 3. resource limitations which turn up in unpredicted situations and finally, 4. the influence of human learning and adaptation. In the present context, the relationship between learning and adaptation and the concept of error appears to be important.

Human Adaptation.

In all work situations constraints are found which must be respected to obtain satisfactory performance. There are, however, also many degrees of freedom which have to be resolved at the worker's discretion. In stable work conditions, know-how will develop which represents prior decisions and choice, and the perceived degrees of freedom will ultimately be very limited, i.e., 'normal ways' of doing things will emerge, and the process of exploration necessary for adaptation will no longer be messing-up the concept of error. In contrast, in modern, flexible and dynamic work conditions, the immediate degrees of freedom will have to be continuously resolved. This implies that effective work performance includes continuous exploration of the available degrees of freedom together with effective strategies for making choice, in addition to the task of controlling the chosen path to a goal.

Therefore, the basis of concept of error is changed in a very fundamental way.

The behaviour in work of individuals (and, consequently, also of) is, by definition, oriented towards the requirements of the work environment as perceived by the individual. Work requirements, *what* should be done, will normally be perceived in terms of control of the state of affairs in the work environment according to a goal, i.e., *why* it should be done. *How* these changes are made to a certain degree is a matter of discretion on part of the agent and cannot be predicted for a flexible environment.

The *alternative, acceptable work activities*, how to work, will be shaped by the work environment which defines the boundaries of the space of *possibilities*, i.e., acceptable work strategies. This space of possibilities will be further bounded by the resource profile of the particular agent in terms of tools available, knowledge (competence), information about state of affairs, and processing capacity. The presence of alternatives for action depends on a many-to-many mapping between means and ends present in the work situation as perceived by the individual; in general, several functions can serve the individual goals and each of the functions can be implemented by different tools and physical processes. If this was not the case, the work environment would be totally predetermined and there would be no need for human choice or decision.

Within the space of acceptable work performance found between the boundaries defined by the work requirements on one side and the individual resource profile on the other side, considerable degrees of freedom are still left for the individual to choose among strategies and to implement them in particular sequences of behaviour. These degrees of freedom must be eliminated by a choice made by an agent to finally enter a particular course of action. The different ways to accomplish work can be categorised in terms of *strategies*, defined as *types* of behavioural sequences which are similar in some well defined aspects, such as the physical process applied in work and the related tools or, for mental strategies, the underlying kind of mental representation and the level of interpretation of perceived information. In actual performance, a particular situation-dependent exemplar of performance, a *token,* will emerge which is an implementation of the chosen strategy under the influence of the complexity of detail in the environment. The particular token of performance will be unique, impossible to predict, whereas the strategy chosen will, in principle, be predictable. This choice made by the individual agents depends on *subjective performance criteria* related to the process of work such as time spent, cognitive strain, joy, cost of failure, etc. Normally, dynamic shifting among alternative strategies is very important for skilled people as a means to resolve resource-demand conflicts met during performance.

ADAPTATION, SELF-ORGANISATION AND ERROR.

It follows directly from this discussion that structuring the work processes by an individual in a flexible environment will be a self-organising, evolutionary process, simply because an optimising search is the only way in which the large number of degrees of freedom in a complex situation can be resolved. The basic synchronisation to the work requirements can be based

on procedures learned from an instructor or a more experienced colleague or it can be planned by the individual on occasion in a knowledge-based mode of reasoning by means of mental experiments. From here, the smoothness and speed characterising high professional skill together with a large repertoire of heuristic know-how rules will evolve through an adaptation process in which 'errors' are unavoidable side effects of the exploration of the boundaries of the envelope of acceptable performance. During this adaptation, performance will be optimised according to the individual's subjective process criteria within the boundary of his individual resources. This complex adaptation of performance to work requirements, eliminating the necessity of continuous choice will result in stereotype practices depending on the individual performance criteria of the agents. These criteria will be significantly influenced by the social norms and culture of the group and organisation. Very likely, conflict will be found between global work goals and the effect of local adaptation according to subjective process criteria. Unfortunately, the perception of *process quality* can be immediate and unconditional while the effect on *product quality* of the choice of an actor can be considerably delayed, obscure and frequently conditional with respect to multiple other factors.

