Cognition, Technology & Work, 5, 272-282 (2003).

A COGNITIVE APPROACH TO SAFE VIOLATIONS.

Denis Besnard & David Greathead
denis.besnard@ncl.ac.uk david.greathead@ncl.ac.uk

Centre for Software Reliability
School of Computing Science
University of Newcastle upon Tyne
Newcastle upon Tyne, NE1 7RU
United Kingdom

Abstract: Classically, humans have been perceived as a source of faults in systems. Modern ergonomic views are promoting a somewhat different idea according to which humans are a factor of safety in unexpected situations. The safety of a system cannot be achieved without taking into account these two sides of cognition which compose what is called cognitive flexibility. In this paper, we will consider the cases of a nuclear accident and a plane crash-landing where human cognitive flexibility has impacted on the final safety of the system. We aim to discuss the violations that humans have performed in these cases with the assumption that they do not always deteriorate system safety. The discussion gravitates around a core argument according to which violations *per se* do not inform on the safety impairments in a system. Some other dimensions have to be taken into account. Among these, we are of the opinion that the accuracy of the operators' mental model plays a key role, allowing some violations to improve system safety in emergency situations.

Keywords: Large-scale systems safety, cognitive ergonomics, violations.

1 INTRODUCTION

Due to the increase of critical functions allocated to automatic agents (e.g. computers), the safety of socio-technical systems is an area where the stakes are continuously being raised. But reducing these systems down to a set of pure technical components would discard a very hot topic: deterministic automatic agents cohabit with non-deterministic human agents. As the actions of the latter can impact very strongly on the final safety of any system where they are present, it is worth questioning ourselves about the integration of humans into socio-technical systems. After Reason (1990; 1997), it is believed here that a combination of organisational arguments added to the identification of local individual factors offers an interesting analytical framework for discussing human actions. As our objective for this paper is analysing the impact of violations, we will look at system safety by linking the local individual cognitive components of actions to the organisational context in which they are embedded.

According to Reason (1990), violations can be seen as deliberate actions that deviate from the practices that designers and regulators have defined as necessary. The position defended here promotes that violations performed by humans at work are too often seen as generators of accidents. We will therefore emphasise a somewhat different view according to which violations can have a positive impact on system safety (Reason, 1997). We will discuss violations in rather neutral terms, as an expression of human cognitive flexibility. In the following section (section 2), we will present some classical concepts related to the role of regulators played by humans in systems. We will then come to the core of the paper and will consider violations (section 3). Precisely, we will

expose two case studies that shed some light on two opposite facets of violations, namely their contribution to impairing or enhancing system safety. The case studies will call for a careful discussion (section 4) where we suggest that the violations and the mental model held by the operators have to be considered together, along with the liberty that violations allow on the system's configuration. This position will drive our set of recommendations (section 5).

2 HUMANS AS COGNITIVE AGENTS IN SYSTEMS

Research in cognitive psychology made a big step forward by quantifying the limits of human memory and reasoning capacities. Important breakthroughs comprise the limits of short-term memory (Miller, 1956), goal and sub-goal decomposition (see the *General Problem Solver* by Newell, Shaw & Simon, 1957), reasoning biases when solving logical problems (Wason, 1966) and expert memory (Chase & Simon, 1973). From this fundamental work, a whole trend of research emerged in the 1980s which focussed on the human factors in the workplace. It then quickly became obvious that there was a need for zooming out from purely individual issues in order to encompass the complexity of the environment in which individuals act. A new approach named *cognitive ergonomics* then became targeted at understanding cognitive acts in the workplace, eventually exploring humans' potential contribution to system safety.

In the cognitive ergonomics' view, the information processing modes that humans implement at work are based on heuristic short-cuts built on top of the experience acquired through a life-long dynamic interaction with a diverse and changing environment. The resulting processing mode prioritises a trade-off between saving cognitive resources and perfect responses to the environment. This trade-off covers a very wide continuum that allows some room for errors and imperfections. However unsuitable to critical processes it may appear, this information processing strategy provides the flexibility that is required to perform and control uncertain actions in response to unknown problems. In the cognitive ergonomics conception, humans are no longer regarded as static components of a system. They are conceived as agents dedicating their mental resources to adapting themselves to varying environments, dealing with unknown situations and, as a result, contributing to system safety.

2.1 Mental models

When they interact with a system, humans need to understand what is currently happening and what is likely to happen next (Sarter & Woods, 1995). For this reason, they maintain a mental representation of the various ongoing and expected processes in a system. This representation is called a *mental model*. However, humans' memory and processing capacities have limits. Consequently, mental models are incomplete representations of reality. They are fed by a) a portion of the knowledge held about the system and updated by b) a selection of the data available in the environment. The selection is driven by the objectives that the operators have in the system (Rasmussen, 1986). For instance, aircraft maintenance crews operate on a different set of data than pilots, although both are actors in the same system. Thus mental models must be seen as incomplete, dynamic, goal-driven representations of reality (Ochanine, 1978) which partially reflect the system acted upon (Moray, 1987).

When they are adapted to a given situation, mental models contain the knowledge that is necessary to conduct an interaction, and data extracted from the environment. Via a proper updating process, valid mental models allow goals to be achieved in a pro-active manner. This is crucial in a dynamic system because its future states and the

consequences of one's actions are then anticipated, allowing operators to stay in step with the process they interact with.
Needless to say however, in today's extremely complex automated systems, humans can have incorrect mental models because of e.g. incomplete or erroneous knowledge, high pace of data flow, time pressure, etc. Moreover, incorrect models are not always detected as such by operators. Revision can be impaired by such mechanisms as fixation errors (De Keyser & Woods, 1990).

2.2 Control modes

Mental models are a concept depicting how humans represent the world. They determine the goals set by the operators and which actions will be performed to reach these goals. How the actions are actually selected, planned and executed depends on the control mode. The general assumption is that the more familiar a situation, the more likely it is that operators will rely on a skill-based mode of control. Conversely, the less familiar a situation, the more likely it is that operators will rely on a knowledge-based mode of control. Half-way between these two modes lies a third rule-based one (Rasmussen, 1986). We will not enter into very deep details about this well-known theoretical framework. Rather, our purpose is to emphasise that the familiarity of a given problem will call for a certain category of solution strategy. Precisely, routine problems call for shortcuts and heuristics. Such problems are dealt with in terms of a quick response to environmental signals (e.g. solving an over-heating problem by choosing the right gauges to read and triggering the right set of actions) with a very low mental load. Therefore, we will associate routine situations to the skill-based mode of control. On the contrary, some understanding has to be built for new or exceptional problems. In this case, the operator has to resort to a more declarative form of knowledge. In this case, the operator is said to treat the information as symbols feeding an inferential, costly, knowledge-based control mode.
The control modes are not exclusive from one another: they cohabit at all times. Only the proportion of the activity controlled under one mode or the other varies. Therefore, it is by pure simplification that a) we will rely on only two of them and that b) we will treat them as discrete notions.

