A COGNITIVE APPROACH TO SAFE VIOLATIONS.

Denis Besnard & David Greathead

Centre for Software Reliability
Department 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 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. However, this *socio-technical* label does not imply that sociology is the sole approach suitable for studying the role of humans in system safety. Even if a view at the system level cannot be avoided, cognitive psychology can be an interesting complementary approach for isolating individual factors. 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. It may be even more suitable for the study of deviant actions since the latter can be generated by factors that lay outside of the boundaries of the physical workplace. As our objective for this paper is highlighting the impact of a class of deviant acts (i.e. 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.

After Reason (1990), violations can be seen as deliberate actions that deviate from the practices that designers and regulators have defined as necessary[1]. The position defended in this paper promotes that violations performed by humans at work are too often seen as generators of accidents. Although this view has reigned in cognitive psychology throughout recent decades and has indeed allowed enrichment of the analysis of large-

scale accidents, we would like to emphasise a somewhat different view according to which violations can have a positive impact on system safety (Reason, 1997). We will thus discuss violations in rather neutral terms, as an expression of human cognitive flexibility. In the following section (section 2), we will very quickly consider some fundamentals about cognitive psychology. In doing so, we will oppose a classical view (that has studied cognitive limitations) to a more recent one (that has highlighted the human contribution to systems' regulation). After the presentation of these two views, we will 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 that operators run 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 HUMAN AGENTS IN SYSTEMS

2.1 The classical view: Humans' cognitive limitations[2]

The *cognitive* approach made a big step forward in psychology by quantifying the limits of human reasoning capacity. Experimentally speaking, the revolution came from a seminal work carried out by Miller (1956) in quantifying the limits of short-term memory. His *magical number seven plus or minus two* has strongly influenced the psychology of memory. Later, on the processing side, Wason (1966) showed that humans are submitted to biases when solving logical problems. This was the beginning of a change in the conception of human reasoning. Humans were no longer simulated as wet machines performing logical operations (see the *General Problem Solver* by Newell, Shaw & Simon, 1957) but considered as fallible information processors obeying biased heuristics. On the social side of psychology, such a demonstration had already been done as early as 1959 when Festinger and Carlsmith published their theory of cognitive dissonance. With this innovative work, they demonstrated that humans can distort their discourse or beliefs in order to make the latter coherent with previously performed unwanted actions. Some years later, Kahneman and Tversky (1973), still in a social psychology line, demonstrated that people's judgement is influenced by subjective perceived likelihood rather than by objective base rates. Many other biases were discovered and studied extensively after that date. Still in 1973, Chase and Simon pioneered some experimental work in a green field area: chess playing. They shook the traditional wisdom according to which experts had the highest performance in a given area. Chase and Simon demonstrated that expert chess players performance at recalling random game configurations, due to their skills being supported by a highly task-specific memory, could drop to the level of novices.

Studying and quantifying human limits in reasoning has provided an extremely valuable amount of knowledge on cognition. From this fundamental work, a whole trend of research emerged in the 1980s which focussed on cognitive factors at the workplace. But it quickly became obvious that there was a need for zooming out from purely individual issues in order to encompass the complexity of the context in which individuals act. A new approach named *cognitive ergonomics* then became targeted at understanding the cognitive acts at the workplace, eventually discovering and exploring humans' potential contribution to system safety. This will be the topic of the next section.

2.2 The cognitive ergonomics view: Humans as regulators of systems

Following the nuclear accident in Chernobyl in 1986, Reason (1987) produced a paper assessing the contribution of the human agents to the disaster. Since then, many other papers have adopted this view about large-scale accidents according to which human errors must be analysed along with systems' failures. At about the same period, Rasmussen (1986) published what was recognised as the bible in human-machine interaction. Due to his engineering background, his view on human agents was much more focussed on their role of regulators and controllers. Rasmussen proposed that the operators' have various types of behaviours (skill-based, rule-based and knowledge-based) relying on different kinds of knowledge. This framework later inspired Reason's (1990) classification of errors, leading to a more unified vision of human agents in dynamic systems (see also Hollnagel, 1993 for a classification of human error models and concepts).

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 anonymous components in 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, participating to system safety. Such researchers as Hollnagel, Amalberti, Cacciabue or Woods have contributed to the enrichment this conception. We will not get into deeper details here as the case studies exposed later will give the opportunity to discuss these views and authors.

