IFE/HR/E – 2005/031

Human error and models of behaviour

Institute for Energy Technology

	KJELLER	HALDEN
Address	NO-2027 Kjeller	NO-1751 Halden
Telephone	+47 63 80 60 00	+47 69 21 22 00
Telefax	+47 63 81 63 56	+47 69 21 24 60

Report number	Date
IFE/HR/E-2005/031	2006-06-09

Report title and subtitle	Number of pages
Human error and models of behaviour	43

Project/Contract no. and name	ISSN
M-8108 Risit feilhandling	0807-5514
Client/Sponsor Organisation and reference	**ISBN**
Norges forskningsråd	82-7017-538-2

Abstract

Twenty years of specialized research on the issue of human error have indicated that the concept human error is far more complicated that originally assumed, to the point that some authors have recently proposed to reject the expression altogether. This report analytically investigates the concept of human error, its causes and manifestations, and the uses of error analysis. The report argues that human errors are not fixed events that can be studied by means of observation alone. Human error is instead a normative concept, which implies a process of comparing empirical events with abstract standards of correct performance. As both this process and the standards are dependent upon the theories of human performance adopted, a review of the dominating theories of human performance in safety-critical systems is provided. A set of suggestions on how to work with standards and models of behaviour is advanced, in order to improve the accountability and quality of error analyses.

Keywords: Human error, human behaviour, human error classification

	Name	Date	Signature
Author(s)	Salvatore Massaiu	2005-08-04	
Reviewed by	Magnhild Kaarstad	2005-09-01	
Approved by	Andreas Bye	2006-06-01	

Contents

1 Human Error

It is virtually impossible to review the issue of human error without finding articles and books that report on the percentage contribution of human errors to system failures. A review of incident surveys by Hollnagel (Hollnagel, 1993) shows that the estimated contribution of "human errors" to incidents ranges from about 20% to around 80%. The fact that the surveys covered a quite short period of time, from 1960 to 1990, makes it unlikely that such huge differences in the estimates can be explained by the transformations in the human-machine environment in those decades (see Hollnagel, 1993).

So what explains this variability? One could point to the heterogeneity of the errors counted which typically concentrate on errors in operation, but also include other phases of human-machine systems interaction, such as design, maintenance, management, etc. A second reason can be attributed to the different industries surveyed. In the review cited: nuclear power plants, aerospace, weapon systems, general studies, etc. These factors do certainly explain a good deal of this variability, but even if we were to concentrate on the same industry and on one particular class of actors (i.e. front line operators), we would still find very different estimates. The reason is that in field of human factors there is no general consensus on the meaning of the expression "human error". There are in contrast various models and perspectives of human performance that incorporate different interpretations of the concept human error. They bring about an extraordinary diversity of notions and applications that they associate with the label "human error". And they produce incident analysis methods and classification systems of errors that are typically only partially compatible with each other.

A further problem with the expression "human error" is that it has been traditionally associated with the attribution of responsibility and blame. In this context, "human error" is typically a judgement of human performance made after an event has occurred. Old views of human errors as dominant causes of accident have influenced the disciplines of accident investigation and error analysis up to our days, to the point that some authors have debased the label "human error" to a "ex post facto judgement made at hindsight" (Woods et al., 1994), with very little or no utility in advancing knowledge about system safety, or even rejected the label altogether: "there is no such thing as human error", (Hollnagel, 1993).

I believe there is still a use for the label "human error", provided we clearly define its meaning and delimit its applications in ways that counter the biases implicit in intuitive and traditional uses of the expression. Consequently, in the following sections I will thoroughly analyse the concept of human error, highlight the areas of misunderstanding, and provide a minimal definition capable of encompassing the majority of uses and applications. I will then review the different models or paradigms to human error analysis, and discuss some examples of classification systems that these approaches have originated. In the words of David Woods this analysis seems necessary since

> "one cannot get onto productive tracks about error, its relationship to technology change, prediction, modelling, and countermeasures, without directly addressing

the various perspectives, assumptions, and misconceptions of the different people interested in the topic of human error" (Woods, Johannesen, Cook, & Sarter, 1994), P. XVII).

1.1 The concept of human error

The concept of human error is not an easy one. There are several reasons for this. In the first place, even limiting the attention to the area of work psychology and human factors, there are different needs and interests in defining human error: human error can be defined, for instance, in order to identify potential threats to system safety, as it is done in human reliability analysis, or in order to identify the causes of an accident. In the former context the definition will probably concentrate on the types of actions an operator can perform within the system and their consequences, on the latter the focus will likely be on the causes of the human actions that were involved in the accident. A second difficulty arises as consequence of the different approaches to the issue: while an engineer will tend to analyse human performance in terms of success and failure in the same way as component elements, a sociologist will describe actions and errors in the context of the socio-technical influences and constraints in which humans operate. The most serious difficulty, however, lies in the concept itself. Human error applies to a large variety of actions (e.g. simple tasks, cognitive operations, motor-skills) it can be attributed to a host of different causes (e.g. internal constitution, external conditions, task demands, volitions) and it can be judged with different criteria (e.g. system parameters, agents intentions, social norms). Hence, it is not an easy task to include all possible conditions and fields of applications in a simple, yet general proposition.

Typically, human error is defined within the theoretical framework provided by a discipline, for a precise scope and to specific fields of application. Available definitions are then working definitions more or less adequate to a scope rather than correct of incorrect in abstract terms. I will, nonetheless, advance a general definition of human error, although not for the sake of a 'correct' definition but because the process will allow identifying and discussing the essential conditions of any definition, and clarifying the meaning of the concept. By providing a definition we will discuss some recurrent ones and therefore appreciate their relative strengths and limitations. This discussion will furthermore make it easier to appreciate the differences and similarities between the various approaches to human error that will be analysed later.

1.2 Errors as normative statements

In order to talk of human error some event or action associated with undesired outcomes or consequences needs to be present. This is the case both for mundane applications of the concept of human error, as a child that fails to report when doing additions, or for work contexts, as when a power plant operator opens the wrong valve. The important point is, however, that this plain consideration contains the two essential elements for a definition of human error: an event and a standard of correct performance. The standard of correctness defines whether the event (or action) is an error or a correct performance, whether the consequences (real or hypothetical) associated with the event are desirable or not. It is important to stress that an action is never an error or an unsafe act by itself but it is so only in comparison with a standard of correctness and a context of execution:

exactly the same action can be exemplary performance in one situation and gross mistake in another.

An error statement is thus a judgment, where a normative propriety (e.g. wrong, too much, too fast etc.), is assigned to a set of descriptive statements of actions and conditions of executions, in virtue of there existing a relevant standard or norm in which those actions and conditions are associated in a different way than the one observed (see Table 1).

Table 1. Type of statements involved in error analysis: two examples

Type of statement	Example 1: Industrial process	Example 2: Road transportation
Descriptive statements		
Action	Operator A opens valve x at t	Driver A pass junction X direction south-north at 11:32:12 pm
Conditions	Valve X is open at t_1 and $t < t_1$	Traffic light at junction X is red direction south-north between 11:31:30 and 11:32:30
Consequences	Release of polluting substances on atmosphere	Increased risk in junction
Standard of correctness		
	Time t to t_1 => valve X is closed	When traffic light red, driver stops
Error judgment		
	Operator A *wrongly* opens valve X at time t	Driver A *wrongly* pass junction X with red light
Causal statements		
Internal causes	Slip of action: operator A intended to open nearby valve Y but fail the execution	Perceptual confusion Circadian rhythms: sub-optimal performance on night time
External causes	Switches of valves X and Y close in position No feedback Unavailable procedures Low lighting in room	Low visibility due to shower Input complexity: left-turn light green
Responsibility judgment		
	Design of work place and working conditions are “error forcing”	Driver A fined

The reference to norms or standards of correct performance becomes of practical relevance when dealing with actions that are not straightforward definable as failures: violations, performance deviations, under-specified instructions, non-procedural practices, etc. That is to say, there are practical circumstances where standards of correct performance, procedures and norms are not clearly specified and a preliminary discussion around them is necessary for the individuation of something as a manifestation of error. In most cases, however, there are straightforward performance criteria and it is natural to agree if an execution has been too short, too late, on the wrong object, omitted etc. The evaluators will have no problem in referring to the time, space and energy proprieties of an action and to characterize it in normative terms as for instance, wrong direction, too fast, repetition, on wrong object and so on.

1.3 Manifestations and causes

Hollnagel (1998) has repeatedly stressed the importance of clearly distinguishing between causes and manifestations in error classification. He claims that few of the

existing error classifications make this distinction clear but mix up observable action characteristics with inferred causes. It is surely undisputable the superiority of a classification that makes the difference between error causes and manifestation explicit and which explains how causes and manifestations are related. However, many classification systems are practical tools developed in well-defined domains where the user would not see themselves making a great deal of inference where, for instance, indicating an “information communication incomplete”, or a “diagnostic error”.

Further, we should not believe that error manifestations or phenotypes are mere descriptions of actions and events. An error manifestation is properly a normative statement in which the time, space and energy dimensions of an action are evaluated against an agreed standard. Clearly, when the standard is obvious the difference has no practical implications. However, as we will see later, this is not always the case. The traditional behavioural categorization of errors in omissions and commission is a clear example of a normative process where the standard of correct performance are assumed to be clearly specified: without a well agreed performance criteria all commissions will also be omissions of something, as well as omissions could be in turn described by a varied phenomenology: an action could be missing, delayed, anticipated, or replaced by another.

Whatever the performance criteria used, the evaluation process depends on the assumption that the event and the consequences are not associated by chance, but there is, instead, a causal connection (1) between the event (action or inaction) and its consequences and/or (2) between the action and the surrounding conditions that preceded it. The latter point shows that a causal explanation of some real or potential unwanted consequences ought not stop to the error manifestations, but may refer to events internal to the subject as well as external characteristics of the situation. This is the level of the causes of the manifestations and it is dependent upon the theory of behaviour underlying the explanation. Therefore, in addition to a standard of correct performance the process of error attribution depends on the theory, or model, of human behaviour adopted. It is generally assumed that there is more than a single cause for any behaviour and that an explanation, or prediction, of a manifestation of error will include a set of causes that are deemed sufficient to have caused it, or to predict it (see Table).

1.4 Slips and mistakes

Definitions of human error are often provided from the point of view of the agent. These definitions are typically not limited to erroneous human actions or behaviours but consider mental processes as well as intentions. Mental processes (such as observation, memory, planning, etc.) are considered relevant because although they can fail without producing unwanted consequences on a particular action, they are likely to explain them on most occasions. In general, mental processes are seen as the mechanisms that underlie human actions and errors. This is a requirement of explanation, of understanding why humans do make errors, but also of description, as a limited set of causes can explain infinite erroneous actions.

Intentions, however, seem to be even more important in defining human error. A simple description of a series of events, although mental events, is insufficient to qualify them as erroneous. For instance, without reference to purposes and intentions the fact that a

person did not achieve a particular goal he/she was supposed to reach could be ascribed equally well to the person's choice of a plan of actions inadequate to reach his/her goal (i.e. the definition of mistake), or to his/her failure to execute an adequate plan (i.e. the definition of a slip), or to his/her purposeful selection of a goal contrary to rules and regulations (i.e. the definition of violation). As the example shows, there are multiple goals implied, both in the form of intentions of the agent and as intentions or standard of correctness of a group, an organisation or a system.

The mismatch between different goals and between goals and results provides the basis of a phenomenology of errors and unsafe acts. When the goals are those of a conscious actor the term intention is used: that a driver had the goal of turning left is synonymous with the driver having the intention to turn left. Several definitions of human error are, in fact, framed on the concept of intention and the difference between intentional and unintentional acts. However, not all approaches do so, and more seriously an over reliance on the common language meaning of intention can be misleading.

To illustrate the point, let's use two well-known definitions of slips and mistakes that rely on the concept of intention. Norman (1983) provides a very concise characterisation of slips and mistakes:

> "If the intention is not appropriate, this is a mistake. If the action is not what was intended, this is a slip".

This statement contains ambiguities in the use of the term intention. Since it aims at defining human error in the context of real work tasks (which are typically characterised by multiple goals, interdependences between goals, time constraints, sub-goals, preconditions, execution conditions, etc.), Norman's meaning of intention is 'plan' (a rule, both actions and goals) in the case of mistake and 'expected outcome of the plan' (actions implied by the plan) in the case of slip. Let's see why. The intention of Norman's definition of slip is clearly the 'goal' of the actor, the expected outcome of his/her plan: the operator intended to push a button but accidentally pushed another one. In the case of mistake the intention cannot strictly be the goal of the actor otherwise it will be the common definition of a violation (including acts of sabotage, suicide attempts etc.). It is instead the plan (goals and actions) that is inappropriate to achieve the intention (overall goal), the plan that is inconsistent. A plan P is inconsistent when:

(1) It does not imply a specific execution E to be put in place: P do not imply E

or

(2) When the execution E that it correctly implies is not adequate to achieve the overall goal OG of the plan: P imply E but E do not imply OG.