Although it is not the only mechanism for it, the so-called flexible reasoning is supported by mental models and control modes that dynamically shape the interaction with the world. These two concepts will be used in the discussion of the two cases described in sections 3.1 and 3.2. So far, we have only relied on the well-known ideas that humans are a central component in systems' safety and that their activity is controlled by a potentially incomplete mental representation of reality. This is the standpoint from which we will investigate violations. We will claim that when reasoning about the safety of one's acts, the operators' intentions and how they address the constraints on a given situation have to be considered. Our purpose will be to defend that violations do not systematically lead to undesired events. When they are coupled with a valid mental model, they can ensure or even increase the safety level of a system. The following sections of this paper will aim at exposing this dual view.

3 VIOLATIONS

Violations have been mentioned or studied in a wide variety of contexts including car driving (Blockey & Hartley, 1995; Parker *et al.*, 1995; Aberg & Rimmo, 1998), aircraft piloting (Air France, 1997), large-scale accidents (Reason, 1990) computer programming (Soloway *et al.*, 1988) and bureaucratic environments (Damania, 2001). They are actions

that intentionally break procedures (Reason, 1987; Parker *et al.*, 1995), usually aiming at easing the execution of a given task. They may reveal the existence of faulty organisational settings when they are the only way to get the work done (Air France, 1997). In this latter case, these violations are the result of latent organisational factors leading to the rules or procedures being broken in order to accomplish a given task. These latent factors are usually implemented by actors who are remote (i.e. managers) from the resulting risks (Reason, 1995).
Reason (1990) distinguishes between several categories of violations. For the scope of our paper, we will focus on one dimension, namely routine violations as opposed to exceptional violations. The first type happens when an operator or a team regularly achieves a set of objectives by means which differ from prescribed procedures. In this case, the violation is often so deeply embedded into the daily practice that it is no longer identified as an illegal act. Such elements as following the path of least effort, managerial *laissez-faire* and badly designed procedures are contributing factors. The second category of violations (exceptional violations) happens when an operator or a team is performing an action in a context identified as a non-routine one, requiring some intentional departure from the prescribed practice. The objective may be to solve a problem that is not identified by the procedures. The cases depicted in the sections 3.1 and 3.2 respectively are instances of theses two types of violations.

As we stated previously, violations must not be directly associated with accidents. The latter take more than violations to happen: they have to be combined with errors. After Hollnagel (1993; see this author for a review of various classifications of errors), we see errors as a behaviour (or knowledge) that fails to produce the expected results and that may lead to unwanted consequences. Errors differ from violations in that the latter are merely behaviours that depart from some form of prescription (procedure, manual, rule, etc.) without necessarily leading to unwanted consequences. However, violations typically create specific unprotected conditions where recovering from or compensating for an error may no longer be possible. Major accidents in large-scale systems exhibit this combination (see for instance Gitus, 1988, about the Chernobyl accident), which is rooted in a variety of cultural, managerial and organisational factors (Cacciabue, 2000). In the following sub-sections, we will defend the idea that violations, under some conditions, can enhance system safety. Because we will analyse this as a cognitive phenomenon, we will account for the role played by mental models in the correctness of illegal actions. Our position is that without a correct mental model of a task and of the future system's states, violations can lead to accidents. On the other hand, a correct mental model may allow one to carry out safe departures from the rules, in emergency conditions for instance. It is this cognitive angle that we investigate in this paper, by focussing on the psychological aspects of systems' safety.
This position will be built upon two opposite case studies providing instances of routine and exceptional violations, respectively. The first case will depict an accident in a nuclear fuel processing plant in 1999, in Tokaimura, Japan. With this case, we will expose the harmful side of violations. We will oppose it with what we call *desirable* violations with the case of the crash-landing of a DC-10 in 1989, in Sioux City, Iowa, USA. However misleading the presentation of the cases may seem, we wish to make clear that our purpose is not to treat routine and exceptional violations as being respectively harmful and desirable. Surely, many examples of safe routine violations (e.g. successful adaptation of badly designed procedures) and dangerous exceptional violations (e.g. driving on the hard shoulder of the motorway to arrive on-time) exist.

3.1 Harmful violations: The Tokaimura criticality accident

There is a limited amount of uranium that can be put together without initiating fission. When this critical mass is exceeded, a chain reaction occurs, generating potentially lethal radiations. On December 30, 1999, in Tokaimura (Japan), a criticality accident occurred at the JCO nuclear fuel processing plant, causing the death of two workers. The immediate cause of the accident was the pouring of approximately 15kg of uranium into a precipitation tank, a procedure requiring mass and volume control (unless otherwise stated, the material in this section is from Furuta *et al.* (2000)).
The workers' task was to process seven batches of uranium in order to produce a uranium solution. The tank required to process this solution is called a buffer column. Its dimensions were 17.5 cm in diameter and 2.2 m high, owing to criticality safe geometry. The inside of this tank was known to be difficult to cleanse. In addition, the bottom of the column was located only 10 cm above the floor, causing the uranium solution to be difficult to collect. Thus, workers illegally opted for using another tank called precipitation tank (see Figure 1). This tank was 50 cm in diameter, 70 cm in depth and situated 1 m above the floor. Moreover, it was equipped with a stir propeller making it easier to use for homogenising the uranium solution.

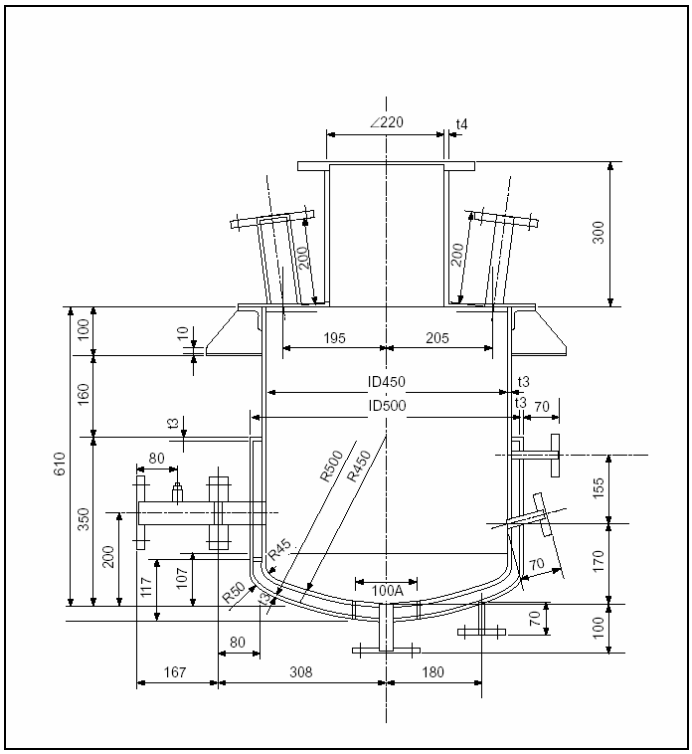

Figure 1: The precipitation tank at the JCO plant.