In this section, we have defended the idea that human cognitive capacities are extremely valuable regarding safety. However, there are factors that are inextricably linked to the validity of the goals pursued and actions taken. These are the knowledge and information that operators keep in memory when performing their task. The latter very strongly impacts on the quality of the dialogue between the operator and the environment. This issue will be addressed in the next section.

2.3 Mental models in action

When they interact with a system, humans need to understand what is going on and what has to be done. For this reason, they maintain a virtual mental representation of the various ongoing and expected processes in a system. This representation is called a *mental model*. But as we have seen in the previous section, humans' memory and processing capacities have limits. So in response to these, mental models are incomplete representations of reality that are fed by a) a portion of the total amount of knowledge on the system and updated by b) only a selection of the data available in the environment. These selections are driven by the objectives that the operators have on 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 very scarce, dynamic, goal-driven representations of reality (Ochanine, 1978) which incompletely 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 mode of control. This is extremely important in dynamic systems because the future states of the system and the consequences of one's actions are then anticipated, allowing operators to keep control of the system they interact with. Needless to say however, humans can perform errors in building or updating mental models, e.g. because of a) erroneous knowledge or b) flaws in data gathering. A main cause of erroneous knowledge at the workplace is the extreme complexity of modern automated systems which unavoidably causes operator's knowledge to be incomplete. We will see in section 3.1 what the consequences can be. There is much more to say about data gathering and far too little room in this paper. Let us just mention lack of or high expertise, strong time pressure and perceived similarity with a known situation as factors that degrade the accuracy of data gathering.

The previous three sub-sections have introduced some cognitive concepts. We have seen in section 2.1 that humans were classically regarded as bounded information processing agents. This facet, although it cannot be denied, does not sufficiently highlight the regulation role that humans have in systems, thanks to the flexibility of their reasoning processes. We have emphasised this position in section 2.2. Although it is not the only mechanism for it, the so-called flexible reasoning is supported by mental models that can be updated dynamically and reshaped depending on the goals maintained by the operator (see section 2.3). But so far, we have only reinforced the well-known idea that humans are a central component in system safety and that their activity is controlled by a scarce mental representation of reality. This is the standpoint from which we will investigate violations and claim that the operators' mental models have to be considered when reasoning about the safety of one's acts. Our purpose will be to defend that violations do not solely 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), plane 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), e.g. 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).

As we stated previously, violations may not be directly associated with accidents. The latter take more than violations to happen: they have to be combined with errors. Typically, a violation creates some specific unprotected conditions where recovering from an error no longer is 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. This position will be built upon two opposite case studies highlighting two opposite sides of the reality of violations. The first case will depict a nuclear 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 to what we call *desirable violations* with the case of the crash-landing of a DC-10 in 1989, in Sioux City, Iowa, USA.

3.1 Harmful violations: The Tokaimura nuclear fuel plant

On December 30, 1999, in Tokaimura (Japan)[3], a criticality accident[4] 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.

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 it was located only 10 cm above the floor, causing the uranium solution to be difficult to collect from the bottom of the column. 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 is equipped with a stir propeller making it easier to use for homogenising the uranium solution.