In a first encounter, if representation of work constraints is not present in the form of instructions from an experienced colleague or a teacher, and know-how from previous experiences is not ready, the constraints of the work have to be explored in a knowledge-based mode from explicit consideration of the actual goal and a functional understanding of the relational structure of the work content. For such initial exploration as well as for problem solving during unusual task conditions, opportunity for test of hypotheses and trial-and-error learning is important. It is typically expected that qualified personnel such as process operators will and can test their diagnostic hypotheses conceptually - by thought experiments - before actual operations if acts are likely to be irreversible and risky. This appears, however, to be an unrealistic assumption, since it may be tempting to test a hypothesis on the physical work environment itself in order to avoid the strain and unreliability related to unsupported reasoning in a complex causal net. For such a task, a designer is supplied with effective tools such as experimental set-ups, simulation programs and computational aids, whereas the operator has only his head and the plant itself. In the actual situation, no explicit stop rule exists to guide the termination of conceptual analysis and the start of action. This means that the definition of error, as seen from the situation of a decision maker, is very arbitrary. Acts which are quite rational and important during the search for information and test of hypothesis may appear to be unacceptable mistakes in hindsight, without access to the details of the situation.

Even if a human actor is 'synchronised' to the basic requirements of work by effective procedures, there will be ample opportunities for refinement of such procedures. Development of expert know-how and rules-of-thumb depends on adaptation governed by subjective process criteria. Opportunities for experiments are necessary to find shortcuts and to identify convenient and reliable cues for action without analytical diagnosis. In other words, effective, professional performance depends on empirical correlation of cues to successful acts. Humans typically seek the way of least effort. Therefore, ex-

perts will not consult the complete set of defining attributes in a familiar situation. Instead it can be expected that no more information will be used than is necessary for *discrimination among the perceived alternatives for action* in the particular situation. This implies that the choice is 'under-specified' (Reason, 1986) outside this situation. When situations change, e.g., due to disturbances or faults in the system to be controlled, reliance on the usual cues which are no longer valid, will cause an error due to inappropriate "expectations." In this way, traps causing systematic mistakes can be designed into the system. Two types of errors are related to this kind of adaptation: The effect of the test of a hypothesis of a cue-action set which turn out negative, and the effects of acts chosen from familiar and tested cues when a change in system conditions make the cue unreliable.

Work according to instructions which take into consideration the possible presence of abnormal conditions that will make certain orders of actions unacceptable, presents an example in which local adaptation very likely will be in conflict with delayed and conditional effect on the outcome. The be safe, the instruction may require a certain sequence of the necessary acts. If this prescribed order is in conflict with the actor's immediate process criteria, modification of the prescribed procedure is very likely and will have no adverse effect in the daily routine. (If, for instance, an actor has to move back and forth between several, distant locations because only that sequence is safe under certain infrequent, hazardous conditions, his process criterion may rapidly teach him to group actions at the same location together because this change in the procedure will not have any visible effect under normal circumstances).

Even within an established, effective sequence of *actions*, evolution of patterns of *movements* will take place according to subconscious perception of certain process qualities. In a manual skill, fine-tuning depends upon a continuous updating of automated patterns of movement to the temporal and spatial features of the task environment. If the optimisation criteria are speed and smoothness, adaptation can only be constrained by the once-in-a-while experience gained when crossing the tolerance limits, i.e. by the experience of errors or near-errors (speed-accuracy trade-off). Some errors, therefore, have a function in maintaining a skill at its proper level, and they cannot be considered a separate category of events in a causal chain because they are integral parts of a feed-back loop. Another effect of increasing skill is the evolution of increasingly long and complex patterns of movements which can run off without conscious control. During such lengthy automated patterns attention may be directed towards review of past experience or planning of future needs and performance becomes sensitive to interference, i.e., capture from very familiar cues.

The basic issue is that human errors cannot be removed in flexible or changing work environments by improved system design or better instruction, nor should they be. Instead, the ability to explore degrees of freedom should be supported and means for recovery from the effects of errors should be found.

SYSTEM SAFETY, ADAPTATION, AND ERROR RECOVERY

The dynamic adaptation to the immediate work requirements both of the individual performance and of the allocation between individuals probably can be combined with a very high degree of reliability but only if errors are observable and reversible (i.e.; critical aspects are visible without excessive delay), and individual process criteria are not overriding critical product criteria.