The workers thought it was not unsafe to pour the seven batches in the precipitation tank. This error caused the accident but the latter was rooted in a complex combination of deviant organisational practices. Among these featured the pressures from the managerial team to increase the production without enough regard to safety implications and crew training. This policy impacted on the safety culture developed by the workers, providing them with excessive liberty, even for critical procedures. The crews' practices were embedded in a work context where routine violations were constantly approved, leading to the implementation of what Westrum (2000) calls a pathological safety culture. Previous successful attempts at reducing the cycle time led uncontrolled actions to

become the norm at JCO (Blackman *et al.*, 2000). These management issues are discussed extensively in Furuta *et al.* (2000).

The JCO criticality accident was caused by a management-enabled violation being coupled with the operators' erroneous processing of uranium batches above the critical mass. This coupling of a violation with an error has been identified by Reason (1990) as a very powerful generator of accidents. Although the causes of this accident, as they are rooted at the managerial level, call for an analysis at the system level (Bieder, 2000), we suggest a complementary individual cognitive approach highlighting the role of violations.

In case of inappropriate use, precipitation tanks have already proven to be dangerous (Paxton, Baker & Reider, 1959). In using this tank for producing so much of the uranium solution, the crews have a) inaccurately assessed the situation, b) developed a flawed set of actions and c) ignored the consequences of such actions. These three components have been identified as important features in the control of dynamic systems (Sundstrom, 1993). In disregarding them, the crews have implemented what Marsden and Hollnagel (1996) have qualified as *opportunistic control*. But we must also acknowledge, after Wagenaar (1987), that accidents are not necessarily caused by humans gambling and losing. Accidents occur because people do not believe that the ongoing scenario is at all possible.

We would now like to point out that humans often operate illegal configurations of their work environment or procedures. In the case of the JCO plant, the workers used an illegal tank because the one they were supposed to use (the buffer column) could not help them respond to the production pressure from the managers. This sort of adaptation, orientated towards easing the work regardless of safety is very common and obeys an implicit rule of least effort to accomplish a given task. Having said that, the critical violations involved in accidents rarely happen instantly. They often are a sum of incremental departures from the prescribed practice. They initially take the form of a slight reconfiguration that eases the work and that is found acceptable by the operators. Modifications are then progressively added to the tools or practice, each increment being assessed as acceptable *per se*. After years of such progressive violations, the work settings can happen to be far beyond the prescribed practice. As Mancini (1987) asserts, large-scale accidents are made of a concatenation of small failures.

With this JCO case, we wish to highlight the workarounds that operators often implement in order to perform daily actions in a less constrained manner (see Gasser, 1986). This can be achieved in a wild manner and depending on the level of awareness, getting the work done sometimes overrides safety concerns. However, violations must not be considered as exceptional actions. They are extremely common practices aimed at saving time and/or effort in performing a given task. They can be seen as shortcuts that bypass some of the steps required by the procedures. They are also one of the features of the cognitive flexibility that allow humans to solve unexpected problems. When the consequences of one's actions are anticipated, violations can help implementing *ad hoc* set of actions allowing people to cope efficiently with exceptional situations. This issue will be addressed in the next section.

3.2 Desirable violations: The Sioux City emergency landing

On July 19, 1989, United Airlines flight 232 bound for Denver crash-landed at Sioux City Airport, Iowa. One hundred and twelve people were killed and 184 survived. The aircraft was forced to land after a metallurgical defect in the fan disc of the tail-mounted engine (#2) caused its catastrophic disintegration. The severity of this failure was such that the

engine's accessory drive system was destroyed, which resulted in a loss of hydraulic control. In addition, 70 pieces of shrapnel damaged the lines of the #1 and #3 engines (see Figure 2), resulting in a complete loss of hydraulic control. At the time of the accident, the loss of all three, independent hydraulic systems was considered a billion to one chance (unless otherwise stated, the material in this section is from NTSB (1990) and from Captain Haynes, pilot on the United Airlines flight 232 (Haynes, 1991)).

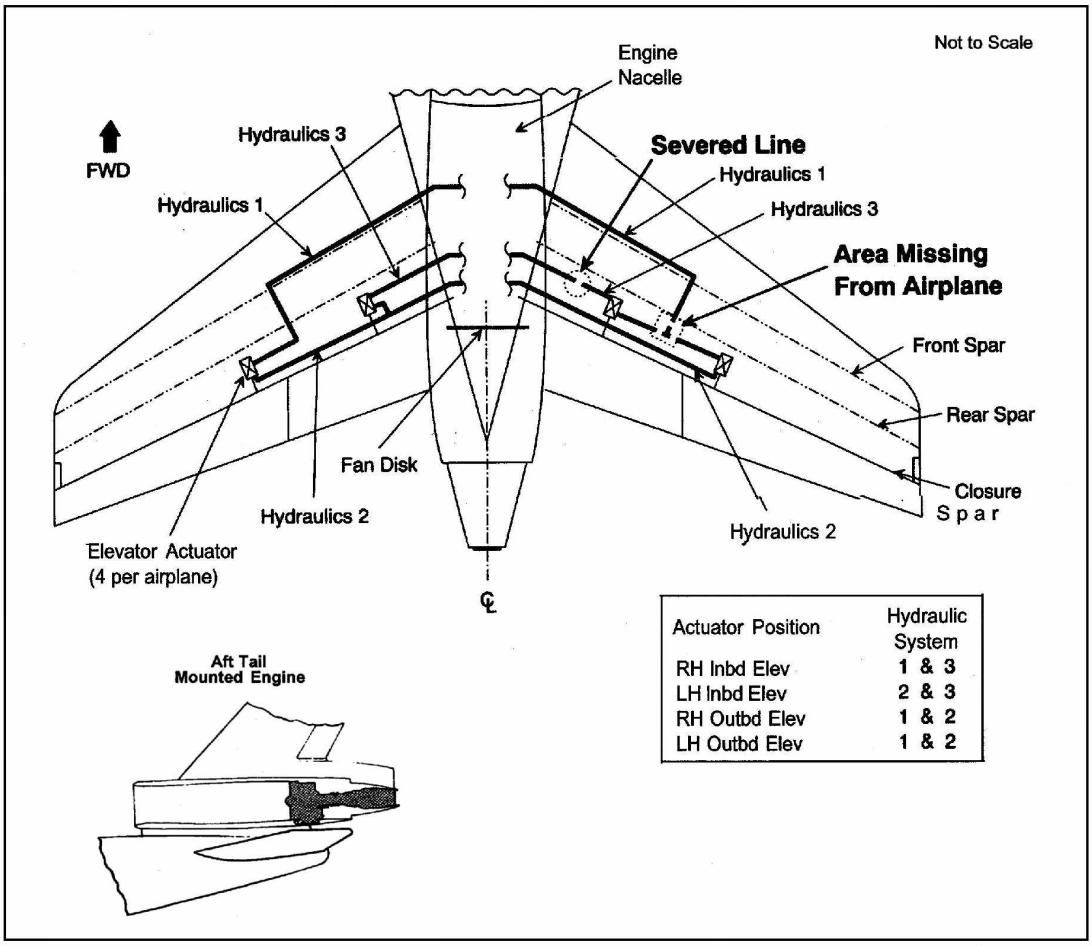

Figure 2: The damaged tail of the DC-10 (adapted from NTSB, 1990).