Generally the selection of a wrong goal, due, for example, lack of skills, is not considered a violation. This is exactly the problem with the meaning of intention: the wrong goal selected here is properly a sub-goal (SG), that is to say, a means to achieve an overall goal (e.g. secure the system). In terms of our definition of inconsistent plan, in this case of complex tasks a plan is inconsistent when:

(1b) It does not imply a specific execution E put in place: (P do not imply E) or (E cannot achieve SG)

or

(2b) When the execution E that it correctly implies is not adequate to achieve the overall goal OG of the plan: (P imply E) and (E imply SG), but (SG do not imply OG).

So again, "choice of a wrong goal" is a mistake not as inappropriate intention, but as inadequateness of the means (the sub-goal) to achieve the top goal (as in (2b) of the above definition). Clearly, Norman did not mean "inappropriate top-goal" in the definition of mistake, but rather inappropriate plan in the sense of inappropriate means for the overall goal. The point is, however, that the different meanings of intention and the different levels of analysis of the task in the definition are not made explicit. As a consequence, Norman statement switches between intentions as expected outcomes (in the case of slips) and intentions as plans (mistakes) (or between simple and complex tasks).

Another example of the difficulties of working with intentions is provided by Reason's (1990) working definition of human error:

> "Error will be taken as a generic term to encompass all those occasions in which a planned sequence of mental or physical activities fails to achieve its intended outcome, and when these failures cannot be attributed to the intervention of same chance agency".

This definition is ambiguous because it does not specify if the outcome is the object of the intention of the actor or of someone else. When mistakes are characterized as "failures of intended actions to achieve their desired consequences" we wouldn't like the "desired consequences" to be exclusively those of the agent. Otherwise we should have to rely (typically) on the actor's subjective experience of having made an error in order to characterise the action as a mistake. To recognise their own mistakes as well as to correct their plans before undesired consequences are reached is a very useful skill of human beings, but it's not a general condition on which to base a definition of human error. When Reason proposes a working definition for mistakes this ambiguity is not resolved:

> "Mistakes may be defined as deficiencies or failures in the judgmental and/or inferential processes involved in the selection of an objective or in the specification of the means to achieve it, irrespective of whether or not the actions directed by this decision-scheme run according to plan.

Leaving aside the issue or referring to the judgemental and inferential processes as the inferred cause the mistakes, mistake is defined as a failure in (1) the selection of an objective or in (2) the specification of the means to achieve it. The first statement of the disjunction is similar to Norman's definition of mistake with a potential ambiguity between objective as overall goal and objective as sub-goal and hence between mistakes and violations. When objective is read as sub-goal a definition is on the whole coherent

with our characterization of inconsistent plan for complex tasks. In other words, if we assume objective to mean sub-goal, that is to say, we are not concerned with violations (as Reason was not in the chapter he put forward the definition) the ambiguity is removed. Still, the intention here seems to be exclusively the one of the agent and we have seen that in many cases failures and unsafe acts are defined in relation to other criteria.

1.5 Violations

When we take the issue of violations seriously into account the story complicates even further. So far we have assumed a simple sense of violation as a deliberate choice of a goal contrary to rules and regulations, as in acts of sabotage and vandalism. A violation is then very easily recognisable as we assume that the individual is capable of choosing between well-understood and unambiguous systems' goals. In reality, this is not straightforward. The individual can disregard but also misunderstand the prescribed task for a variety of reasons: because of a lack of knowledge, because the goals are poorly defined, because the system contains conflicting goals and principles, because the conditions of executions do not make possible to perform the task in all situations, etc. Following a scheme proposed by Leplat (1993), in all these cases there is a divergence between the prescribed task, or "task for the expert", and the re-defined task, or "task for the subject". When the subject knows the prescribed task, but for some reasons does not want to execute it we would normally call it a violation. This is true from the point of view the agent's intentions, and is the common interpretation of a violation. In this view some degree of intentionality or deliberation must be present to qualify a divergence between prescribed task and redefined task as a violation.

However, the term violation can also describe cases of deliberate choices of goals contrary to rules and regulations, but in which the agent's intention was not in contrast with overall systems goals (e.g. safety), or the violating behaviour did not lead to negative outcomes. This is a consequence of the fact that the "task for the expert" is an ideal and by definition correct prescription, while the actual prescriptions embedded into work procedures, rules and orders might sometimes be inadequate or neutral towards the realization of overall systems goals.

On the other hand, if we shift to the point of view of the "expert", or, in general, an external point of view, we would probably call violations also those cases where the subject did not know, or did not deliberately choose not to follow, the prescribed task (actual or ideal). One can find in literature examples of rule violations that are attributed to lack of training or understanding, which clearly point to the fact that the subject did not know he/she was not following the prescribed task. Also, it is common in the violation literature to talk of routine violations, behaviour contrary to rule and regulations that has become the norm, that is, executed automatically and without deliberation. Once an external point of view is taken, the realm of application of the concept violation might extend to include all behaviours that diverge from procedures, rules, instructions, 'missions', as well as from the principles to be considered in the evaluation of a task, together with their conditions of execution (possibly everything not due to impairment, as in legal terms). In other terms, the conceptual distinction between errors and violation is far from being clear-cut. What is certain is that in defining and evaluating unsafe human actions we must be aware of the differences in relying on

internal versus external points of views as well as on the consequences of assuming actual versus ideal standards of correct executions. Table 2 illustrates how a "phenomenology" of violations can be obtained by considering these two dimensions.

Table 2. Phenomenology of violations

Standard of correct execution	**Point of view**	
	Internal: looking at the deliberate choice of goal	**External: looking at actual behaviour**
Ideal: objectives and principles	Malevolent and irresponsible intention, incorporating certain or likely negative outcome: – Sabotage, vandalism, etc.	Misunderstanding or ignorance of system's objectives, principles and conditions of application: – Mistakes as violations – Reason's "erroneous violations"
Actual: existing rules and procedures	Goal conflicts, i.e. choice of *system's or personal* objectives and principles which conflicts with *known* existing rules and procedures, but also conflicts between rules; both positive and negative outcomes: – Non formalised best practices and recoveries – Non harmful short-cuts, strategies, etc. – Situational violations, case adaptation of inapplicable or conflicting rules – Exceptional and optimising violations	Ignorance of existing rules and procedures, but also non-deliberate behaviour contrary to rules: – Behaviour dictated by *system's* objectives and principles which conflicts with *ignored* existing rules and procedures; generally associated with positive outcomes – Routine violations

1.6 Intentions

It is now clear that the problems and ambiguities discussed in relation to Norman's and Reason's definitions, and the various interpretations of the concept violation, rotate around the meaning of intention and intentional behaviour, and between the difference between the intentions of the expert, i.e. system designer, management or society – and the intention of the subject. It should also be stressed, however, that Reason's and Norman's definitions are working definitions, and as such their appropriateness is their utility. The problems discussed stem from the multiple meanings of the concept intention and the fact that it is present, as standard of correct performance, in the definition of all types of unsafe acts, either they are called errors or violations.

The term intention can have three different meanings that can be outlined by recalling the history of the philosophical use of the concept. In Latin intentio had originally the same meaning as concept but was used by medieval philosophers, first of all by Thomas Aquinas (1225-1274), to indicate both the reference of the concept (on objective entity) and the act to refer. The concept was reintroduced in the nineteenth century by Austrian philosopher and psychologist Franz Brentano (1838-1917) to define all psychological phenomena, as opposed to the physical ones. For Brentano all psychological events are intentional in the sense that they are directed to some object, they relate to some content. In addition, all psychic acts, insofar they are intentional, are completely present to the consciousness, they can be entirely known. These aspects of the concept of

intention are still present in the common use of the word, as we have seen in the definitions above. To summarise, the concept intention is thought to have the following properties:

(1) It is the expected outcome of an activity, the goal (parallel to the referred object)

(2) It is the outcome and the actions to achieve it, the plan (parallel to Aquinas' intentio)

(3) It is a mental phenomenon present to the actor's consciousness, e.g. the violation from the subject point of view (as in Brentano's psychic act).

When the three aspects of the concept intention are clearly recognized it becomes easier to understand the concept of human error and unsafe act as well as to interpret the definitions that make use of it. And it would probably be less misleading to think of mistakes in terms of the second meaning of intention above, i.e. as inappropriate plans or inconsistent plans of actions. Keeping in mind the previous discussion, we could define three classes of unsafe acts from the actors' point of view in their most basic form: (a) slips as wrong executions; (b) mistakes as wrong plans of actions; and (c) violations as wrong intentions (as top goals).

It is also true that some characterizations of human error do not refer to agents' intentions at all, as for instance the engineering tradition. It is however necessary to refer to intentions, volitions or reasons in order to provide psychologically tenable definitions of human errors and violations. It is also natural to refer to intentions when error is defined from the point of view of the agent. It must however be pointed out again that when analysing actions with unwanted consequences, the intentions are not always those of the agents but the standard of correctness for the actions can be external (e.g. procedures, expectations of the organisations, etc.) and may or may not coincide with those of the agents.

1.7 A general definition of human error

The previous discussion has outlined the fundamental dimensions necessary in defining human error: the goal or intention, as the standard of correctness, and the action to be evaluated. We have concluded on the importance to restrict the meaning of intention in order to differentiate between plans and goals. Connected to this is the level of application of the definitions: primitive tasks versus complex ones provides yet another way to confound between intentions as plan, intentions as overall goal and intentions as sub-goals. Keeping in mind these distinctions, I define human error in the following way:

> Human error is the failure to reach an intended goal, the divergence of a fact from a standard.

This definition is able to include all Reason's types of unsafe act (slips, lapse, mistakes and violations), by way of selecting the appropriate goals and intentions. The standard of correctness can be internal to the person (the person's intention or expected consequences of his/her action), or external (the expectations that other persons or

organizations place on the agent). When reference is made to internal standards of correctness it is not required that we have to rely on the person's own experience of having made an error. This experience can be valuable or not depending on the circumstances, but it is not necessary. There are, in fact, external or public criteria that, trough inference, allow for the ascription of intentions to the agents, in the way it is typically done in cognitive psychology. By reference to goal structures, volitions and intentions it has been possible, for example, to distinguish between mistakes and slips, that is, between actions that followed an inadequate plan and actions that followed an adequate one but failed to reach their goal.

It may be questioned whether the goal has to be present to the agent's consciousness, that is, as an explicit goal, or it can be sub- or unconscious, that is, as implicit goal, as in the case of lower level cognitive tasks such as motor skills. The point is clearly related to the difference between errors and violations, that is, to the degree of deliberation of the action being evaluated. The answer is that since the process of error attribution is a normative one that normally is not performed by the subjects who committed the actions at issue, the difference is not important, as external criteria or internal attributions are employed as standards of correctness. It becomes important in terms of error psychology where the internal mechanisms of error are the issue of study (for this respect see Reason, 1990, pp. 5-8).

Finally, it should be noted that this definition of human error would correspond to a definition of error in general were not for the nature of the goals. It is the cognitive and intentional nature of the goals that make these errors "human".

1.8 Practical aims of error analysis

All theories and techniques that investigate the issue of human error necessarily refer to some combination of the following three causal factors: 1) person related/psychological; 2) environmental-external; and 3) task characteristics. Differences in characterisation, importance and interactions assigned to these three elements result in different theories, models or approaches to human error, as we will see in Section 2. The relative importance of the causal factors present in an explanation of human error is moreover dependent on the main research question. It is not difficult to indicate the three most common issues in retrospective and perspective error analysis:

1. The event is the cause of the unwanted consequences

2. The actions are caused by some internal and external factors

3. The actor is the responsible of the unwanted consequences.

The three issues are traditionally associated with different disciplines. The first is exemplified by the engineering approach. The traditional engineering approach (before the Second World War) identified incident causes into 'unsafe acts' and 'unsafe conditions', that is, attributed the cause of system failure to human or equipment. Accident prevention manuals of that time attributed 80% of incidents to humans and 20% to equipment (Heinrich, 1931). The human and technical causes were seen as

independent of each other and the prevention strategy to be the modification of either one.

More recent engineering approaches known as human reliability assessment still start from the distinction between human and technical failures but have enriched the analysis. After the human or technical source of system failure is identified the analysis can go further in identifying the components' sub-systems or operators' functions that failed. The decomposition will stop at the sub-components or human functions that failed over which reliable failure probabilities data are available. For instance, in the case of an operator that failed to start the auxiliary feed-water system a fault tree diagram will be produced where the operator failure is represented in terms of combinations of elementary task functions necessary to accomplish the task e.g. read an analog meter, diagnosing an abnormal event within 10 minutes. What still is common to the old approach is that the human failures are defined in terms of unfulfilled operator functions, or unperformed tasks assigned, and not from the point of view of the subjects.