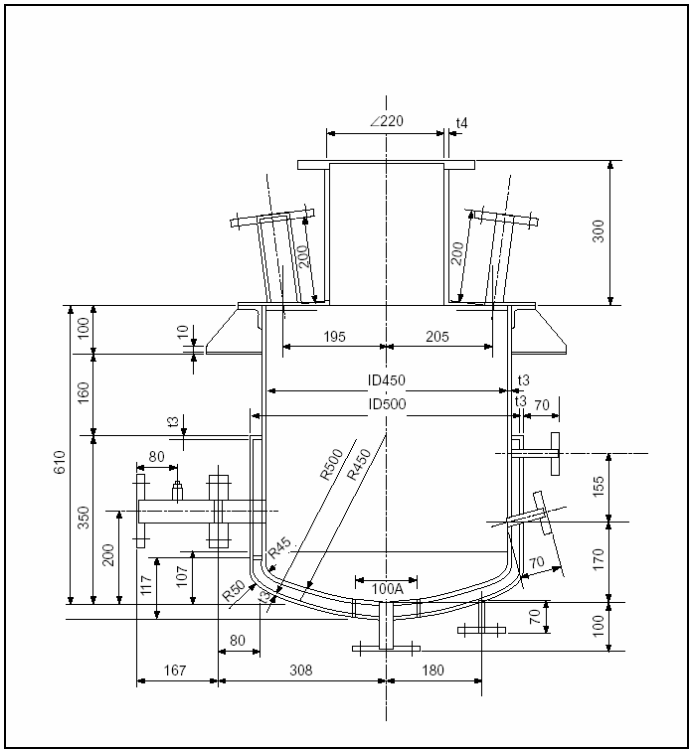

Figure 1: The precipitation tank at the JCO plant

The workers thought it was not unsafe to pour the seven batches in the precipitation tank. This false belief directly caused the accident but 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 deviations 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 criticality JCO accident was caused by a management-enabled violation being coupled with the operators' erroneous knowledge about the uranium 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 potentially 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 deviation, 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 deviations that trigger accidents rarely happen instantly. They often depart incrementally 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 deviations, the work settings can happen to be far beyond the prescribed practice. Large-scale accidents are made of a concatenation of small failures (Mancini, 1987).

With this JCO case, we want 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 in 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* control modes allowing 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[5]. 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.

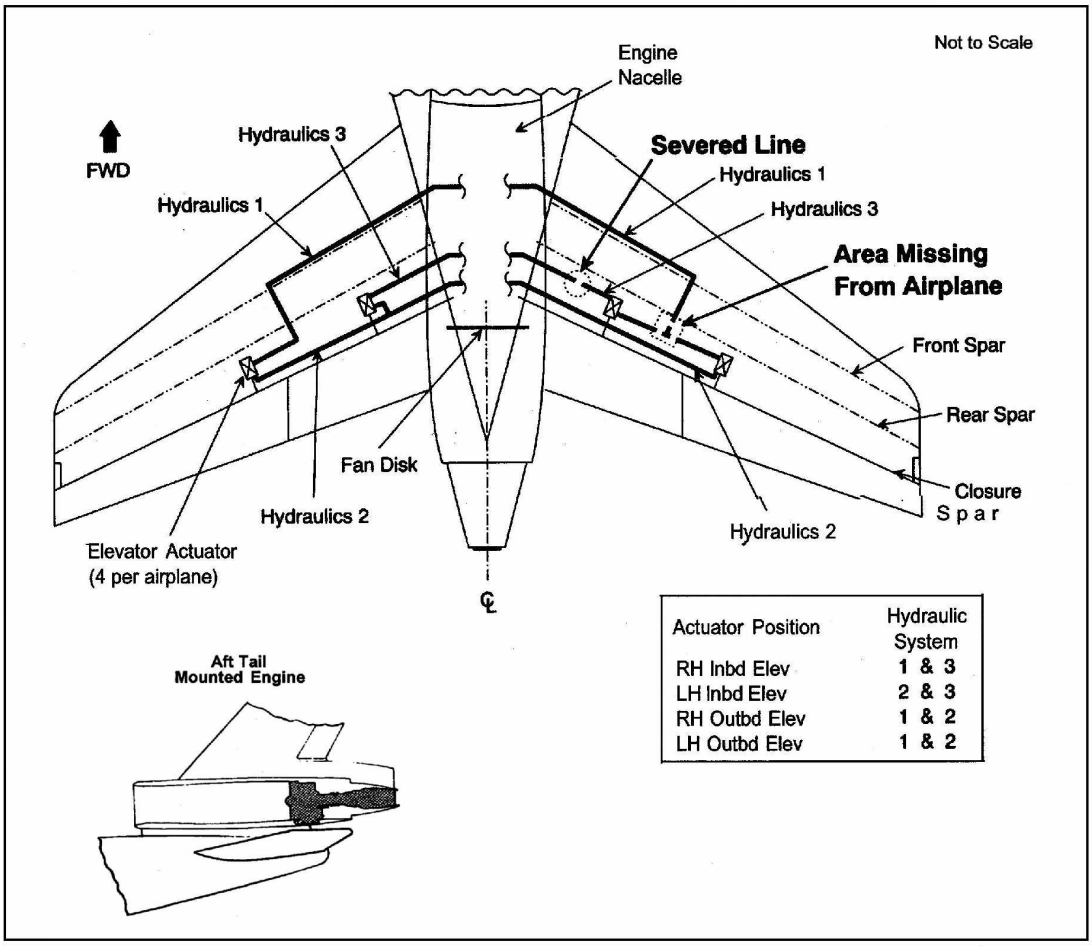

Figure 2: The damaged tail of the DC-10 (adapted from NTSB, 1990)

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[6]. 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.