The damage to the hydraulic lines resulted in the crew having no control over ailerons, rudder, elevators, flaps, slats, spoilers, or steering and braking of the wheels. The only control which the crew had was over the throttle controls of the two, wing-mounted engines. By varying these throttle controls, they were able, to a certain extent, to control the aircraft. However, as revealed by the radar plot diagram (see Figure 3), the control over the vertical and horizontal axes were dramatically impaired. For instance, in order to correct a bank and stop the aircraft turning onto its back, they had to cut one throttle completely and increase the other. In addition to this problem, the crew also had to react to phugoids, i.e. oscillations in the vertical flight path whereby the aircraft repeatedly climbs and dives in association with fluctuations of speed. This was brought about as cutting the power to turn the aircraft caused the speed to drop. In turn, this caused the nose to drop and the aircraft to dive. The crew had to attempt to control this oscillation throughout the 41 minutes between the engine failure and the crash-landing. They needed to cut the throttles when the aircraft was climbing and approaching a stall (as increasing the throttles would cause the nose to raise further still). The crew also had to

increase the throttles when the aircraft began to dive (to increase the speed and bring the nose up). As both the pilot and the co-pilot were struggling with the yoke, they could not control the throttles. It is usually possible to control all three throttles with one hand. However, as the #2 engine had been destroyed, its throttle lever was locked and the remaining two levers, on either side of the jammed lever, had to be controlled with one hand each. Fortunately, another DC10 pilot was onboard as a passenger and was brought to the cockpit. This second pilot could then control the throttles allowing the pilot and co-pilot to control the yoke and the co-pilot to maintain communication with the ground. This is, understandably not common flying practice and several flying procedures were obviously violated on this flight.

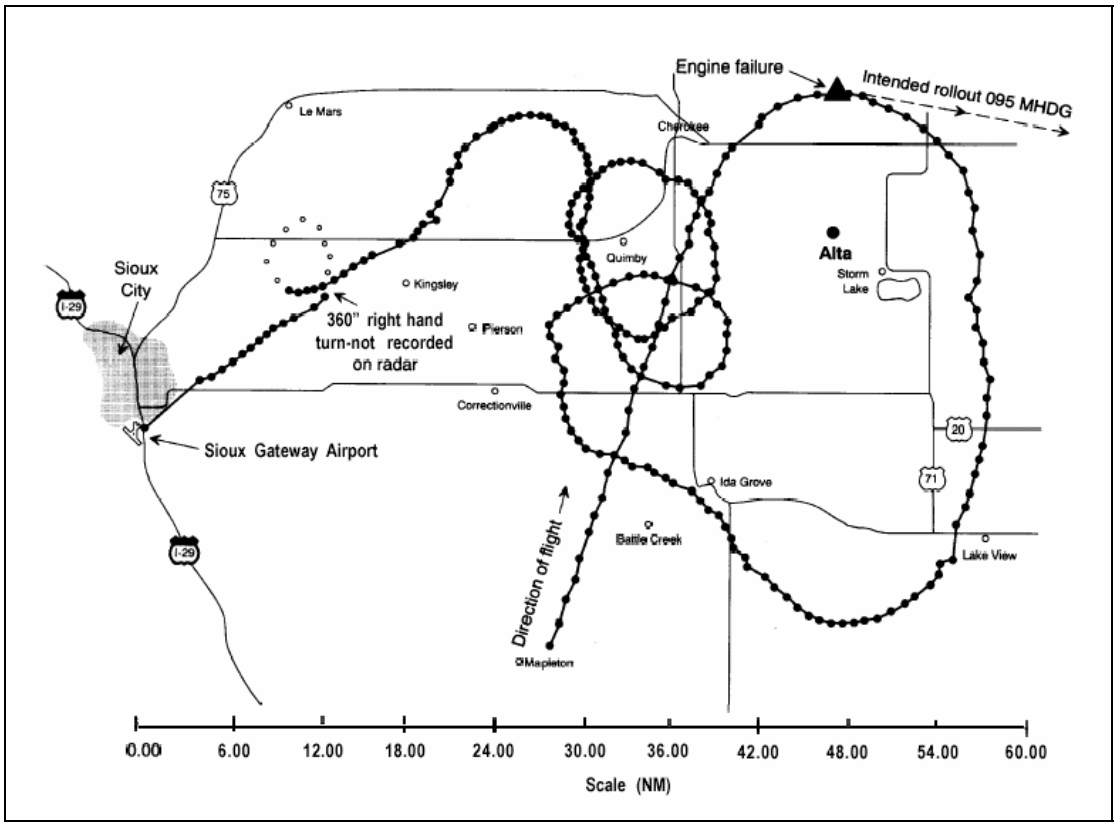

Figure 3: Radar plot diagram (NTSB, 1990).

By performing these violations, the crew were able to reach the airport –where the rescue teams where on standby- and save so many lives. It is unfortunate that the DC10 was on a 'down' phase of the phugoid when it landed as this resulted in the impact force being much greater. Nevertheless, this event exhibits the neutral nature of violations. These can be beneficial to system safety when they are coupled with a valid mental model. They allow operators to implement *ad hoc* control modes and to some extent, cope with unknown configurations.

In section 2.1, we have seen that mental models partially reflect the knowledge operators have of a system and are refined through a selection of environmental data. In this crash-landing case, the pilots used their knowledge of the aircraft's hardware to make the data displayed by the instruments converge towards a sensible representation of the situation. Updating a mental model is a crucial step in this kind of diagnosis-like activity and it can be flawed even among expert operators. This has been experimentally demonstrated among mechanics and electronics operators (Besnard, 2000; Besnard & Cacitti, 2001) and has been the cause of other air crashes (NTSB, 1997; METT, 1993). So it is fair to

say that the pilots of the DC-10 achieved a high level of performance. In comparison to the JCO operators, the pilots developed a more anticipative mode of control coupled with a more global and more functional view of the situation (Cellier, Eyrolle & Mariné, 1997).
Another contributing factor in the relative success of this crash-landing probably relies on the mental model sharing that the pilots established. This component of distributed decision taking (see Hollan, Hutchins & Kirsh, 2000) is a core activity in flight tasks (Doireau, Wioland & Amalberti, 1997). The transcripts of the dialogues inside the cockpit reveal at least two instances of such a distribution:

At 1552:34, the controller asked how steep a right turn the flight could make. The captain responded that they were trying to make a 30° bank. A cockpit crewmember commented, "I can't handle that steep of bank ...can't handle that steep of bank." (NTSB, 1990, p 22).

At 1559:58, the captain stated "close the throttles." At 1600:01, the check airman stated "nah I can't pull'em off or we'll lose it that's what's turnin' ya." (NTSB, 1990, p 23).