The second issue is the core of the discipline of error psychology. The interest here is on the psychological causes of the action that failed (independently of it having negative consequences on a particular occasion). Clearly, this perspective complement the previous one by providing the theoretical basis for a fault tree specification, and, at least in principle, failure data. Error psychology investigates the psychological mechanisms that control cognitive activities and identifies internal mechanisms, psychological functions or global performance control modes together with tasks conditions as causes of failures.

The perspective represented by the third statement is typically a juridical or moral one. It aims at establishing the degree of involvement and the margins of choice of the agent in the causal process that led to the unwanted consequences. The themes of intentions, comprehension and autonomy are central in answering this research question. This issue is related to accident investigation, although most techniques limits their scope to the multiple causes of an accident and leave the issue of personal responsibility to the prosecutors. As practical enterprises, accident investigation techniques use methods and models from the two previous approaches.

It should be stressed that the issue of responsibility has a bearing on the topic of human error at work well beyond an accident investigation perspective, since the degree of responsibility associated with a task influences the behaviour of the agent in purely cognitive and behavioural terms (Skitka, Mosier, & Burdick, 1999; Skitka, Mosier, & Burdick, 2000). The issue of responsibility thus can itself be a causal factor of accidents and should be considered in the design process (e.g. function allocation, support systems, error tolerance)

It is clear from this discussion that different approaches and research questions focus on different aspects of the causal explanation of human errors although they all necessarily include, at least implicitly, the reference to the three levels mentioned before: psychological, environmental, and task. Yet, the reference to the core factor permits to differentiate the different approaches of human error modelling and classification.

1.9 Errors, accidents and safety

The classical paradigm of safety science maintains that in order to achieve safety hardware failures and human errors must be reduced or eliminated. The study of accidents and incidents is one natural place learn about errors, since the analysis of past events makes it possible to identify systems failures, discover their causes, and in this way to generate general knowledge. Not surprisingly studies in this direction started already at the beginning of the 19th century and were directed by the assumptions that (1) there were two paths towards incidents, that is, technology failures and human errors; and (2) that the two were quite independent from each other. These two assumptions have been the hallmark of safety science up until the 1980s, and their influence is still strong (as one can easily see by the media treatment of technological accidents, which typically ask whether the cause of the accident was a technical failure or a "human error").

As technological progress in the 20th century advanced faster than human factors science, this traditional view on safety, which maintained two independent causes of accidents, ended up placing considerable emphasis upon the negative influence of the human element, and in particular of "front line" operators of the systems: pilots, air traffic managers, ships' officers, control room crews, anaesthetists and so on. The major system safety challenge soon became the reduction of the potential for human errors as the dominating cause of accidents. A first solution was envisaged in designing the human out of the systems by mechanisation and automation. When this was not possible, and hence the human element had to be left a place, the inclination was to apply to the human the same theories and methods as to the hardware elements of the system. An example of this propensity is the Fitts' list, which compares humans and automatic machines against the type of task they can perform, as a means to allocate functions in a system. As we will see later (see Section 2.1) such early approaches did not contribute much to the reduction of accidents nor to the understanding of the human role for the system safety. They lent ideological support to the so-called 80:20 rule, an unproven assumption that stated that 80% of the accidents were human caused and 20% equipment caused, to the extent that this became common wisdom in accident prevention manuals of the time.

The reason why these early approaches did not advance knowledge on risk and safety was that they had serious methodological flaws. Incident analysis, framed into the human-machines dichotomy, did not allow finding general patterns out of the particular incidents. As incidents are typically the results of unique mixtures of factors, the reliance on a simplified causal model made it impossible to identify the real determinants of accidents, to the point that even the distinction between causes and effects became arbitrary. In fact, these early attempts lumped together incidents independently of their characteristics and especially independently of the human contribution to the events. The role of the individual in the accidents was not really modelled, but for the psychophysical characteristics of the victims.

1.10 Modern safety science

Safety research thus tried to understand why incidents occurred as well as to envisage remedies for the accident prevention. However, the study of human error as a specific

topic only came to the forefront of industrial research late in the twentieth century and as a consequence of large-scale accidents such as the Tenerife aircrafts collision, Three Mile Island, Chernobyl, and the Space Shuttle, to only mention a few.

The old dichotomy between technology failures and human errors was replaced by system thinking. The modern approach considers safety as the result of the interplay between individuals, technology and organisations, a perspective that in Scandinavia is typically referred as Man-Technology-Organization (MTO) model. The new safety science recognised the inadequacy of treating the human with the same tools and methods used for the hardware elements, and special emphasis was put into the disciplines of human factors, applied psychology and organisational research. The leading findings of the about 20 years of cross-disciplinary research on the role human error for system safety have modified the intuitive assumptions normally associated to the relation between errors and accidents. They can be summarised as follows:

1. Human errors have to be viewed in a system perspective in order to assess their contribution to safety. Individual errors can and do occur without resulting into accidents: most human-machine systems incorporate barrier functions or safety nets that bring the system back to safe operating conditions in case of deviations caused by initial failures. Amalberti (2001) provides a quantitative estimate of one human error out of 1000 that have unacceptable severe consequences. It is now accepted knowledge that accidents in ultra-safe production and transport systems (i.e., systems with less than one accident per 100 000 events) are usually the result of unforeseen combinations of errors happening at different level of the man-technology-organisation complex. The ideas of defence-in-depth (Reasons' Swiss cheese model) and high reliability organisations (Rochlin, 1993) were developed in this context.

2. Human errors could not and should not be eliminated completely. As it became clear since the firsts international conferences on the issue (Senders & Moray, 1991), human error could not be treated in exactly the same terms as technical failures. It was noticed in the first place that errors are an essential component of learning, and that they even seem to display positive roles, e.g. creativity, exploration, adaptation. Even more importantly, although humans often produce errors that result in accidents, they more often perform correctly and, in particular, are capable of detecting and recovering both system's and their own errors. Detection and recovery of error might even be considered as better indicators of performance than error production.

3. Individuals recover the majority of their own errors before they result into incidents. Error control is part of the broader performance control, the cognitive regulation of performance where operators dynamically optimise performance objectives and costs. Cognitive control includes activities as: awareness of performance goals and difficulty at the required level; style of control used (conscious or automatic); choice of mechanism to detect and recover errors; and tolerance to produced errors. The ideas of cognitive control and recovery potential resulted in two classes of approaches. The first class is known as error management, error handling or simply error recovery. System safety is pursued by without concentrating on errors per se but on the generation and propagation of system hazards and on the way these can be prevented to results into accidents. These approaches, that in the literature go under the names of error management (Bove, 2002), treat management (Helmreich, Klinect, & Wilhelm, 1999),

and control of danger (Hale & Glendon, 1987), provide models and classification of human error different from those that concentrate on the human errors production mechanisms. The second class studies the cognitive control of global performance and individuates for example cognitive control modes (Hollnagel, 1993), or the meta-knowledge and confidence that ground cognitive risk control (Amalberti, 1992).

1.11 Classification

A classification of error is a structured way of reducing the multiplicity of error manifestations to a smaller set of fundamental manifestations or to a set of causal mechanisms. In principle, error classifications or taxonomies are not different from those found in the natural sciences. In practice, error taxonomies lack the internal systematic order of the natural taxonomies which are organised around few and simple principles. The problem is that in the field of human error there are not either agreed definitions of what constitutes the manifestations that are to be organised, nor simple causal relationships between causes and between causes and manifestations.

The causal explanation of behaviour (and thereby error) is the base of a classification system. Without a causal model a classification scheme is arbitrary since it is the underlying model that determines how the scheme is organised, what is cause and what manifestation, how the terms are to be interpreted and applied, and what combinations are meaningful. As different causal models can describe the complexity of human behaviour, so there are differences in the description of human errors between and among taxonomies. In general, there are two level of description of human error. The basic level of description is the overt behaviour or manifestations of errors as discussed above (for example, omission and commission, wrong timing, too much force). Classifications that include characteristics of the individual, of the internal psychological mechanisms and of the external environment refer to the causes of behaviour and not only to manifestations. Such causes can be observable, e.g. feature of the situation such as glare, noise, equipment, availability of procedures, years of service, etc. or theoretical constructs hypothesized to explain cognitive processes, e.g. decision, diagnosing, capacity limitations, observation etc. Errors as causes can be divided in terms of such internal functions, e.g. errors of detection, decisions errors, or can be related to the features of the situation, e.g. stress related error, poor illumination, glare etc.

Beside the causal model adopted, error classification can be organized around the principle of risk management or control of danger mentioned before. In this case the classification and modelling will not be limited to errors causes and manifestations but will include the wider process of successful and unsuccessful performance. This process is centred on the way errors and hazards are handled more than on the way errors came about. It should be noticed that error producing and error management approaches are not theoretically contrasting views but rather the difference is in the task performance level used as unit of analysis. The point can be illustrated by contrasting risk management in air traffic control with human reliability analysis in the nuclear sector. The latter has the main goal of quantifying the reliability of man-machine system, typically a nuclear power plant. System experts write down a PRA/PSA (Probability Risk/Safety Assessment) event tree model, a logical representation of how a set of disturbances (initiating event) can develop into a serious negative outcome (e.g. core

damage). Operators' activities are usually represented as recovery behaviours that need to be assigned a failure/success probability in the same way as all other failures represented in the event tree model. Similar logical models, called fault trees, are used to calculate failure probabilities. In the case of human failures the required recovery behaviours are typically decomposed into the logical combinations of operations and cognitive activities necessary for their success. Human error probabilities for the un-decomposed events are obtained from published sources or estimated by the experts and are adjusted for the effects of contextual factors present during the performance (performance shaping factors). The example shows that HRA models only the human error production phase while error management is incorporated in the system analysis, the PRA, which properly provides the starting point of the HRA. The system experts thus perform the task of modelling system and risk scenario dynamics in the PRA before the HRA is performed. This rigidity in the modelling of a dynamic system has been repeatedly criticised (Hollnagel, 1998) but is dependent in part on the, at least assumed, predictability of the process of nuclear power production and in part on the quantification requirements.

In human machine systems where risk scenarios and dynamics are less predictable and the focus is not on risk quantification but rather on identification and reduction, the phases of error production and error management are typically analysed as parts of a single process. This is generally the case in aviation, air traffic control, and road traffic, and is the hallmark of incident investigation. Here it is customary to analyse large performance segments or series of events where (possibly) different actors perform many activities, and where errors are committed, recovered or exacerbated in the risk management process. We will return later to risk management models and classifications, suffice it here to say that, besides the focus on whole performance success or failure, these approaches emphasise the positive side of human performance and the active and anticipating role of the operators.

This discussion also shows that there is a strong relationship between the theoretical approach, the practical purpose and the domain of application, which determines the level of description and the shape of the classification systems. If we concentrate on the purposes for classifying human error we can specify four main classes:

1. Incident investigation. To identify and classify what types of error have occurred when investigating specific incidents (by interviewing people, analysing logs and voice recordings, etc.).

2. Incident analysis. To classify what types of error have occurred on the basis of incident reports; this will typically involve the collection of human error data to detect trends over time and differences in recorded error types between different systems and areas.

3. Error identification. To identify errors that may affect present and future systems. This is termed Human Error Identification (HEI).

4. Error quantification. To use existing data and identified human errors for predictive quantification, i.e. determining how likely certain errors will be. Human error quantification can be used for safety/risk assessment purposes.

Incident investigation and incident analysis are retrospective activities where the classification system will help explaining an event that has already happened. Most classification schemes are developed for retrospective analysis. Error identification and error quantification are predictive analysis, where the interest is on events that can happen. Predictive analysis has been the concern of system designers and reliability practitioners. Although the explanation of past events and the prediction of future ones are the basic features of any scientific theorizing, the exchange of methods and of classification schemes between the two directions has been rather limited. This is due to lack of comprehensive theories of human behaviour and the consequent need to delimit the scope of the analysis to the prevailing interest. Another point of difference between prediction and retrospection is that while reliability studies have centred classification at the observable level of behaviour (omission and commission), incident investigators and system designer have favoured descriptions at a deeper causal level.

2 Models of behaviour and human error

In the process of error attribution, or equivalently, of evaluation of normative statements, it is essential to specify the standards of correctness adopted as well as the model of human performance that controls the application of the standards to the conditions of execution under investigation. In the words of Woods & Cook (2003): “the standard chosen is a kind of model of what it means to practice before outcome is known. A scientific analysis of human performance makes those models explicit and debatable”. It is in this spirit that this section will describe the main models of human performance that have been used in the study of human error.