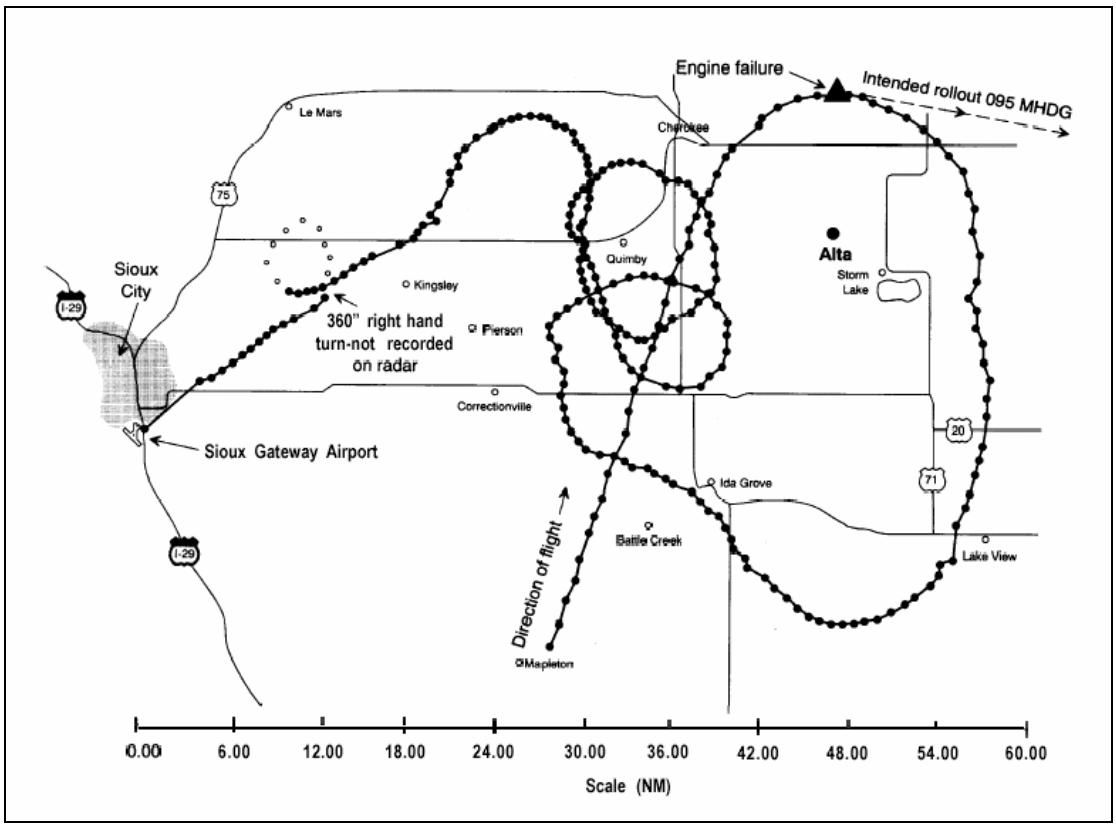

Figure 3: Radar plot diagram (NTSB, 1990).

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 than could have been achieved. 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 2.3, we have seen that mental models are built upon the knowledge operators have of a system and refined through the 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. Building and 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 understand very accurately the consequences of their actions although they are 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[7]. 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. Deviant acts were situated against the procedures but nevertheless exhibited a high degree of relevance.

4 IMPLICATIONS FOR THE SAFETY OF SOCIO-TECHNICAL SYSTEMS

As Van der Schaaf (1992, quoted by Rauterberg, 1995) puts it, when a system's unexpected configurations restore or enhance the reliability level, then these positive deviations 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. Although exceptions to the following statement exist[8], we think 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 in systems, one has to take into account the mental model that operators run. The two case studies exposed in this paper are two opposite instances of this argument.

We obviously accept the idea that many lives are not put in danger thanks to pilots and operators correctly applying well-designed procedures. But these procedures rely on probabilities and this introduces a bias: In high-pace, high-critical systems, marginal emergency conditions for which no procedures exist imply such a narrow span of legal

actions that violations are sometimes the only way to control the system. Following procedures under nominal or even expected emergency settings 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 the 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 if it implies implementing 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 Violations as reconfigurations

In our view, violations are actions that can be interpreted as *ad hoc* reconfigurations. In non-emergency situations, we conceive them as deviant acts 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).
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 operate safe *ad hoc* modes of control 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.
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.

4.2 Violations and 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 mention 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).

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. The intriguing fact is 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. They must simply 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 or 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. Some hints about what we think can improve this balance are given in the next section.