These two transcripts show that the pilots have a shared understanding of the situation. Each operator interprets the statements of the captain with regards to the limits of the controls that this pilot is acting upon. The decisions are shared among the crew members and the mental model that is supporting the piloting activity is composed of the knowledge of several agents. Indeed, at this time, United Airlines were advocating a policy whereby flight crews were encouraged to share information and opinions and not merely obey the captain without question. Finally, contrary to the JCO operators, the pilots understood very accurately the consequences of their actions although they were under strong time pressure. In the last extract of the transcripts, 18 seconds before touchdown, the captain asks for the throttles to be closed. This is the normal practice for landing a plane and this statement was probably released as a side effect of a rule-based behaviour. Interestingly enough, the operator controlling the throttles rejected the statement, arguing that the throttles were steering the aircraft. This is an example of a safe violation supported by a valid mental model. By implementing an action contrary to the usual procedure, one can nevertheless keep an already degraded system's state in reasonably safe boundaries.

The pilots' accurate mental model has led them to define viable boundaries for possible actions and allowed them to restore some form of control on the trajectory under strong time pressure and high risks. Controlling the aircraft on the basis of such a model afforded the implementation of positive desirable violations. The latter, although situated against the procedures, nevertheless exhibited a high degree of relevance.

4 IMPLICATIONS FOR THE SAFETY OF SOCIO-TECHNICAL SYSTEMS

One may object that JCO and the landing at Sioux City are different on too many dimensions to be compared to one another. For instance, JCO is an organisational failure leading to a routine violation whereas the Sioux City landing is a material failure leading to an exceptional violation. This makes two dimensions to consider, namely organisational *vs.* material and routine *vs.* exception handling.

- As far as the organisational *vs.* material dimension of the cases is concerned, if one a) has an accurate knowledge about the task at hand, b) has enough information available, c) and processes this information accurately, there are very few reasons why violations should be hazardous. Anticipating the correct future system's state *is* what matters. Whether or not crews are following the rules is of

little concern. This applies to the two cases we have considered. It is obvious for the Sioux city landing. But even at JCO where the managerial laissez-faire was a significant factor, knowing what kind of mistake they were about to make would have saved the operators.

- Let us now consider the routine *vs.* exception handling. The settings in which the operators were working at JCO were routine ones. The plant had run several production campaigns in the past and uranium solution production was just one of the services offered by the company. On the other hand, the conditions in which the DC-10 had to be landed were quite remote from nominal ones due to the many hydraulic systems' failures and the reconfigurations needed to continue the flight. Despite this difference in nature, the two cases remain comparable as far as the cognitive processes are concerned: The representation of the process at hand and the understanding of the consequences of actions were a matter of accident or (relative) success.

If one accepts to consider JCO and the Sioux City landing under the angle that we have adopted in this paper (i.e. the link between valid mental models and successful violations), it then seems acceptable to treat these two cases as comparable, despite their differences in location, time, technology, etc. In this respect, we believe that the mental processes involved in these two cases are relatively context-independent.

As Van der Schaaf (1992, quoted by Rauterberg, 1995) puts it, when an unexpected configuration restores or enhances the reliability level, then this positive system's state must be analysed to improve the functioning of the system. This is the spirit of this paper, supported by the example of the DC-10 crash-landing. And inevitably, in the context of this research, violations *per se* are not considered as harmful. Exceptions to this statement exist, e.g. in a system considered to be lost and upon which one performs a command or action whose consequences are not known, assuming that the system's state cannot be worse anyway. But we nonetheless believe that what is harmful is an action, legal or not, carried out without a full understanding of its consequences. So when discussing the impact of violations on systems safety, one has to take into account the mental model that operators hold. The two case studies exposed in this paper are two opposite instances of this argument.

We obviously accept the idea that many lives are saved thanks to pilots and operators correctly applying well-designed procedures for known emergency situations. This safety principle relies on experience: procedures account for past occurrences of events that have fallen into common knowledge. These procedures then prescribe a series of actions designed as an answer to already known conditions. This introduces a bias since high-pace, highly-critical systems, sometimes exhibit unexpected emergency settings for which no procedure exists. These conditions impose such a narrow span of legal actions that violating elementary rules is sometimes the only way to control the system. Following procedures under nominal, expected emergency conditions is a good interaction principle. However, if we think of low-probability, high-risk, unexpected situations, then the rules that stand for expected, standard situations may not always apply.

One lesson that can be learnt from violations in systems is that one should not expect humans to always act as prescribed. Procedures themselves do not rule human behaviour (Fujita, 2000) and there are many ways in which humans can configure a system and use it in unexpected and/or unprotected modes. The motivation for doing so may be based on a heuristic evaluation. If the intuitive cost/benefit trade-off in reconfiguring a system allows an operator to ease the accomplishment of a task, then it is likely that this reconfiguration will be performed, even at the cost of a violation. In this trade-off,

factors such as safety culture and risk perception are key notions. And again, whether or not the operator has a relevant knowledge of the potential consequences of his/her actions is what determines the level of risk involved.

4.1 A conservative safety culture

We have seen in section 3.1 that when flawed mental models combine themselves with violations, they can lead to serious impairments in safety. We have qualified these violations as harmful. As Reason (1990; 2000) and many others have pointed out, the existence of such violations is often caused by management flaws that propagate through the various layers of an organisation. As a consequence, a front-line operator causing an accident must not be regarded as an individual cognitive error but as a wider system failure. Even if the latter is not the approach we have adopted in this paper, we have to acknowledge that operators are too often blamed for having performed actions that a flawed cultural context or a bad management policy made inevitable. The picture may be even worse. According to Van der Schaaf (2000), rules in organisations are often developed simply to protect management from legal actions. Such alarming issues have already been raised by Rame (1995) who asserts that some incidents even lead to data obfuscation when human factors are involved. The legal side of enquiries and the individual blame policy that still prevail in the western European society can be put into question as well, especially when they clearly disregard non-individual factors leading to accidents (see for instance Svenson, Lekberg & Johansson, 1999).

4.2 *Ad hoc* reconfigurations

In our view, violations are actions that can be interpreted as *ad hoc* reconfigurations. In non-emergency situations, we conceive them as departures from the rules that informally express a need for different working practices or tools. But violations also occur in emergency situations where they help implementing recovery control modes on a system. So a strong warning has to be given to systems designers. If the human agents of a system are not able to perform violations, it may reveal that the protections against human undesired actions have risen up to the point where the human cognitive flexibility cannot be exploited any more. This is probably the kind of situation that inspired Bainbridge's (1983) *ironies of automation*. She suggests that the more advanced a control system, the more critical the role of human agents. This is potentially caused by the impossibility, beyond some point, to design perfect automated systems. This impossibility implies keeping the human agents inside the control loop in order to cope with potential unexpected events (Amalberti, 1996). Including humans in a system implies the acceptance of having them interacting with it in a manner that diverges from the specifications. Although it induces a risk, it exploits their capacity to handle these unexpected events that require *ad hoc* reconfiguration. This is a function that is extremely difficult to implement in machines and is widely accepted as being a typical human skill.

It is an intriguing fact that we seem to be more prepared to accept these violations when they lead to a happy end rather than when they cause an accident. Instead, they should be seen as the two facets of the same coin. In the end, as Woods and Shattuck (2000) suggest, the design options range from a centralised control inhibiting actors' adaptation to variability, to local actors' complete autonomy disconnecting the hierarchy from any decision taking. Obviously, the final safety of a system will rely on the right balance between these two extreme points.