Behavioural theories have always used models and metaphors to explain the complexity of human mind and behaviour. A number of these have been borrowed from the prevailing scientific and technical paradigms: mechanics and steam power in the nineteen century, animal learning and telephony in the early twentieth century, computers after the second world war, and, more recently, cybernetics and artificial intelligence. The dominant psychological schools of the early twentieth century were psychoanalysis and animal learning. The former exerted its influence in therapy and the media, while the latter dominated academic psychology, particularly in the United States. Here the most influential psychologist became John B. Watson with his Behaviourist Manifesto of 1913 (Watson, 1913) where he banned the mentalist tradition, that is, the discourse over mental concepts such as intention, volition, and particularly consciousness and introspection. Parallel to the animal learning and behaviourist psychology was the controversy over heredity and environment, nature versus nurture that framed the investigations into industrial accidents. Those who believed in the centrality of heredity developed theories that explained behaviour in terms of observable individual characteristics. In criminology theories were developed that classified criminal types by physiognomy. Similarly, in industrial accident investigations individual characteristics, such as sensory capacity, speed reaction and personality, were looked upon as determinants of the likelihood of a person being involved in accidents. This early approach to describe human behaviour at work went under the name of accident proneness theory, and is the first model of behaviour we will describe in this section.

2.1 Accident proneness model

The accident proneness model was developed in Great Britain at the end of the 19th century and the beginning of the 20th to explain the increased accident rate in industrial production. The theoretical context was the heredity versus environment controversy, which in turn was rooted in Darwin's evolutionary theory. Two explanations were advanced to explain the increasing rate of industrial accident: the first one stressed the importance of the environment, that is, the growing speed of production and the more demanding work tasks; the second reputed the individual differences to be more important and was historically concerned with the consequences of drafting regular workers for the first world war and the employment of (assumedly) less competent young and women. The dispute was never resolved and was probably irresolvable in the way it was posed. The premise of debate was, in fact, that the two explanations were independent from each other, so that the individual characteristics would make some persons more dangerous independently of the technical environment. As a matter of fact, the environmental perspective succeeded in guiding health and safety regulation, as documented by accident prevention manuals of the time. The accident proneness model on the other hand guided accident investigations and research, becoming popular among insurance companies.

The accident proneness model claimed that individual differences made some persons more likely to incur in accidents. Consequently, it researched individual differences in sensory (e.g. visual capacity), psycho-physical (e.g. speed of reaction), and psychological (e.g. personality) characters. The results of the research were generally poor and no psychological classification of accidents was produced. Hale and Glendon (1987), summarise the shortcomings of the accident proneness research:

(1) The proneness could be 'proved' only after the incident, hence statistics emphasised the characteristics of the individuals rather than those of the accident.

(2) Accidents were lumped together for statistical analysis independently of their characteristics and on the real involvement of the victim in the accident causation.

(3) The preventive actions proposed by the model were (a) excluding some individuals from performing dangerous work or (b) modifying mutable traits by training, counselling and motivation.

(4) Different groups of individuals, however defined, could and were found to have higher accident rate but no psychological characteristic was able to explain more than 20% of the variance in accident rate.

(5) The theory offered the opportunity for blaming the victims for the accidents lifting employers from responsibility.

The failure of the accident proneness model to find a valid set of explanatory individual factors, which could be used for accident prevention, discredited not only the model but also any psychological attempt to provide a practical basis for system safety

improvement. Designer and engineers, lacking the basis for differentiating between the normal and the accident prone, assumed the worst-case scenario, that is, assumed all humans to be unreliable, and sought system safety by reducing the human role and increasing automation. This conviction was furthermore reinforced by the indirect support provided by the accident proneness model to the 80:20 rule, which stated 80% of the accident to be human caused and 20% equipment caused: the reduction of the 20% of technical causes became a measurable objective for safety research.

2.2 Traditional Human Factors and Engineering models

Engineering approaches to system safety have maintained the dichotomy of human versus technical failures. There are however two schools of thought regarding how to treat human failure: the first considers human failures in all stages of system life cycle – specification, design, manufacturing, installation, maintenance, modification and, not least, operation, as systematic error, that is, error with identifiable and modifiable causes which is in essence a non-quantitative phenomenon. The alternative approach considers human failures, and in particular human errors during operation, to be random, at least at the elementary task level, and hence to be quantifiable. This approach is the essence of techniques of human reliability analysis (HRA) that are part of systems' probability risk assessment (PRA), which we will discuss later.

2.2.1 Traditional Human factors

Human factors, or human factors engineering, can be defined as applied research on the physical and mental characteristics, capabilities, limitations, and propensities of people at workplaces and the use of this information to design and evaluate the work environment in order to increase efficiency, comfort, and safety (Kelly, 1999). Human factors became firmly established as a separate discipline during the Second World War as a consequence of the proliferation of highly complex systems (most particularly aviation systems) that stretched human capacities to their limits. Human factors practices and standards have since become a major consideration in many design areas, particularly those in which the human/system interface is critical to overall system safety. Human factors research and recommendations address such issues as automation and control, military system design, nuclear power plant regulation and evaluation, as well as consumer usability issues such as the layout of automobile dashboards. Human factors research maintains that most active monitoring and intervention by operators in complex systems involves cognitive (mental) functioning. Typical study issues are fatigue, memory, attention, situation awareness, workload, cooperation, training, manpower, crew management and decision-making.

Insofar the discipline of human factors is concerned with the problem of system design and production of standards and regulations, the focus is on the global and qualitative aspects of human performance. Human error is treated indirectly assuming that improved system design solutions will aid human activities and hence reduce the occurrences of errors. The definitions of human error are typically framed from the point of view of the subject with reference to cognitive processes and the context of execution. That is, human error is viewed as degraded performance determined by a complex set of causal factors.

The view is exemplified by a joint European effort to harmonize the safety standards of railway signalling by the European Committee for Electrotechnical Standardization CENELEC. The CENELEC standards assume that safety relies both on adequate measures to prevent or tolerate faults as safeguards against systematic failure (man-made failures in specification, design, manufacturing, installation, operation, maintenance, and modification) and on adequate measures to control random failures (hardware faults due to the finite reliability of the components). Given that the CENELEC considers unfeasible to quantify systematic failures, safety integrity levels are used to group methods, tools and techniques which, when used effectively, are considered to provide an appropriate level of confidence in the achievement of a stated integrity level by a system. The required safety levels connected with the 'man-made' unreliability are achieved through the satisfaction of standards of quality and safety management (CENELEC REPORT, 1999). The safety balance of the system is assessed through the concept of Safety Integrity Levels, a measure of four discrete levels that enables the comparison of the qualitative and quantitative estimation of risks. The CENELEC standard provides tables where safety integrity levels correspond to intervals bands for hazard rates, which are the result of the estimation of the quantitative assessment. Safety levels and risk tolerability criteria depends on legislative principles, such as the Minimum Endogenous Mortality (MEM) or the French Globalement Au Moins Aussi Bon (GAMAB).

2.2.2 Human Reliability Assessment

Human Reliability Assessment is a discipline that provides methods for analysing and estimating risks caused by unavoidable human errors, as well as assessing how to reduce the impact of such errors on the system. Three functions of HRA are identified (Kirwan, 1994):

1. Human error identification: What errors can occur?

2. Human error quantification: How probable is it that the errors occur?

3. Human error reduction: How can the probability that errors occur be reduced?

HRA is regarded as a hybrid discipline, founded on both a technical, engineering perspective (to provide understanding of the technical aspects of systems) and psychological perspective (to provide understanding of the psychological basis of human error). The combination of these perspectives provides a foundation for assessing a total risk-picture of a system, and to determine which factors impose most risk (human or technical).

HRA dates back to the early seventies, when the nuclear industry developed systematic tools for the analysis and the estimation of the operators' contribution to plant risk and safety. There are nowadays many HRA methods available and several general approaches to HRA in the nuclear sector, with some being developed or adapted to other industries as well –such as petrochemical, aviation and air traffic management.

HRA has been a purely quantitative method and human error probability (HEP) was defined as:

HEP = Number of times an error has occurred / Number of opportunities for an error to occur

As we will see later, this quantitative approach was chosen to make HRA applicable to (quantitative) Probabilistic Safety or Risk Assessment (PSA/PRA). The quantitative approach was therefore necessary to ensure that human errors were included in the total risk-picture. The focus has however shifted from a pure quantitative approach, by recognising the importance of understanding the complexity and diversity of human error and its causes.

Independently of general approach and industry domain, all Human Reliability Analyses are nowadays concerned with the variability of operator's action and in particular of those actions (or lack of actions) that may initiate or influence a system event in a positive or negative way. Unpredicted human performance variability, in fact, often becomes part of the causal generation of incidents and accidents (Hollnagel, 1998). In HRA Human-machine systems are analysed in terms of the interactions between hardware elements and human operators. In the case of hardware, errors are described in terms of basic components (such pumps and valves) failing to perform the function they were designed to perform. As humans are concerned, errors are represented as failures to perform a particular task at a particular time. Tasks are in turn decomposed into basic types (such as reading an analogue meter or diagnosing an abnormal event within 10 minutes, THERP) that are associated with nominal failure probabilities, i.e. estimated failure probabilities before some environmental and personal factors (i.e. performance shaping factors) have been taken into account. Although human tasks are specified in relation to basic psychological functions (e.g. observing, diagnosing) and contextual elements are considered in adjusting the failures probabilities (both environmental and personal), in such descriptions the essence of human error remains the random human error variability associated with the basic task-function assigned – which can ultimately either be performed at the wrong time (error of commission) or not performed at all (error of omission).

One of the major criticisms to first-generation human reliability techniques can be stated in terms of the failure of giving a proper answer to the causes of human error: the failure of explaining the causes of the human variability in performing the identified tasks would undermine the validity of the proposed human error probabilities. The criticism maintains, for instance, that in order to calculate the failure probability of diagnosing an abnormal event within 10 minutes, the nature of the diagnosis, its associated task complexity and attention demands are relevant, as well as the training of the operators, the availability of procedures, the task familiarity and so on.

The issue however should not be so much that the analysis stops at some basic task and its associated failure probability, but that without an adequate description of the psychological and contextual factors it is impossible to estimate meaningful failure probabilities for the tasks typically included in reliability analyses. In other terms, if we had reliable human failure probabilities of basic tasks, and if the fault-tree model were a valid model of a situation, then we shouldn't bother on investigating the causes of, let's say, failed diagnosis in terms of internal psychological error mechanisms, since the purpose of the analysis was a quantification of system risk calculated upon the consequences of combinations of hardware and human failures. The question remains

whether current human reliability techniques provide an adequate description of the psychological and contextual factors to estimate meaningful failure probabilities and whether event and fault trees are a valid way of modelling dynamic systems.

2.2.3 Classification

A market standard in the classification of human error in human reliability analysis is the scheme proposed in the early sixties by Alan Swain (1963) (Table 3). Operator failures can be described in terms of:

1. Error of omissions: the operator fails to perform the required operation

2. Error of commission: the operator wrongly performs the required action

3. Extraneous errors: the operator performs an extraneous action.

The scheme refers to the behavioural level of human error in the performance of prescribed work tasks in a well-specified human-machine system. Operators are system elements that perform certain functions for the correct functioning of the overall system. The system design specifications provide the standard of correct functioning of humans as well as technical elements in all possible operating states. Without this premise of full predictably and, possibly, of predefined procedures for operators' tasks it is not possible to distinguish between the three basic categories. For instance it is logically impossible to distinguish between omission and commission since all commissions are by definition omissions of something. (Hollnagel, 1998). Analysts classify malfunctions in the context of the prescribed functions in a practical fashion. In particular, error of commission are distinguished from error of omission by assuming that a required task is initiated but incorrectly carried out.

Table 3. Categories of incorrect human outputs (Swain & Guttmann, 83, p. 2-16)

Category	**Sub-category**	**Examples**
Errors of commission	Omits entire step	
	Omits a step in task	
Errors of commission	Selection error	Selects wrong control Mispositions control (includes reversal errors, improperly made connections, etc.) Issues wrong command or information
	Error of sequence	
	Time error	Too early Too late
	Qualitative error	Too much Too little

2.3 Information processing

The roots of the information processing paradigm are the animal learning tradition, the stimulus-response behaviourist paradigm championed by Watson, communication theories, and finally, the cognitive program derived from Tolman's stimulus-organism-response model. The central view in the paradigm is that humans are processors of information and psychology's task is to discover and describe the functions, mechanisms, stages and limits of the human processing capabilities. The information processing metaphor is clearly analogous to the technical model used since the 1940s for the description of the first computers and later extended to all automated systems. Concepts and vocabulary are permuted from this tradition and from communication theory. Examples are central processor, working memory, bottleneck, overload, and capacity limits.