System break-down and accidents are the reflections of loss of control of the work environment in some way or another. If the hypothesis is accepted that humans tend to close their degrees of freedom to get rid of choice and decision during normal work and that errors are a necessary part of this adaptation, the trick in design of reliable systems is to make sure that human actors maintain sufficient flexibility to cope with system aberrations, i.e., not to constrain them by an inadequate rule system. In addition, it appears to be essential that actors maintain 'contact' with hazards in a way that they will be familiar with the boundary to loss of control and will learn to recover (see the study of high reliable organisations described by Rochlin et al., 1989). In 'safe' system in which the margins between normal operation and loss of control are made as wide as possible the odds are that the actors will not be able to sense the boundaries and, frequently, the boundaries will then be more abrupt and irreversible. When radar was introduced to increase safety at sea, the result was not increased safety but more efficient transportation under bad weather conditions. Will anti-blocking car brakes increase safety or give more efficient transport together with more abrupt and irreversible boundaries to loss of control? A basic design question is: How can boundaries of acceptable performance be established that will give feedback to a learning mode in a reversible way, i.e., absorb violations in a mode of graceful degradation of the opportunity for recovery?

Under certain conditions self-organising and adaptive features will necessarily lead to 'catastrophic' system behaviour unless certain organisational criteria are met. Adaptation will normally be governed by local criteria, related to an individual's perception of process qualities in order to resolve the perceived degrees of freedom in the immediate situation. Some critical product criteria (e.g., safety) are conditionally related to higher level combination or coincidence of effects of several activities, allocated different agents and probably, in different time slots. *The violation of such high level, conditional criteria cannot be monitored and detected at the local criterion level,* and monitoring by their ultimate criterion effect will be too late and unacceptable. Catastrophic effects of adaptation can only be avoided if local activities are tightly monitored with reference to a *prediction* of their role in the ultimate, conditional effect, i.e., *the boundaries at the local activities are necessarily defined by formal prescriptions, not active, functional conditions.* (As argued below, the only possible source of this formal prescription is a quantitative risk analysis which, consequently, should be used as a risk management tool, not only as the basis for system acceptance.

This feature of adaptation to local work requirements probably constitutes the fallacy of the defence-in-depth design principle normally applied in high risk industries (Rasmussen, 1988b). In systems designed according to this principle, an accident is dependent on simultaneous violation of several

lines of defence: an operational disturbance (technical fault or operator error) must coincide with a latent faulty maintenance condition in protective systems, with inadequacies in protective barriers, with inadequate control of the location of people close to the installation etc. The activities threatening the various conditions normally belong to different branches of the organisation. The presence of potential of *a catastrophic combination of effects of local adaptation* to performance criteria can only be detected at a level in the organisation with the proper overview. However, at this level of the control hierarchy (organisation), the required understanding of conditionally dangerous relations cannot be maintained through longer periods because the required functional and technical knowledge is foreign to the normal management tasks at this level.

The conclusion of this discussion is that catastrophic system breakdown is a normal feature of systems which have self-organising features and at the same time, depend on protection against rare combination of conditions which are individually affected by adaptation. Safety of such systems depends on the introduction of locally visible boundaries of acceptable adaptation and introduction of related control mechanisms. What does this mean in terms of organisational structures? What kind of top-down influence from 'management culture' and bottom-up technological constraints can be used to guide and limit adaptation? *How can we model and predict evolution of organisational structure and the influence on system safety?*

CONTROL OF SAFETY IN HIGH HAZARD SYSTEMS

The trend towards large-scale industrial process plants and the related defence-in-depth design practice point attention to the need for a better integration of the organisation of plant design and of its operation. For large-scale, hazardous systems, the actual level of safety cannot be directly controlled from empirical evidence. For such installations, design cannot be based on experience gained from accidents, as it has been the case for accidents in minor separate systems when, for instance, considering work and traffic safety. Consequently, the days of extensive pilot plant tests for demonstration of the feasibility of a design are also gone and safety targets have to be assessed by analytical means based on empirical data from incidents and near misses, i.e., data on individual, simple faults and errors. For this purpose, large efforts have been spent on developing methods for probabilistic risk analysis for industrial process plants.

Typically, however, such risk analysis is considered only for the initial acceptance of a particular plant design. It is generally not fully realised that a risk analysis is only a theoretical construct relating a plant model and a number of assumptions concerning its operation and maintenance to a risk figure. This fact implies that following the acceptance of a plant on the basis of the calculated risk, the model and assumptions underlying the risk analysis should be considered to be specifications of the preconditions for safe operation which, in turn, should be carefully monitored by the operating organisation through the entire plant life (Rasmussen and Pedersen, 1984).