As far as the actual design is concerned, Woods (1993, quoting his 1986 work) suggests a two-fold view. *"The tool maker may exhibit intelligence in shaping the potential of the artefact relative to a field of practice. The practitioner may exhibit intelligence in tailoring his activity and the artefact to*

the contingencies of the field of activity given his goals". This highlights the dual view that one has to have about human agents in systems. Some people design tools, others use and reshape them so that the latter fit their intentions better, so to speak. This reshaping activity by users has been identified by Wimmer, Rizzo & Sujan (1999) as a source of valuable data that design teams must try to capture.

Although not all violations are desirable, preventing humans from performing any is not the issue. The point is letting them configure the system at the condition that they are trained and have enough understanding of the risks associated with their actions (Fujita, 2000). This correlates with Reasons' (2000) view about high-reliability organisations: Human compensations and adaptations to changing events is one of the most important safeguards. In this conception, violations can contribute to make a system safer. If operators have sufficient knowledge and available cognitive resources, they can implement an anticipative mode of control which is a pre-requisite for a safe interaction with dynamic real-time systems. In such conditions, human agents are able to conduct a safe *ad hoc* interaction in the case of e.g. emergency situations that were not expected by designers (Cf. section 3.2). Then, the flexibility of the human operator can maintain or improve the safety of a given system by enlarging the span of the control that he or she has on it.

4.3 Violations and control modes

JCO and Sioux City are, to some extent, cases where operators respectively needed to reduce production time and gain a higher level of control. This has been achieved by transgressing rules and procedures. From a control mode point of view, the JCO routine violations might be associated with a skill-based control mode whereas the emergency Sioux City landing might have relied on a knowledge-based control mode. The rationale for this classification derives from the very routinised job of diluting uranium at JCO, and the extremely unlikelihood of landing a plane without hydraulics. From this standpoint, routine violations appear to be extremely hazardous since they involve highly encapsulated cognitive processes that require very little control. The consequence may be to routinely depart from procedure in undetected exceptional settings. This is an already documented phenomenon whereby experienced operators (e.g. trouble-shooters; see Besnard, 2000) fall into a frequency bias and fail to identify non-routine conditions when solving a problem. On the other hand, the violations implemented aboard the DC-10 are safer in our opinion, due to the crew being aware that they were in exceptional flying conditions.

Classically, the skill-based control mode has been said to be reliable since it is built from the experience gained in a given class of problems. On the other hand, the knowledge-based control mode has been introduced as a fallible one due to cognitive cost. Although the above appears to be true in the vast majority of cases, the two accidents presented in this paper provide counter-examples to these assumptions. One explanation could be as follows. Although skill-driven violations and knowledge-driven violations trigger the same level of hazard, the difference lies in the fact that that expert operators resorting to a knowledge-based level of control are doing so as an answer to the recognised exceptional nature of the problem at hand. Therefore, it may be that they are more aware of the impact of an error and invest more effort into understanding the consequences of their acts. Having said this, and as stated in section 3, it is doubtless that counter-examples exist where routine violations are successful and exceptional violations are catastrophic.

5 RECOMMENDATIONS

After Cacciabue and Kjaer-Hansen (1993), we think that a design team designing systems interacting with humans should bring together a variety of skills, including engineers, computer scientists and psychologists. As this research originates from an interdisciplinary research project on dependable computer-based-systems (visit DIRC at http://www.dirc.org.uk) the authors are rather sensitive to this kind of argument and will make the following recommendations rely on this principle.

Within the scope of this paper, we do not believe that humans are deterministic agents. If they indeed are, it then seems reasonable to say that we are not yet able to understand the underlying deterministic mechanism. It follows that humans sometimes act in ways that are beyond our prediction power. We must therefore design the workplace as a whole accordingly and expect humans to adopt unanticipated configurations of the system, including its tools and procedures. In this paper, we have defended the idea that whether a violation increases the control over a given task depends on the extent to which the operator's mental model accurately reflects and predicts the system's present and future states of the process. This defines two potential areas for improvements: the rationale for the initial violation and the content of the mental model itself. The two following recommendations address these issues.

5.1 Design workable instead of exhaustive procedures

Organisations must understand the reasons behind the gap between procedures and practice (Dekker, 2003). From this standpoint, errors that are coupled with violations must not always be interpreted as incompetence (Rizzo, Ferrante & Bagnara, 1995). Instead, they sometimes highlight the need for improvements such as more workable procedures. Wishing to design perfect and exhaustive procedures is not feasible. If this were the case, we could replace human agents with one form of automation or another. The reality is that procedures and rules are intrinsically incomplete in the sense that there will always be an unforeseen event for which the safety of a system will rely on human intervention. Although designing exhaustive rules and procedures can stem from the assumption that it will enhance systems' reliability, the outcome may actually be the opposite. Humans will implement workarounds (Gasser, 1986) in order to by-pass rules that are too costly to follow or unworkable. In our opinion, a step towards a solution is to design more usability-centred rules and procedures i.e. ones for which humans will understand the rationale and will conform to. In this respect, the workplace should be constrained by and designed according to the way humans think and act, not the other way around. Several ways to "do the job" should be supported, accounting for different styles and/or levels of expertise. For instance, different types of assistance and protection should be provided to the operators, depending on their experience in a given task. Ultimately, it is only by designing artefacts and procedures that match the operator's operational objectives and practices (a counter-example is the buffer column at JCO) that humans will accept to act within defined boundaries.

5.2 Operators must have accurate mental models, not only rules to follow

When hazardous unanticipated events arise, the human agent is the last barrier before the accident. It is therefore eminently important that operators could hold a representation of the system's behaviour which is compatible with what is actually happening. This is where mental models impact safety. The JCO case shows that violations alone are not the problem. The coupling of the latter with a flawed mental model *is* the problem since

it prevents the operator from understanding the current state of the system, therefore degrading the anticipation of future events. In this context, improving the accuracy of mental models is mandatory. It implies that a) adequate education is given to the operator through e.g. training schemes, and that b) the system provides a realistic, workable picture of its behaviour. This second point seems of high relevance to the reliability of human-automation interaction and is specifically targeted to systems designers. One difficulty that the JCO operators were faced with is the lack of visibility of the critical state the system had entered. So one suggestion is to create systems whose complexity can be made compatible with some form of transparency. For instance, systems with short delays of feedback and that display tangible changes in state may improve the representation that operators have of the behaviour of this system. The fact that the degraded behaviour of the DC-10 was obvious to the pilots helped them to adopt the right control mode.