Information processing models human cognition as a process where a flow of information begins with a stimulus, passes several stages of processing (e.g. perception, attention, memory storage), which transforms the information demanding some time, and results in a response. This model maintains the stimulus-response assumptions of (a) a defined sequence and (b) a basic logical-temporal dependence on a stimulus. At the same time it elaborates the 'o' of the stimulus-organism-response paradigm by formulating empirical hypothesis regarding the proprieties of the processing systems and the communication channels.

The information-processing paradigm has been declined in two directions, a quantitative and a qualitative one. The quantitative tradition has been conducted in the laboratory within the natural science tradition of experimental psychology. It has focused on the resource limited aspects of human cognition: attention and memory. Theories are advanced to explain well-defined and manipulable phenomena, such as learning lists of nonsense syllables to quantify short-term memory, or presenting different messages to different ears and asking the listener either to repeat all words or to monitor the message of one channel in order to investigate parallel processing and interference. The drawback of these types of studies is that there is always a very strong trade-off between internal and external validity, that is, while contrasting theories can well explain controllable and paradigmatic situations, none of them tell us a lot of how people work in more complex, real tasks. Hollnagel defines this line of research as micro-cognition whose "emphasis is more on predictability within a narrow paradigm than on regularity across conditions, and more on models and theories that go in depth than in breath" (Hollnagel, 1998). Errors in this tradition are the result of capacity limitations.

The qualitative tradition has emphasised the process of cognition rather than the resources limitations. Models have elaborated on the stages of information processing, such as diagnosing and decision-making, in more complex situations than those analysed in the quantitative orientation. Qualitative information processing has been carried out within the cognitive tradition. Although this is a very recent discipline without generally established aims and principles, it is probably correct to locate as central themes (1) the representation and application of stored knowledge, (2) the study of realistic tasks where the interaction with the environment is taken into account, and (3) the consideration of controlled, conscious processing modes as opposed to automatic, unconscious ones.

The cognitive tradition became popular in the mid seventies when a number of different research trends in psychology, the work of Minsky (1975) on computer vision, Rumelhart (1975) on the interpretation of stories, and Schmidt (1975) on motor skill learning, reintroduced Bartlett's concept of schema (Bartlett, 1932) and the active role of cognition initiated by the Gestalt tradition. These theories subscribe to the principle that the total is more than the sum of is parts thus rejecting an atomistic view of mental processes. Although differences in language, schema theorists agree that there are high-level knowledge structures already active at the early stages of information processing. Perception, for instance, is not a passive collection of external bits of information but is dependent on the preconceived knowledge structures that anticipate much of what will occur, and which are activated by combinations of external and internal triggers. Missing features of the environment can thus be provided by memory or can simply assume default values. Errors can occur either by the activation of wrong schemata or by the faulty assignment of values to some schema variables. As this process is not limited to perception, but concerns all cognitive activities, errors can generally be conceived as over-adapted responses, a psychological tendency to over rely on past experience. The idea of "default assignment" is the central thesis of Reason's influential model of human error.

The most referred information processing model is of Wickens (1992) (Figure 1). Wickens's model is a good illustration of the orderly sequence of stages of processing in the information paradigm. The starting point are the stimuli from the environment that are received by the senses. The sensed stimuli are initially processed through sensory modality (e.g. visual, auditory, kinaesthetic) specific short term sensory stores (STSS) where the representation of physical cues are prolonged for a short period after the stimulus has physically terminated. Wickens attributes the following properties to the sensory stores: (1) they do not require conscious attention to prolong the presence of the stimuli after these have ceased, (2) they are relatively veridical, meaning that they preserve details of the stimuli, and (3) decay times are dependent on sensory modality, but are always rapid (one to eight seconds).

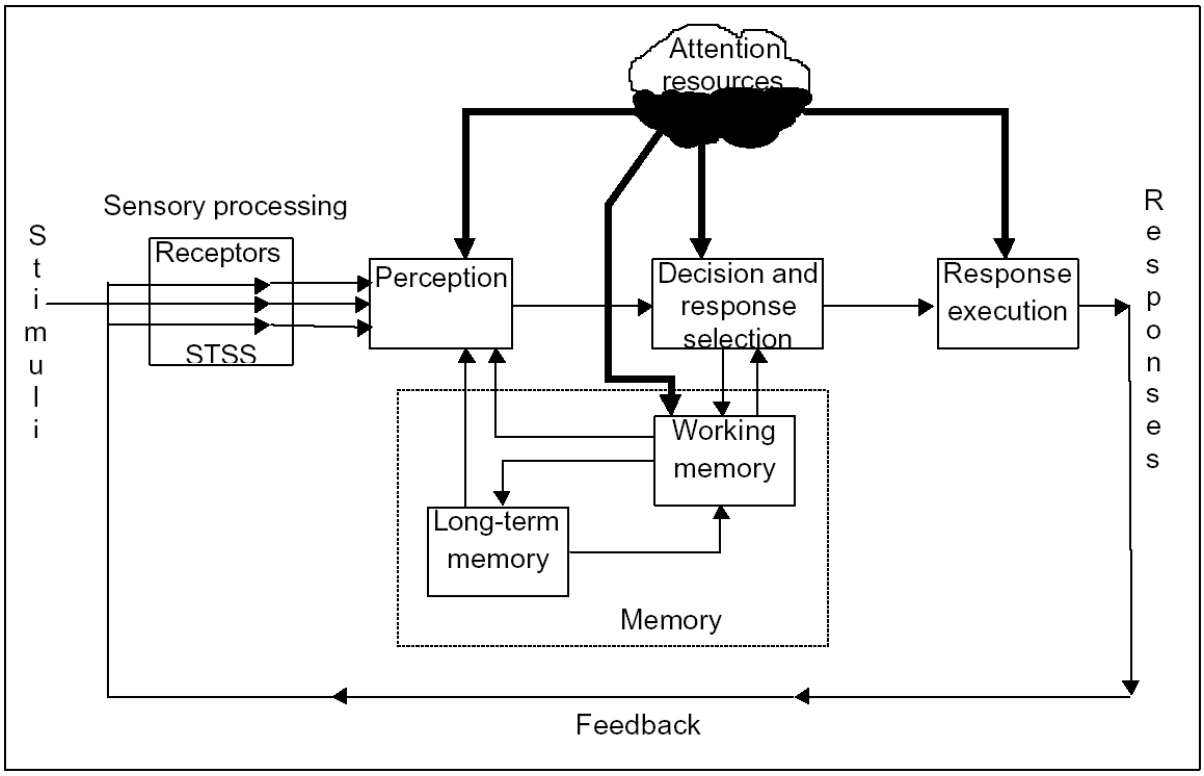

Figure 1. Wickens model of information processing

Perception concerns integrating and assigning meaning to the sensory inputs. The most basic form of perception is detection, which concerns determining whether a sign or target is present. At a higher level of processing the targets are assigned to the class they belongs to - a process referred to as recognition or identification. Perceptual judgements are distinguished between absolute and relative. Absolute judgement concerns identifying a stimulus on the basis of its position along one or several stimulus dimensions (e.g. length of an object) whereas relative judgement concerns determining the relative difference between two or more stimuli (e.g. which object is longer).

The next stage in the model is called decision and response selection and occurs after meanings to the physical cues have been assigned (a stimulus becomes information once it has been assigned meaning). Decisions may be rapid or thoughtful, and the individual may choose to execute a response. Alternatively, information may be stored in memory for a short period (seconds to minutes) in working memory by rehearsal. According to Baddeley and Hitch (Baddeley & Hitch, 1974), working memory consists of three stores:

1. A modality-free central executive - this has limited capacity and is used when dealing with cognitively demanding tasks.

2. An articulatory loop (a ‘slave system’) - this holds information in a speech-based form.

3. A visuo-spatial scratch pad or sketch pad (a ‘slave system’) - this is specialised for spatial and/or visual memory.

Information can be transferred for a longer period (hours to years) in long-term memory by learning.

The model distinguishes between the decision to initiate a chosen response from its execution. The latter phase is denoted response execution. In this phase errors are typically associated with problems of automaticity, which refers to the fact that people are able to execute highly practised action sequences with little attention. Such activities are associated with a specific type of error, namely slips.

The outcome of the decision can function as a basis for further pick-up of cues and decision-making. This is represented by the feedback loop in the model. However, the information flow does not always start with external stimuli, but can be triggered by internally as, for example, by thoughts in working memory. In addition, the flow needs not be necessarily from left to right, as in the case where immediate experience represented in working memory affects perception.

An important aspect of information processing models is that the stages of perception, decision-making and response selection and execution are, as illustrated in Wickens model, largely dependent on the available attention resources. It is hypothesised that there exist a limited amount of attentional resources that can be distributed among the different processing components.

2.3.1 Classification in information processing

As an example of classification in the information processing tradition I present The Human Error Reduction in ATM (HERA) technique, developed by EUROCONTROL within the scope of the EATCHIP/EATMP (European Air Traffic Control Harmonisation and Integration programme/ European Air Traffic Management Programme) work programme.

2.3.1.1 Model

HERA is based on Wickens’ (1992) model of information processing. However, a number of modifications were introduced to tailor it to ATM and to resolve various criticism. Working memory follows from perception and contains the controllers’ mental representation of the traffic situation. A ‘self-picture’ is introduced, i.e. thoughts that controllers have about themselves and their ability to cope with the traffic situation. In addition, decision and response selection are two separate processes (Figure 2).

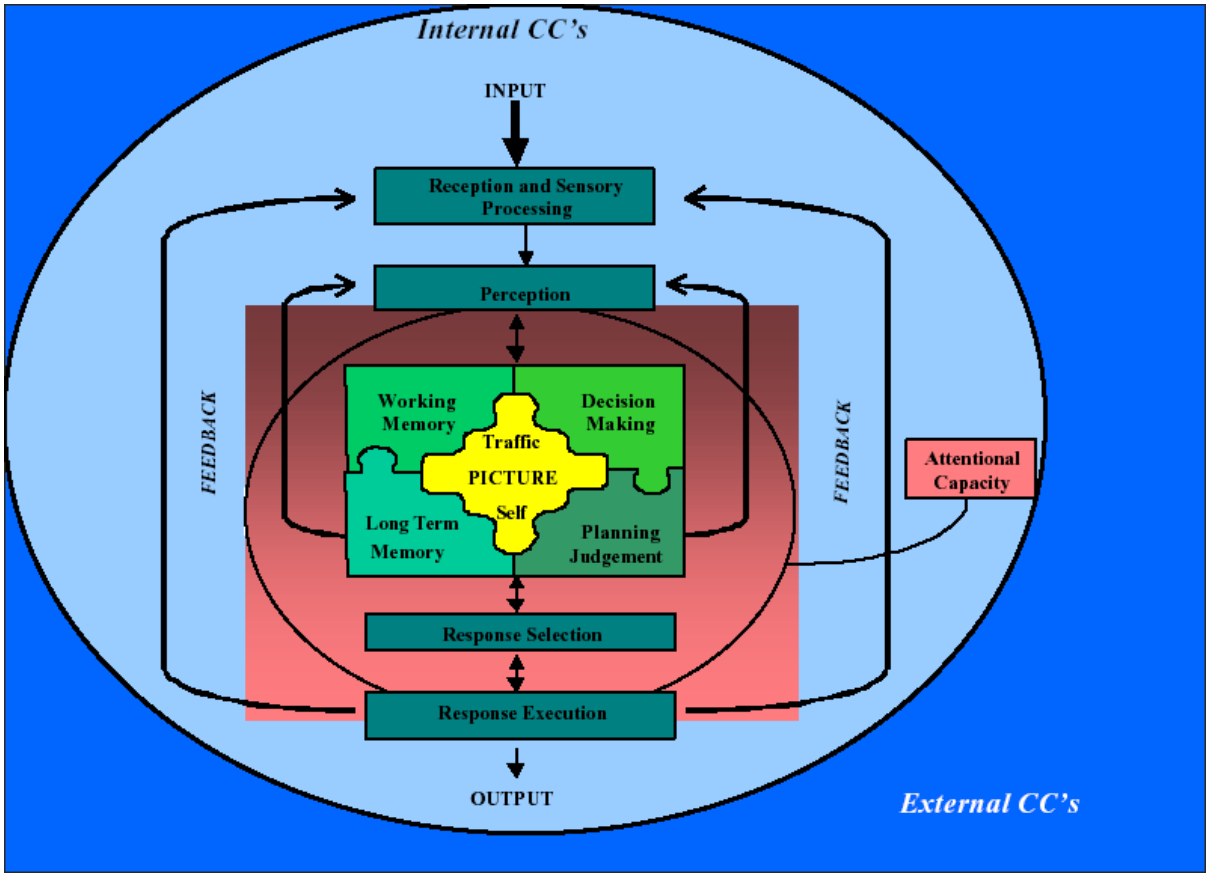

Figure 2. Hera model of information processing (Isaac et al., 2002).