5 RECOMMENDATIONS

After Cacciabue and Kjaer-Hansen (1993), we think that a design team designing machines 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[9], the authors are rather sensitive to this kind of argument and will make the following recommendations rely on this principle.

Following one aspect of the JCO case, we would like to emphasise the fact that violations that cause accidents are sometimes design or procedures failures. So the errors that are coupled with these violations must not always be interpreted as incompetence (Rizzo, Ferrante & Bagnara, 1995). Instead, they sometimes highlight the need for improvements on a given system. Although training is an obvious response to incidents, we will clearly focus on design issues, promoting a view according to which tools should fully support human decision making and improve system's safety (Cacciabue, 1991; Hollnagel, 1987).

The JCO and DC-10 cases (see section 3.1 and 3.2 respectively) tell us that violations can be inevitable to operators due to e.g. faulty organisational settings or emergency. So as a design principle, we suggest that violations should be expected and supported in order not to leave operators in a risky non-assisted mode of control. The considerations below follow this assumption.

Interface designers should expect almost-impossible cases to happen. Even if some very rare events are not worth addressing by a specific design decision, interfaces should be at least highly configurable to each user's requirements (Hussey, 1999) and working methods (Bainbridge, 1998). We suggest that this configurability should be in accordance with the criticality of a system. For a safety-degraded critical system, safety constraints could thus be relaxed in order to allow variations in expertise, preferred control modes, etc. to perform emergency violations with maximal degrees of liberty. Such a design policy would offer to let the human be in full control of the system for adjusting the configuration to exceptional circumstances. The drawback is that the mental model of the operator has to be accurate whereas we know that, due to cognitive limitations, it is highly fallible. Moreover, as this option underexploits the potential of computer-based assistance systems, a complementary approach follows.

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 machine as smooth and efficient as the interaction between two persons. But it is an essential part of human communication that each participant is able to continuously 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 dynamically support the operators' reasoning, provide synthetic shots on the system's state, anticipate which action is now required, which information will be needed next, etc. So we think the next generation of aid tools should be able to provide a) assistance for unexpected emergency situations as well as b) anticipative protection measures for dangerous acts. Operators need more help on these exceptional situations for which they have not been trained rather than on nominal settings. Although this specific topic is out of the scope of the paper, one possibility could be to design systems that have a model of themselves, the latter being coupled with the aforementioned screening functions. The objective would be to let the system match the operator's attempted actions with the available system's functions.

6 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. The Tokaimura and Sioux City cases show that violations can generate very different situations depending whether they are coupled with a valid or invalid mental model. These two views are not mutually exclusive in system's lives. They cohabit at all times. As such, their double status i.e. contributing to or impairing system safety must be acknowledged by systems designers. Going further, the facts that some violations are possible warrants the presence of enough degrees of liberty for humans to exploit their innate cognitive flexibility.

7 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.

Bainbridge, L. (1998). *Difficulties and errors in complex dynamic tasks*. Discussion paper available at http://www.bainbrdg.demon.co.uk/Papers/CogDiffErr.html

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, Heildelberg (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

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

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.

Festinger, L. & Carlsmith, J. M (1959). Cognitive consequences of forced compliance. *Journal of Abnormal and Social Psychology*, 58, 203-210.

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.

Hussey, A. (1999). Patterns for safer human-computer interfaces. In M. Felici, K. Kanoun & A. Pasquini (Eds) *SAFECOMP'99*, Springer-Verlag, Heidelberg (pp. 103-112).

Kahneman, D & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80, 237-51.

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.

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).

8 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.

Footnotes

1. Although the concept of violation could lead to a wide discussion, we will use this simple definition for the purpose of this paper.
2. An immense amount of studies should be reported here. As the humble purpose of this section is only to give a flavour of a classical trend in psychology, exhaustiveness will clearly not be the objective.
3. Unless otherwise stated, the material in this section is from Furuta *et al.* (2000).
4. There is a limited amount of uranium that can be put together without initiating fission. When this amount is exceeded, a chain reaction occurs, generating potentially lethal radiations
5. 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).
6. A phugoid is an oscillation in the vertical flightpath of an aircraft whereby it repeatedly climbs and dives in association with fluctuations of speed.
7. Although it is out of scope for this paper, it is worth mentioning the relative weakness of rule-based behaviour in marginal situations. Rules can be applied even if the all the conditions required for them are not present (Besnard, 1999).
8. This could be the case of a system that is 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.
9. http://www.dirc.org.uk