The next generation of support tools should be anticipative by nature. As suggested by Kanno *et al.* (2003), they should be pro-actively driven by the dialogue between the operator and the system instead of being reactively displaying information on request. This line of thought is addressed in the next section

6 A LOOK FORWARD: ANTICIPATIVE SYSTEMS NEEDED

Hollnagel and Woods (1999) assert that the goal of designing a man-machine system should be that of making the interaction between the operator and the system as smooth and efficient as the interaction between two persons. An essential part of human communication is that each participant is able to continuously anticipate and modify his or her model of the other. So after Amalberti's (1992) concerns, we think machines should account for human operators' context dependency. There may be enough knowledge in ergonomics and enough computational resources available in modern control systems to allow the implementation of screening functions dedicated to analyse human actions (as already suggested by Rasmussen, 1991). Such screening functions could lead to e.g. the building and loading of a cognitive profile of the various users of the system. This kind of assistance tool would dynamically and specifically support a given operators' reasoning, provide synthetic shots on the system's state, anticipate which action is now required, which information will be needed next, etc. Also, we believe the next generation of support tools should be able to provide a) assistance for unexpected emergency situations as well as b) anticipative protection measures for acts identified as hazardous. This last point may be a way to foresee human-machine interaction flaws such as mode confusion (Crow, Javaux & Rushby, 2000; Leveson *et al.*, 1997; Rushby 2001; Rushby, Crow & Palmer, 1999) by predicting mismatches between the operator's mental model and the system's state.

Operators need more help on exceptional situations for which they have not been trained rather than on nominal settings. However ambitious it may seem, this vision of the world is just one where classical ideas are stretched beyond our current knowledge. According to Cacciabue (1991) and Hollnagel (1987), tools should fully support human decision making and improve system's safety. The future reliability of the interaction between human agents and critical systems may depend on how far we succeed in extending this kind of principles and how diverse are the systems in which we can turn these principles into a tangible design policy.

7 CONCLUSION

In this paper, after recalling some views on the contribution of human agents to system safety, we have been concerned with violations and the way they impact on systems. We

have defended the idea that violations *per se* do not explain the occurrence of accidents. The accuracy of operators' mental models is a key factor. To support our claim, we have exploited the Tokaimura criticality accident and the Sioux City crash-landing cases. Our analysis revealed that violations can generate very different outcomes depending on whether they are coupled with valid or invalid representations of reality. Therefore, we believe that violations cannot be assessed without taking into account the level of understanding of the operators. As a consequence, violations should be supported rather than prevented as they are a real opportunity for human operators to recover from degraded situations. As improvements to socio-technical systems' dependability, we recommend that a) rules and procedures should not aim at being exhaustive but rather workable, and that b) training associated to system's transparency can improve the validity of mental models. Lastly, we believe that the next generation of support tools should be anticipative. In this respect, it is obvious to us that multiple disciplines (e.g. ergonomics, cognitive psychology, computer science, engineering) are required in design teams if we want humans' cognitive abilities to be supported in unexpected, critical conditions.

8 REFERENCES

Aberg, L. & Rimmö, P.-A. (1998). Dimensions of aberrant behaviour. *Ergonomics*, 41, 39-56.

Air France (1997). Anatomie d'un accident. F-28 Dryden, Canada, Mars 1989. *Bulletin d'information sur la Sécurité des Vols*, 36, 2-7.

Amalberti, R. (1992). Safety and process control: An operator-centered point of view. *Reliability Engineering and System Safety*, 38, 99-108.

Amalberti, R. (1996). *La conduite de systèmes à risques*. Paris, Presses Universitaires de France.

Bainbridge, L. (1983). Ironies of automation. *Automatica*, 19, 775-779.

Besnard, D. & Cacitti, L. (2001). Troubleshooting in mechanics. A heuristic matching process. *Cognition, Technology & Work*, 3, 150-160.

Besnard, D. (2000). Troubleshooting in electronics. In F. Kornneef & M. van der Meulen (Eds). *Computer safety, reliability and security*. Proceedings of SAFECOMP 2000, Springer-Verlag, Heidelberg (pp. 74-85).

Bieder, C. (2000). Comments on the JCO accident. *Cognition, Technology & Work*, 2, 204-205.

Blackman, H. S., Gertman, D. & Hallbert, B. (2000). The need for organisational analysis. *Cognition, Technology & Work*, 2, 206-208.

Blockey, P. N. & Hartley, L. R. (1995). Aberrant driving behaviour: Errors and violations. *Ergonomics*, 38, 1759-1771.

Cacciabue, P. C. & Kjaer-Hansen, J. (1993). Cognitive modelling and human-machine interactions in dynamic environments. *Le Travail Humain*, 56, 1-26.

Cacciabue, P. C. (1991). Cognitive ergonomics: A key issue for human-machine systems. *Le Travail Humain*, 54, 359-364.

Cacciabue, P. C. (2000). Comments on the HF analysis of the JCO criticality accident. *Cognition, Technology & Work*, 2, 209-211.

Cellier, J. M., Eyrolle, H. & Mariné, C (1997). Expertise in dynamic systems. *Ergonomics*, 40, 28-50.

Chase, W. G. & Simon, H. A (1973). Perception in chess. *Cognitive Psychology*, 4, 55-81.

Crow, J., Javaux, D. & Rushby, J. (2000). Models of mechanised methods that integrate human factors into automation design. *International Conference on Human-Computer Interaction in Aeronautics: HCI-Aero 2000*, Toulouse, France.

Damania, R. (2002). Environmental policies with corrupt bureaucrats. *Environment and Development Economics*, 7, 407-427.

De Keyser, V. & Woods, D. D. (1990). Fixation errors: Failures to revise situation assessment in dynamic and risky systems. In A. G. Colombo & A. Saiz de Bustamante (Eds.) *Systems reliability assessment*, ECSC, EEC, EAEC, Brussels and Luxembourg (pp. 231-251).

Dekker, S. (2003). Failure to adapt or adaptations that fail: contrasting models on procedures and safety. *Applied Ergonomics*, 34, 133-238.

Doireau, P., Wioland, L. & Amalberti, R. (1997). La détection d'erreurs humaines par des opérateurs extérieurs à l'action: le cas du pilotage d'avion. *Le Travail Humain*, 60, 131-153.

Fujita, Y. (2000). Actualities need to be captured. *Cognition, Technology & Work*, 2, 212-214.

Furuta, K., Sasou, K., Kubota, R., Ujita, H., Shuto, Y. & Yagi, E. (2000). Analysis report. *Cognition, Technology & Work*, 2, 182-203.

Gasser, L. (1986). The integration of computing and routine work. *ACM Transactions on Office Information Systems*, 4, 205-225.

Gitus, J. H. (1988). *The Chernobyl accident and its consequence*s. London, United Kingdom Atomic Energy Authority.

Haynes, A. (1991). Transcript of the presentation given at the NASA Ames Research Centre, May 24th, 1991. http://www.panix.com/~jac/aviation/haynes.html

Hollan, J., Hutchins, E. & Kirsh, D. (2000). Distributed cognition: toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction*, 7, 174-196.

Hollnagel, E. & Woods, D. (1999). Cognitive system engineering: New wine in new bottles. *International Journal of Human-Computer Studies*, 51, 339-356.

Hollnagel, E. (1987). Information and reasoning in intelligent decision support systems. *International Journal of Man-Machine Studies*, 27, 665-678.

Hollnagel, E. (1993). The phenotype of erroneous actions. *International Journal of Man-Machine Studies*, 39, 1-32.