2.3.1.2 Classification scheme

The classification system of HERA consider two factors:

- The error, i.e. what error occurred (type), how did the error occur (mechanisms).
- The context, i.e. when did the event occur, who was involved and what was their involvement (including the organisation factors), where did it occur, what tasks were being performed, how did the event occur, and what information or topic did the error involve.

HERA employs a quadripartite distinction between error: Error/Violation Types (ETs), Error Detail (ED), Error Mechanisms (EMs) and Information Processing Levels (IPs).

The Error/Violation Types (ETs) are the way the action manifested itself externally. In order to decide what is the 'right' or wrong' action all relevant procedures and expected actions will be considered. ETs include errors of omission, timing, sequence, quality, selection and communication, (examples are: omission, action too late, mis-ordering, extraneous act, right action on wrong object, and incorrect information transmitted).

Procedural violations are actions that contravene a rule, procedure or operating instruction. Procedural violations are more complex than errors and are defined in terms

of: procedures, controller intention and awareness, working practices and circumstances.

The Error Detail (ED), as well as the Error Mechanism and the Information Processing levels, describe the error from a psychological perspective. There are four ED domains covering all the information processing activities:

- Perception and vigilance
- Memory – working and long-term
- Planning and decision-making
- Response execution.

The EDs classify the error at a gross level (e.g. error of 'working memory' or 'response execution') and direct the user to a subset of errors within the relevant ED domain - the Error Mechanisms (EMs). The Error Mechanisms (EMs) describe the internal manifestations of the ED (e.g. misidentification, late detection, misjudgement). They refer to the cognitive function that has failed. The Information Processing Levels (IPs) describe how the psychological cause influences the Error Mechanism EM within each Error Detail (ED) level. These 'psychological causes' refer to inherent human fallibility which influence behaviour, such as visual discrimination, expectations, working memory capacity, confusion, habit, etc. For example, the IPs within the ED 'perception and vigilance' include 'expectation bias' (i.e. seeing or hearing what one expects to hear), 'information confusion' (i.e. confusing two things that look or sound alike), and 'distraction/preoccupation' (i.e. temporary interruption by an external event or more prolonged loss of concentration due to internal thoughts).

There are three categories in HERA that describe the context at the time of the error: the Task, The information & Equipment, and the Contextual Conditions (CCs). The Task describes the function(s) that the controller was performing at the time the error was made. Example tasks include: coordination, tower observation, planning, R/T communications and instruction, control room communications, strip work, materials checking, radar monitoring, HMI input & functions, handover/relief briefing, takeover, training, supervision, and examination. The Information/Equipment lists describe the environment in which the error occurred. Examples of HERA information/equipment elements are: procedures, coordination, aircraft type, geographical position, airport, flight rules, secondary radar, visual approach aids, aerodrome equipment, flight information displays, and Input devices.

Contextual Conditions (CCs) can be defined as factors, internal or external to the controller, which influence the controller's performance of ATM tasks. Contextual Conditions (CCs) can help to explain why the error occurred. CCs include the following sub-categories: pilot-controller communications (e.g pilot breach of R/T standards/phraseology), pilot actions (e.g. responding to TCAS alert), traffic and airspace (e.g. excessive traffic load / complex traffic mix), weather (e.g. extreme wind at high altitude), documentation and procedures (e.g. inappropriate regulations and standards), training and experience (e.g. controller under training), workplace design,

HMI and equipment factors (e.g. R/T failure), environment (e.g. lighting - illumination, glare), personal factors (e.g. high anxiety/panic), team factors (e.g. poor/unclear coordination), organisational factors (e.g. problems in the work environment), and administrative workload problems. There may be more than one CC for an error. Contextual factors are also important for the creation of an error database.

2.3.1.3 Method

The analysis uses incident investigation reports, proceeds from the beginning of the description and moves forward in time, and creates a description of how human errors propagate and result in incident. The classification system of HERA was developed in two forms, a tabular format and a series of decision flowchart diagrams. Tables are available for Tasks, Information and Equipment, Error/Violation Types (ETs), and Contextual conditions (CCs). The diagrams allow the analysts to identify errors by answering a series of 'Yes/No' questions. There are separate decision flow diagrams for: Error Detail (ED), Error Mechanisms (EMs) for each error detail, Information Processing levels (IPs) for each error detail, and Contextual Conditions (CCs) sub-categories. Each decision flow diagram starts at a different Error Detail (ED) domain. This allows the analyst to start at the applicable ED and makes the technique more resource-efficient. The decision flow diagrams allow the analyst to begin at any ED domain. Also, the format allows the analyst to skip ED domains where they are confident that the error did not occur within that area, or where the analyst is directed to 'jump' to another ED domain by following the diagram.

HERA's internal structure of Error/Violation Types ETs, Error Mechanisms EMs and Information Processing Levels IPs allows the analyst or incident investigator to classify errors at three levels of detail. There should almost always be sufficient information to classify the ET, and usually there will be enough information to classify the EMs. IPs add value to the analysis, but are the most difficult 'level' to classify, because there is sometimes insufficient information to determine the psychological cause (See Figure 3 for a pictorial description of an incident).

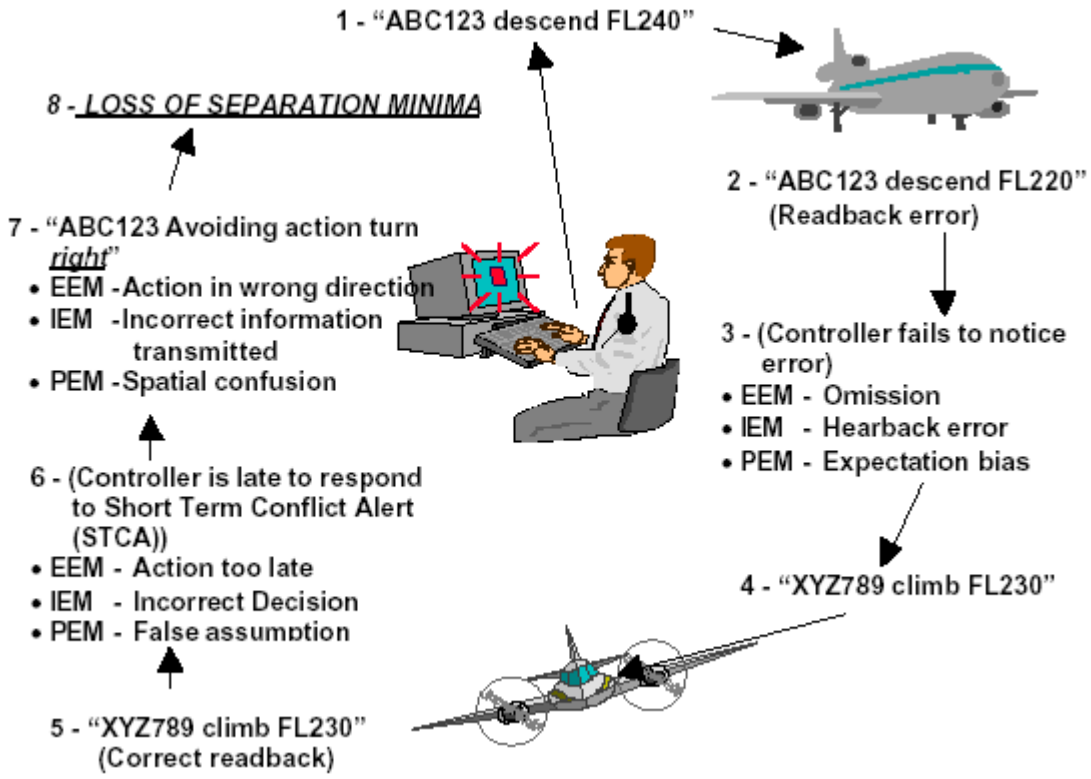

Figure 3. Pictorial description of an ATM incident and associated HERA error classifications (Isaac et al., 2001b).

2.4 Cognitive System Engineering

Cognitive System Engineering (CSE) implies a distinctive approach to human performance and human error in complex human-machine systems. Cognitive system engineering is a theoretical perspective on human-machine systems design and human performance modelling that has been advanced in the works of Erik Hollnagel and David Woods (Hollnagel & Woods, 1999; Hollnagel, 1993; Woods, 1986; Woods, 1988). Cognitive system engineering is a response to the rapid technological progress produced by the digital revolution. Since the late seventies, the evolution of machines towards becoming 'intelligent' agents as well as the occurrence of certain events, as the Three Mile Island incident, demonstrated the deficient status of Man-Machine Systems knowledge at the time.

In cognitive system engineering the human factor problem of assisting performance is translated into the one of achieving human-machine integration, where the role of machines has changed due to increased computer power and the potentialities of artificial intelligence.

The central tenet of this perspective is that contemporary human-machine systems are best viewed in terms of joint cognitive systems, and should therefore be designed, analysed and evaluated as such. A physical or biological system is considered a cognitive system when it satisfies the following criteria (Hollnagel et al., 1999):

1. Its behaviour is goal oriented

2. Its behaviour is based on symbol manipulation

3. It uses knowledge of the world

4. It is adaptive to new circumstances and can see problems in more than one way.

A cognitive system possesses knowledge about itself and the environment and is able to plan and modify its actions on the basis of that knowledge. It is in this sense concept driven as opposed to data driven non-cognitive systems whose actions are simpler responses to stimuli. Concept driven behaviour is produced by means of models of representations of the environment. Humans are clearly cognitive systems and use models of the environment in intelligent behaviours such as planning and deciding. Human-Machines systems are also cognitive systems, even when the machine part alone does not satisfy all the criteria to qualify.

The operators and the process or machine have to be modelled on equal terms and the interaction between the human and machine parts of the human-machine cognitive ensemble is the focus of CSE. In particular, this interaction is thought to be more complex and dynamic of the way the environment is set to interact with the human operator in the information processing model (i.e. signals, executions, disturbances). In the words of David Wood (Woods, 1988):

> "The configuration or organization of the human and machine components is a critical determinant of the performance of the system as a whole. The joint cognitive system paradigm demands a problem-driven, rather than a technology-driven, approach where the requirements and bottlenecks in cognitive task performance drive the development of tools to support the human problem solver." (p. 153).

There are two extremes in the way this interaction can be configured. Intelligent machines can be used as tools that expand human capabilities. The machine amplifies the fundamentally correct human capabilities, overcoming some of their limitations (e.g. memory, attention resources). Human performance is shifted in a different yet higher level. In a man-machine system designed in this way the locus of control is the human part. At the other extreme, machines are conceived as prostheses, replacements or remedies for human limitations. The machine compensates the deficiencies in human problem solving and reasoning. In a cognitive support system designed in this way the locus of control of the joint system is the machine, while the human is the data gatherer and action implementer of the stand-alone problem-solver machine.

While the latter view of machines as prostheses is sometimes endorsed by information processing approaches, Cognitive System Engineering is definitely oriented towards the former view of machines as tools. The joint cognitive system perspective defines a computer consultant as

> "a reference or source of information for the problem solver. The problem solver is in charge; the consultant functions more as a staff member. As a result, the joint cognitive system viewpoint stresses the need to use computational

> technology to aid the user in the process of solving his problem." (Woods, 1986), p. 161).

The consultant does not provide a solution plus a solution justification, but performs the role of a good advisor. "Good advisory interaction aid problem formulation, plan generation, (especially with regard to obstacles, side effects, interactions and tradeoffs), help determine the right question to ask and how to look for or to evaluate possible answers." (Woods, 1986), p. 160).

Methodologically CSE investigates human performance on a global level and in realistic tasks. The focus is on the overt phenomena, on how tasks (controlling a process, conducting a car) and cognitive activities (such as planning, decision making) are performed and achieve their goals, rather than on the underlying psychological mechanisms of the cognitive activities implied. CSE can thus be defined as ecological, in the sense that the study of humane performance and problem solving behaviour is considered meaningful only in relation to the tools (i.e. support systems) and the world that drive the behaviour under study. This is in contrast with most laboratory research which analyse problem solving without tools and hence with the risk of eliminating critical features of performance (Woods, 1988).

2.4.1 Human error in CSE

The problem-driven approach and the view of support systems as tools lead to the second central assumption of the Cognitive System Engineering perspective: human cognition is an active process dependent on two equally strong determinants, the operator's goals and the situation or context. The influence of the environmental circumstances, the context of performance, is the core of the description and analysis of human error, as performance failures are not traced back to the internal information processing malfunctions but interpreted as mismatch between cognition and working conditions. This mismatch is inevitable in complex systems and it is due to the unanticipated variability that characterizes all highly dynamic and highly coupled worlds. Unanticipated variability is the result of "underspecified instructions, special conditions or contexts (violations of boundary conditions (...) or impasses where the plan's assumptions about the world are not true), human execution errors, bugs in the plan, multiple failures, and novel situations (incidents not planned for)" (Woods, 1988).