This use of a risk analysis raises some important problems. Risk analysis and, in particular, the underlying hazard identification are at present an art rather than a systematic science. We have systematic methods for analysing

specific accidental courses of events, the tokens. However, identification and characterisation of the types of hazards to analyse, in particular related to the influence of human activities during operation, maintenance and plants management, to a large extent depend upon the creativity and intuition of the analyst as it will be the case in any causal analysis. It is, therefore, difficult to make explicit the strategy used for hazard identification, the model of the system and its operating staff used for analysis, and the assumptions made regarding its operating conditions. Even if communication of causal arguments is unreliable between groups having different intuition such as designers and operations management, progress can be made, considering that the documentation of a risk analysis today is not designed for use during operations and maintenance planning and therefore is less accessible for practical operations management (Rasmussen, 1988b).

Another problem is the changing requirements to system management. Present organisation structures and management strategies in industry still reflect a tradition which has evolved through a period when safety could be controlled directly and empirically. The new requirements for safety control based on risk analyses have not yet had the necessary influence on the predominant organisational philosophy. The basic problem is that empirical evidence from improved functionality and efficiency is likely to be direct and unconditional, when changes are made to meet economic pressure in a competitive environment. In contrast, decrease of safety margin in a 'safe' system caused by local sub-optimisation, tends to be delayed and conditional and to require careful monitoring at higher management levels. Risk management requires a supplement of the traditional empirical management strategies by analytical strategies based on technical understanding and formal analysis.

CONCLUSION

The conclusion of the arguments presented are that the present rapid trend towards very large and complex systems, process plants as well as systems for financial operations and for information storage and communication, calls for a reconsideration of some of the basic approaches to system design and operation and to the role of human error in system safety. Some of the deficiencies presently attributed to operator or management deficiencies may very well be structural problems which have to be considered at a much more fundamental level than efforts to improve human reliability.

REFERENCES

Fitzgerald, P. J. (1961): Voluntary and Involuntary Acts. In: A. C. Guest (Ed.): Oxford Essays in Jurisprudence, Clarendon Press. Reprinted in: A. R. White (Ed.): The Philosophy of Action. Oxford Univ. Press

Feinberg, F. (1965): Action and Responsibility. In: M. Black(Ed.): Philosophy in America. Allen and Unwinn. Reprinted in: A.R. White (Ed.): The Philosophy of Action. Oxford Univ. Press

Kuhn, T. (1962): "The Structure of Scientific Revolution." University of Chicago Press, 1962.

Mackie, J. L. (1975): "Causes and Conditions." American Philosophical Quarterly, Vol. 2.4 pp. 245-255 & 261-264 Reprinted in: E. Sosa (Ed.): Causation and Conditionals, Oxford University Press.

Ramussen, J.. and J. Leplat: Analysis of Human Errors in Industrial Incidents and Accidents for Improvement of Work Safety. Accid. Anal. and Prev. Vol. 16, No. 2, pp.77-88, Pergamon Press Ltd., 1984.

Rasmussen, J. (1988a): Coping Safely with Complex Systems. American Association for Advancement of Science, invited paper for Annual Meeting, Boston, Februar 1988; In: Risø-M-2769.

Rasmussen, J. (1988b): Safety Control and Risk Management: Topics for Cross-Disciplinary Research and Development. Invited Key Note Presentation in International Conference on Preventing Major Chemical and Related Accidents. In: IChemE Publication Series No. 110; Washington: Hemisphere Publishing Corporation, 1988.

Rasmussen, J. and O. M. Pedersen (1984): Human Factors in Probabilistic Risk Analysis and in Risk Management. In: Operational Safety of Nuclear Power Plants. Vol. 1, pp. 181-194, IAEA, Wien, 1984.

Russell, B. (1913): "On the Notion of Cause". Proc. Aristotelean Society, Vol. 13, pp. 1-25.

Reason, J. (1989): Human Error: Causes and Consequences. New York: Cambridge University Press. In Preparation.

Reason, J. (1986): Cognitive Under-Specification: Its Varieties and Consequences. In B. Baars (Ed), The Psychology of Error: A Window on the Mind. New York: Plenum

Reason, J. and Wagenaar, W, (1989): Types and Tokens in Accident Causation, CEC workshop Workshop on Errors in Operation of Transport Systems; MRC-Applied Psychology Unit, Cambridge May 1989; To be published.

Rosch, E. (1975): Human Categorization. In: N. Warren (Ed.): Advances in Cross-Cultural Psychology. New York: Halsted Press.

Rochlin, G.I., La Porte, T.R., and Roberts, K.H., (1987): The Self-Designing High-Reliability Organization: Aircraft Carrier Flight Operations at Sea, Naval War College Review, Autom 1987.