Kanno, T., Nakate, K. & Furuta, K. (2003). A method for team intention inference. *International Journal of Human-Computer Studies*, 393-413.

Leveson, N., Pinnel, L. D., Sandys, S. D., Koga, S. & Reese, J. D. (1997). Analysing software specifications for mode confusion potential. in C. W. Johnson (Ed) *Proceedings of a workshop on human error and system development*, Glasgow, Scotland (pp. 132-146).

Mancini, G. (1987) Commentary: Models of the decision maker in unforeseen accidents. *International Journal of Man-Machine Studies*, 27, 631-639.

Marsden, P. & Hollnagel, E. (1996). Human interaction with technology. The accidental user. *Acta Psychologica*, 345-358.

METT (1993). *Rapport de la commission d'enquête sur l'accident survenu le 20 Janvier 1992 près du Mont Sainte-Odile a l'Airbus A.320 immatriculé F-GGED exploité par la compagnie Air Inter.* Ministère de l'Equipement, des Transports et du Tourisme (French Ministry of Equipment, Transports and Tourism).

Miller, G. A. (1956). The magical number seven plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, 63, 81-97.

Moray, N. (1987). Intelligent aids, mental models, and the theory of machines. *International Journal of Man-Machine Studies*, 27, 619-629.

Newell, A., Shaw, J. C. & Simon, H. A. (1957). *Preliminary description of General Problem Solving* -I (GPS-I). Technical Report CIP, Working Paper 7, Carnegie Institute of Technology, Pittsburgh, PA, USA.

NTSB (1990). *Aircraft accident report. United Airlines flight 232. Mc Donnell Douglas DC-10-10. Sioux Gateway airport. Sioux City, Iowa, July 19, 1989*. National Transportation Safety Board, Washington DC, USA.

NTSB (1997). Wheels-up landing, Continental Airlines flight 1943, Douglas DC-9 N10556, Houston, Texas, February 19, 1996. National Transportation Safety Board, Washington DC, USA. http://www.ntsb.gov/Publictn/1997/AAR9701.pdf

Ochanine, D. (1978). Le rôle des images opératives dans la régulation des activités de travail. *Psychologie et Education*, 2, 63-72.

Parker, D., Reason, J., Manstead, S. R., & Stradling, S. G. (1995). Driving errors, driving violations and accident involvement. *Ergonomics*, 38, 1036-1048.

Paxton, H. C., Baker, R. D. & Reider, W. J. (1959). Los Alamos criticality accident. *Nucleonics*, 17, 107-.

Rame, J.-M. (1995). Rôle des industriels dans la prévention des accidents. *Pilote de ligne*, 5, 20-21.

Rasmussen, J. (1986). *Information processing and human-machine interaction*. Amsterdam, North Holland.

Rasmussen, J. (1991). Technologie de l'information et analyse de l'activité cognitive. In R. Amalberti, M. de Montmollin & J. Theureau. *Modèles en analyse du travail*. Liège, Mardaga (pp. 49-73).

Rauterberg, M. (1995). About faults, errors and other dangerous things. In H. Stassen & P. Wieringa (Eds) *Proceedings of XIV European Annual Conference on Human Decision Making and Manual Control* (Session 3-4, pp. 1-7). Delft, Delft University of Technology.

Reason, J. (1987). Chernobyl errors. *Bulletin of the British Psychological Society*, 40, 201-206.

Reason, J. (1990). *Human error*. Cambridge, Cambridge University Press.

Reason, J. (1995). A systems approach to organisational errors. *Ergonomics*, 1708-1721.

Reason, J. (1997). *Managing the risks of organisational accidents*. Aldershot, Ashgate.

Reason, J. (2000). Human error: Models and management. *British Medical Journal*, 320, 768-770.

Rizzo, A., Ferrante, D. & Bagnara, S. (1995). Handling human error. In J.-M. Hoc, P. C. Cacciabue & E. Hollnagel (Eds) Expertise and technology. *Cognition and human-computer interaction*. Hillsdale, N. J., Lawrence Erlbaum.

Rushby, J. (2001). Modelling the human in human factors. Invited paper, *Safecomp 2001*, Budapest, Hungary (pp. 86-91).

Rushby, J., Crow, J. & Palmer, E. (1999). An automated method to detect potential mode confusions. Proceedings of the *18th AIAA/IEEE Digital Avionics Systems Conference*, St Louis, MO, USA.

Sarter, N. & Woods, D. D. (1995). How in the world did we ever get into that mode? Mode error and awareness in supervisory control. *Human Factors*, 37, 5-19.

Soloway, E., Adelson, B. & Ehrlich, K. (1988). Knowledge and processes in the comprehension of computer programs. in M. T. H. Chi, R. Glaser & M. J. Farr *The nature of expertise*. Hillsdale, NJ : Lawrence Erlbaum.

Sundstrom, G. A. (1993). Towards models of tasks and task complexity in supervisory control applications. *Ergonomics*, 11, 1413-1423.

Svenson, O., Lekberg, A. & Johansson, A. E. L. (1999). On perspective, expertise and differences in accident analyses: Arguments for a multidisciplinary approach. *Ergonomics*, 42, 1567-1571.

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

Van der Schaaf, T. (2000). Near miss reporting changes the safety culture (Report after a visit to the University of Wisconsin-Madison), *The Human Element*, 5, 1-2. http://www.engr.wisc.edu/centers/chpra/newsletter/CHPCS_vol5.1.pdf

Wagenaar, W. A. & Groeneweg, J. (1987). Accidents at sea. Multiple causes and impossible consequences. *International Journal of Man-Machine Studies*, 27, 587-598.

Wason, P. C. (1966). Reasoning. In B. M Foss (Ed). *New horizons in psychology*. Harmondsworth, UK. Penguin.

Westrum, R. (2000). Safety planning and safety culture in the JCO criticality accident: Interpretative comments. *Cognition, Technology & Work*, 2, 240-241.

Wimmer, M., Rizzo, A. & Sujan, M. (1999). A holistic design concept to improve safety-related control systems. In M. Felici, K. Kanoun & A. Pasquini (Eds) *SAFECOMP'99*, Springer-Verlag, Heidelberg (pp. 297-309).

Woods, D. D. & Shattuck, L. G. (2000). Distant supervision-local action given the potential for surprise. *Cognition, Technology & Work*, 2, 242-245.

Woods, D. D. (1986). Paradigms for intelligent decision support. In E. Hollnagel, G, Mancini & D. D. Woods (Eds) *Intelligent decision support in process environments*, New-York, Springer Verlag.

Woods, D. D. (1993). The price of flexibility. *Proceedings of the International Workshop on Intelligent User Interfaces*, Orlando, Florida (pp. 19-25).

9 ACKNOWLEDGEMENTS

This paper was written at the University of Newcastle upon Tyne within the DIRC project (`http://www.dirc.org.uk`), a UK-based interdisciplinary research collaboration on the dependability of computer-based systems. The authors wish to thank Gordon Baxter (University of York) and anonymous reviewers for useful comments and the sponsor EPSRC for funding this research.