Human error in CSE, or in Hollnagel's terms, erroneous human action, is thus the consequence of two determinants: the human-machine mismatch and the inherent human variability. Human machine mismatch can thus in principle be eliminated or minimized through design, although it is impossible to predict all possible operating situations in advance and therefore no totally matched man-machine system is possible. On the other hand, even a hypothetically completely matched system will not be free from human errors since there always remain the possibility of residual erroneous actions, which are due to inherent human variability, i.e. "random fluctuations in how the mind works, subtle (or even less subtle) influences from the environment, forgetting, loss of attention, associative jumps, etc." (Hollnagel, 1993). Since residual erroneous actions and knowledge mismatches naturally blend in daily operations it's empirically difficult to isolate them. This would create a problem for a quantification of human error if performance were decomposed in assumedly basic actions outcomes.

It follows that the ideas of failure mechanisms and processing limitations that are central in the information processing approach to human error are considered unnecessary. In the first place these mechanisms are hypothetical constructs. In the second place, they can be studied only in isolation or laboratory tasks (micro-cognition). And finally, the contextual effects on human performance are underrepresented and underestimated, as a result of the previous two points. An alternative approach for the classification of the causes of human erroneous actions and the assessment of human reliability is developed by adopting a more holistic approach that starts with a thorough evaluation of the global context of performance. In this case, the situation and the task are set to determine the goals of the agent, which in turn govern the control of actions as well as the number and sequence of cognitive functions that will be active in the process. Each resultant control mode active in the particular task is then associated with an estimated level of human reliability or intrinsic performance variability.

For CSE the response to unanticipated variability is not creative problem solving but coordination by resource management. This is a process where background knowledge is gradually unfolded in the process of monitoring and adapting plan execution in reaction to deviations from pre-planned responses. CSE individuates a series of strategies to help the users deploy the background or meta-knowledge they possess, or in other words, to enhance demand-resources match:

(1) The study of human-human cooperation, where the machine is thought of as an cognitive agent

(2) The supervisory control model where the human has ultimate authority. Shared representation is required by the two agents, as well as the supervisor must be able to redirect the lower machine.

(3) The view of machines as extensions and expansions, where people are tool builder and tool users. Tools magnify capacity for physical work, perceptual range, and, as CSE is particularly concerned, cognitive environment. The latter is enhanced by calculation power, search and deductive power, end economy of representation in terms of conceptualisation power. This is the ability to experiment, to visualise the abstract, and the enhancement of error tolerance by feedback about effects/results.

2.5 Risk management models

By risk management models I indicate a class of approaches that in the literature goes under the various names of error management (Bove, 2002), treat management (Helmreich et al., 1999), resource management (Amalberti, 1992), and control of danger (Hale et al., 1987), to name a few. Common denominator for these approaches is that they pursue system safety without concentrating on errors per se but on the generation and propagation of system hazards, including human generated hazards, and on the way these are prevented to result into accidents. In describing such processes risk management approaches take into account the whole process of successful and unsuccessful performance, and clarify how hazards and errors are anticipated and detected, how they are recovered, and in general, how humans achieve and maintain control in risk situations.

Similarly to human cognitive engineering these approaches underlie the active and anticipative role of the individuals in their strategies to maintain safety. Amalberti (1992) has proposed a risk management model, or "model of anticipation", to describe the way individuals control safety-critical processes. At the core of the model are anticipation and action, as the means by which risk, which is a normal task component and part of operators' competence, is managed. Operators know their resources are limited, and are therefore forced to take risks. However, they have a representation of their competence (meta-knowledge) that allows them to continuously maintain risk at the lower level compatible with expected performance. They thus develop strategies to adapt to task demands, i.e. to manage risks, to maintain performance within safety margins, to switch between short-term activities and long-term reasoning, etc. These strategies typically involve two stages. The first stage involves anticipating the course of the situation. Based on simplified heuristics (e.g. application of known schema, hypothesis testing, saliency and frequency of occurrence) and previous experience, actions are continually taken from a repertoire of known solutions in order to force risky situations within known safety margins, i.e. within pre-planned models of the situations. It is clear at this stage that risks are managed by taking other risks. The second stage is directed at the definitive solution of the problem with the identification of the causes of the deviations. Depending on the success of the first stage the risks are finally minimised at the second stage. The two stages involve different cognitive mechanisms: the first stage involves primarily skill and rule-based reasoning but "reactive behaviours are kept at minimum in applying prepared responses and escaping, freeing time in order to optimise response" (Amalberti, 1992, p. 103). The second requires pre-eminently knowledge-based reasoning, and requires time for elaboration and mental effort.

The error classifications inspired by this class of models are not conceptually different from those based on information processing. That is, the human error is still a cognitive category, while the difference resides on the fact that the classification is not focused on the human error per se, but rather on the entire human performance in safety critical situations. In other terms these approaches, similarly to CSE, are coupled tightly to a system safety model that assign a role to the human performance that is as much interested in explicitly modelling the positive contributions to safety as it is the negative ones. A very well developed classification system based on a error recovery model is the one developed by Bove (2002).

2.6 Violation models

We have seen that the conceptual difference between errors and violations is less clear-cut than one might intuitively assume. The fact that a liberal conceptualisation of the concept of violations allows to include a large number of unsafe acts has been exploited to develop an alternative approach to the study and management of systems safety. This perspective parallels the traditional study of human error insofar it also identify few external manifestations of violations and a more articulated set of their causes. There is an important difference however, in the fact that the causes of violations are not (essentially) identified based on cognitive theories of human performance but on motivational and attitudinal frameworks. The most important consequence of this change of perspective is that their classifications system are (at present state) not as much tools for learning from past and predicting future human contributions to system failures, but theoretical models to support techniques and methods that can help

mapping and reducing the potential for a special class of human hazardous behaviour: individual and group violations.

In the previous discussion of human error in section 1, I have pointed out the potential risk in confounding between two types of unsafe acts: errors and violations. This is not surprising since the conceptual boundaries between these concepts are by no means rigid. In everyday language the word violation covers both intentional and unintentional acts. Road transportation provides easy examples of this: a car driver can pass a red traffic light because he didn't notice it or because he was in a hurry. In the human error literature, instead, there is a prevalence of defining violations in terms of intentional behaviour. Reason (1990), for example, defines violations as "deliberate –but not necessarily reprehensible– deviations from those practices deemed necessary (by designers, managers and regulatory agencies) to maintain the safe operation of a potentially hazardous system" (p. 195). Similarly, Mason (1997) describes violations as 'deliberate deviations from the rules, procedures, instructions or regulations introduced for the safe or efficient operation and maintenance of the equipment" (p. 288).

However, in the same tradition, Collier, (2000), considers violations to "span the full range of intention ... (as they) ... include acts that are totally habitual and 'unconscious', as well as fully deliberate acts" (pp. 3-4). Collier's position is in fact more consistent with human factors practice. Both Reason and Mason accept a classification of violations that includes 'routine violations', a behaviour contrary to the rules that has become the norm, and which is executed in a 'automatic and unconscious' (Mason, 1997) manner. Reason, in the road traffic context (Reason et al., 1990), labels the one he reputes the most interesting class of violations as 'erroneous or unintended violations', a type of activity that is deliberate but which has not the prior intention to cause injury or damage. So it seems that beside acts that intend to cause harm or damage, i.e. acts of sabotage and vandalism, there is a potential overlap between what would be defined as error and what as violation. Following Leplat (1993), in the case of errors:

> "one finds a distinction analogous to that of objective/subjective responsibility, but translated in terms of error for the expert/error for the subject. The error for the expert defines the divergence between what it is expected by a subject (result, procedure...) and what it is really done. The error for the subject is the divergence between what he/she wanted to do and what he/she thinks to have actually done."

In this scheme violations would be the divergence between what the 'expert' expects, i.e. the prescribed task, and what the subject intends to do, the re-defined task, when this divergence is not due to a misunderstanding of the task, since in this case we would rather talk of a mistake (for the expert). The subject would generally be aware of the divergence (i.e. that there is some different expectation on his/her behaviour), at least at some level of consciousness. In other terms, he/she knows the prescribed task but cannot or does not want to follow it in some circumstances.

The interesting part in Leplat's scheme is that the 'task prescribed' does not only include previously specified procedures, rules and instructions, but it is to be intended in a broad sense, including vaguely specified 'missions', the principles to be considered in the evaluation of a task, as well as the conditions of execution. Violations are such only

if the context of rules, values and procedures determine them to be so. As Reason (1990) puts it, “violations can be described only in relation to a social context in which behaviour is governed by operating procedures, codes of practice, rules and the like” (p. 195).

2.6.1 Causes of violations

The causes of violating behaviours point to the motivational, emotional and attitudinal dimensions of human behaviour. In contrast with cognitive approaches, violation based research approaches for the reduction of system accidents and damages do not focus on the individual’s elaboration of information. Instead, they consider most human errors and unsafe acts as violations of existing rules, group norms and safe practices, and therefore investigate the group, organisational and social dimensions of behaviour. The accent is then on why individuals do not follow rules and procedures, on why prescribed tasks and redefined tasks diverge, and on the factors that influence such redefinitions.

Violations can be organised around the principle of goal conflict. In this way three classes of real or perceived goal conflict can be identified:

1. Between the individual and the organization

2. Within the organisations’ goals system

3. Within the individual’s goals system

An example of conflict between the goals of the individual and the organisation is when individual and organisation are maximising different and conflicting things (Hollywell & Corrie, 2000). For example, while an organisation may wish to constantly maximize safety, an individual might decide in particular situations to maximise productivity, speed, comfort, financial benefit, or social conformance.

However, there are cases where individuals commit violations in order to maximise safety. In extreme and abnormal cases, rules and procedures might not be adequate, applicable or even available. Examples of this kind of situations can be found in nuclear domain (e.g. the Davis-Besse incident, see Kirwan, 1994) and from transportation (e.g. the Clapham Junction incident, Hidden, 1989). Hollywell & Corrie (2000) maintain that “It can be argued that in particular (and hopefully, extremely rare) situations the violation of a rule or procedure could lead to increased safety” (p. 3). They point to the fact that in those occasions operators have to choose between obeying a rule or using their knowledge of the system, and reflecting of the effect of recent changes on the UK Railway system, they conclude that one “cannot have both prescription and adaptability in the same context; one approach has to be chosen to manage safety” (p. 3). This conflict might not be as clear-cut as it seems in this sentence, as all safety relevant systems contain both elements, yet the problem of the optimal configuration of prescription and adaptability might be rightfully considered one of the major challenges in safety-systems design.

One supporter of the adaptability alternative is Ruiz Quintanilla (Ruiz Quintanilla, 1987) who reputes that allowing the operators the possibility of individual choice of

problem solving strategies would result in the avoidance of “monotony, satiation and error proneness” (p. 127), and, in this context of violation, one would add, to the reduction of violating behaviour. Ruiz Quintanilla grounds his claim in psychosocial research on the meaning of working in different social groups. Workers identify with their work and profession provided they dispose of discretionary freedom and the potential for self-regulative activities within the organisation. System and job design should take into account not only to the cognitive capabilities of the operators but also preference structures, and job expectations, in general, work-related value orientations.

Connected with the dilemma between prescription and adaptability is the role of automation in socio technical systems. Automation can constrain the individuals’ safety control strategies and plans by introducing a system rigidity that contradicts human flexibility and anticipation. Violations here take the form of overriding the automaton by deactivating some protection systems in order ensure safety on the individual’s own premises. This kind of violation is more generalised than one would expect, as Amalberti (1992) found in the aviation context:

> ... Strategies of detouring systems from their standard uses are quasi-systematic with expert pilots, but with various levels of frequency and risk taking (in a questionnaire, 86% of population of fighter pilots responded that they were using such strategies frequently or fairly frequently) (p. 104).

A third related potential source of conflict between the individual and the organisation lies in the justification of rules and regulations: although their formulation attempt to ensure safety and other systems’ goals, their efficiency is not necessary self-evident. That is to say, as there is not guarantee that following procedures in any case will guarantee the system goals, there is likewise no guarantee that not following them will result in negative systems consequences. The links between regulations and systems negative outcomes are at best logical and empirical and at worse taboos and superstitions. Insofar as individuals will retain the freedom to choose behaviour the potential for rule violations will always be present, even to the point of becoming normal behaviour. It is well documented in the literature that informal group norms and behaviours develop as result of working groups interactions with technical systems, and these might deviate from the norms prescribed by the designer or the organisation.

The shift of attention from the individual to the social dimension of unsafe acts makes a violation approach particularly attractive from the point of view of the measures an organization can put in place to reduce the potential for undesired consequences. It is clear, in fact, that violations are strongly influenced by company management; they are created by, accepted by or condoned by management.

2.6.2 Violation and risk-taking behaviour

Performing tasks in a manner that violate rules and regulations, which are introduced to guarantee the safety of the socio-technical systems, would expectedly affect the level of risk connected with such activities. It is, therefore, not surprisingly to find individuals’ judgments and perceptions on risk and safety as determinants of violating behaviour.

In road traffic research the concept of risk compensation has been introduced to explain why certain safety measures achieved less than expected results. According to the risk compensation theories the individual perceives a level of risk connected with a particular activity and adapts his/her behaviour in order to reach certain personal goals at a certain target level of safety (Wilde, 1974). For example, the motivation for saving time may imply that improvements of roads (e.g. wider lanes or shoulders) result in higher speed, with no net effect on safety as a consequence.

However, the idea of perceiving and weighing risks in safety relevant behaviour is controversial. For instance, Taylor (1987) claims that incidents occur as a result of risk taking behaviours, but these are not rational decision making situations. Taking risk, in this perspective, is not balancing situations, as negative consequences cannot be weighted. The stochastic property of accidents and "divergence" in catastrophe theory, i.e. many consequences from same antecedents, cause negative stochastic events to be perceived as "black holes", something incomparable and without meaning, in contrast with positive stochastic ones, i.e. which motivate the intentions to take risks. Instead, negative consequences take the form of

> "fears of no-conformity, to respected opinions, social deviance, or lawbreaking. A driver venturing on a short journey at night without functioning lights probably does not feel that his life is at risk, but may well fear that the police will stop him, or that other drivers will protest, or that if an accident does happen he will be made to take the blame, or even that to behave in such a way is to act in a foolish and uncaring manner inconsistent with his image of himself. The balance may, in other words be a moral one." (Taylor, 1997)

This 'balance' might alternatively considered a motivational and attitudinal one rather than a moral one, and theoretical explanations could be provided in psychological terms although not prominently cognitive ones. At any rate, it should be stressed that, although risk perceptions and attitudes are probably present in all violating behaviours, there seems to be a much richer combination of determinants involved in the choice and execution of 'redefined' tasks, at least insofar many such redefined task would not be perceived as changing (and even would not change) the level of risk connected to the activity.

2.6.3 Classification in the violations framework

Mason (1997) describes violations as "deliberate deviations from the rules, procedures, instructions or regulations introduced for safe or efficient operation and maintenance of equipment" (p. 288), and estimates that "up to 90% of accidents occur when an individual, or individuals, deliberately contravenes established and known safety rules" (Mason, 2000).

He has proposed a classification of violation based on the factors that influence a person's decision to break rules. These are considered at two levels: factors that motivate the violation and factors that influence the decisions to violate. Factors of the first type are called 'direct motivators', which directly motivate management and operating and maintenance personnel to break the rules. Factors of the second type are named 'behaviour modifiers', which could increase or reduce the probability of any

individual deciding to commit a violation. Table 4 lists the direct motivators and behaviour modifiers identified by Mason.

Table 4. Mason's classification of violations.

Direct Motivators	Behaviour Modifiers
• Making life easier • Financial gain • Saving time • Impractical safety procedures • Unrealistic operating instructions or maintenance schedules • Demonstrating skill and enhancing self-esteem *There could also be:* • Real and perceived pressure from the 'boss' to cut corners; • Real and perceived pressure from the workforce: (a) To break rules, (b) To work safely.	• Poor perception of the safety risks • Enhanced perception of the benefits • Low perceptions of resulting injury or damage to plant • Inadequate management and supervisory attitudes • Low chance of detection due to inadequate supervision • Poor management or supervisory style • Poor accountability • Complacency caused by accident environments • Ineffective disciplinary procedures • Inadequate positive rewards for adopting approved work practices

Mason's classification includes both individual and organisational factors in both classes of factors, but the emphasis is on the organisational level.

The UK's Human Factors in Reliability Group (HFRG) published, jointly with the UK Health and Safety Executive (HSE, 1995), a methodology for addressing procedural violations that is grounded on Mason's classifications of direct motivators and behaviour modifiers. This approach identifies organisational factors that increase the potential for violations. Such organisational factors include safety commitment, training, management and supervision, job design and equipment design. The approach can be applied by the non-specialist and is applicable to a wide range of industries.

By means of interviews and questionnaires applied to generic or specific sets of rules, the approach is designed to identify those set of rules and procedures within an organisation which could have the greatest potential impact on safety, if not followed. Each set of rules or procedures is assessed using a checklist. 48 questions have been created to map generic rule sets, as it would be impractical to complete a checklist for every safety rule of an organisation.

Examples from the checklist include:

- The rules does not always describe the correct way of working
- Supervision recognises that deviations from the rules are unavoidable
- The rules are not written in a simple language
- Rules commonly refer to other rules
- I have found better ways of doing my job than those given in the rules
- I often encounter situations where no prescribed actions are available.

A selection of the workforce is asked to rate the ‘degree of agreement’ with forty-eight statements with a score between zero (disagree) and six (strongly agree). The methods provide ways of analysing the answers and linking them to appropriate management strategies for minimising the potential for violations.

3 Conclusion

Although modern safety science has abandoned the obsession to precisely define, count and eliminate human error, it is still impossible to completely remove the need for the concept when working with safety-critical systems. Yet, if we want to go on productive grounds on human error, we have to go beyond folk views of the concept. We then need to recognize that human error is not a natural category, that is, human errors statements are not about state of affairs or events that we can observe in the empirical world. Instead, human error is a normative category, judgments about the conformity of actions to standards. The verification of these judgments is a process where empirical events and abstract standards are compared. These standards of correct performance and the criteria for their applicability to practical circumstances are, moreover, dependent on the models of human behaviour adopted.

A set of suggestions to those engaged with the topic of human error can be derived as conclusion of this report:

1. *Specify the standard of correctness.* This follows the normative nature of the concept human error. Any time something is defined as error, violation or unsafe act the standard assumed should be mentioned.

2. *Differentiate between ideal versus actual standard.* Ideal standards are missions, principles of correct functioning, common practices, etc. whereas actual standards are existing rules, procedures and instructions.

3. *Specify the point of view: internal vs. external*. An internal point of view is the point of view of the agent, i.e. by looking at intentions and choice of goals. An external standard concentrates on the actual behaviour. As we have seen, this difference is relevant for defining errors.

4. *Indicate the model of behaviour adopted for the analysis*. The model of behaviour determines how the standards are applied to real situations, and what types of errors and behaviours are meaningful.

5. *Describe actions and conditions and distinguish them from the assumed causes*. This point stresses the importance of the context for action, as well as the importance of clearly differentiating between causes and manifestations.

These suggestions are more a summary of the central theses held in the report than practical rules for error analysis. Nonetheless it seems that keeping them in mind, or even using them as a checklist, will improve the accountability of one’s analysis (may it be defining, classifying or predicting something as human error), as all essential elements of error analysis will be made public. At the same time, when working with

retrospective analyses, hindsight problems could be reduced as this process forces to go through one's own assumptions and reasoning.

Needless to say, these suggestions will not make the job easy, as standards of correct performance are seldom obvious and their application granted. Nor will these suggestions guarantee direct generability of findings in the case of incidents' analysis and error classifications, as there are several valid models of behaviour and various classification systems.

References

Bartlett, F. C. (1932). *Remembering: a study in experimental and social psychology*. Cambridge: Cambridge University Press.

Bove, T. (2002). *Development and Validation of a Human Error Management Taxonomy in Air Traffic Control*. PhD Risø National Laboratory, Roskilde.

CENELEC REPORT (1999). *Railway Applications: Systematic Allocation of Safety Integrity Requirements* (Rep. No. prR009-004:1999 E). Brussels: CENELEC, European Committee for Electrotechnical Standardization.

Collier, S. (2000). *A Framework for Understanding Violations and Errors of Commission* (Rep. No. HWR-634). Halden: IFE.

Hale, A. R. & Glendon, A. (1987). *Individual behaviour in the control of danger*. Amsterdam: Elsevier.

Heinrich, H. (1931). *Industrial accident prevention*. New York: McGraw-Hill

Helmreich, R. L., Klinect, J. R., & Wilhelm, J. A. (1999). Models of threat, error and CRM in flight operations. In (pp. 677-682). Austin, Texas, USA: The University of Texas at Austin.

Hidden, A. (1989). Investigation into the Clapham Junction railway accident. UK department of Transport report. London: Her Majesty's Stationery Office.

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

Hollnagel, E. (1993). *Human reliability analysis: context and control*. London: Academic Press.

Hollnagel, E. (1998). *Cognitive reliability and error analysis method: CREAM*. Oxford: Elsevier.

Hollywell, P. & Corrie, J. D. (2000). Reducing Violations on the Railways. What only Experience Can Teach. In London: Energy and Safety Division, IBC Global Conferences Limited.

HSE (1995). *Improving Compliance with Safety Procedures – Reducing Industrial Violations*. HSE Books.

Isaac, A., Shorrock S.T., Kennedy R., Kirwan B., Andersen H., & Bove T. (2002). *Short Report on Human Performance Models and Taxonomies of Human Error in ATM (HERA)*, report HRS/HSP-022-REP-02, EATMP-EUROCONTROL, Brussels.

Kirwan, B. (1994). *A guide to practical human reliability assessment*. London: Taylor & Francis.

Leplat, J. (1993). Intention and Error - A Contribution to the Study of Responsibility. *European Review of Applied Psychology-Revue Europeenne de Psychologie Appliquee*, 43, 279-287.

Mason, S. (1997). Procedural violations: causes, costs and cures. In F.Redmill & J. Rajan (Eds.), *Human Factors in Safety-Critical Systems* (pp. 287-318). Oxford: Butterworth-Heinemann.

Mason, S. (2000). Easy Way to Tackle Violations. In London: Energy and Safety Division, IBC Global Conferences Limited.

Minsky, M. A. (1975). A framework for representing knowledge. In P.Winston (Ed.), *The Psychology of Computer Vision*. New York: McGraw-Hill.

Rasmussen, J., Duncan, K., & Leplat, J. (1987). *New Technology and human error*. John Wiley & Sons.

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

Ruiz Quintanilla, A. S. (1987). New Technology and Human Error: Social and Organizational Factors. In J.Rasmussen, K. Duncan, & J. Leplat (Eds.), *New Technology and Human Error* (pp. 125-128). Chichester, UK: John Wiley & Sons.

Rumelhart, D. E. (1975). Notes on a schema for stories. In D.Bobrow & A. Collins (Eds.), *Representation and Understanding: Studies in Cognitive Science*. New York: Academin Press.

Schmidt, R. A. (1975). A schema theory of discrete motor skill learning. *Psychological Review*, 82, 225-260.

Senders, J. W. & Moray, N. P. (1991). *Human Error: Cause, Prediction, and reduction*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Skitka, L. J., Mosier, K., & Burdick, M. D. (2000). Accountability and automation bias. *International Journal of Human-Computer Studies*, 52, 701-717.

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51, 991-1006.

Swain, A. (1963). A method for performing a human factors reliability analysis, Monograph SCR-685, Sandia National Laboratories, Albuquerque, US.

Swain A. & Guttmann, H. (1983). *Handbook of Human Reliability Analysis with Emphasis on Nuclear Power Plant Applications*. NUREG/CR-1278, Nuclear Regulatory Commission, U.S.

Taylor, D. H. (1987). The Hermeneutics of Accidents and Safety. In J.Rasmussen, K. Duncan, & J. Leplat (Eds.), *New Technology and Human Error* (pp. 31-41). Chichester, UK: John Wiley & Sons.

Watson, J. B. (1913). Psychology as the behaviorist views it. *Psychological Review*, 20, pp. 158-177.

Wickens, C. D. (1992). *Engineering psychology and human performance*. New York: HarperCollins.

Wilde, G. J. S. (1974). Wirkung und Nutzen von Verkehrssicherheitskampagnen: Ergebnisse und Forderungen - ein Überblick. *Zeitschrift für Verkehrssicherheit*, 20, 227-238.

Woods, D. D., Johannesen, L. J., Cook, R. I., & Sarter, N. B. (1994). *Behind Human Error: Cognitive Systems, Computers, and Hindsight*. CSERIAC Program Office.

Woods, D. D. (1986). Paradigms for Intelligent Decision Support. In E.Hollnagel, G. Mancini, & D. D. Woods (Eds.), *Intelligent decision support in process environments* (pp. 153-173). Berlin: Springer.

Woods, D. D. (1988). Commentary: Cognitive Engineering in Complex and Dynamic Worlds. In E.Hollnagel, G. Mancini, & D. D. Woods (Eds.), *Cognitive engineering in complex dynamic worlds* (pp. 115-129). London ; San Diego: Academic Press.