2nd Symposium on Resilience Engineering
Juan-les-Pins, France November 8-10 2006
Resilience Engineering

The dilemma of ill-defining the safety performance of systems if using a non-resilient safety assessment approach [1]

Experiences based on the design of the future Air Traffic Management System

Oliver Straeter [1], Jörg Leonhardt [2], Diana Durrett [2], & Jürgen Hartung [3]

[1] Eurocontrol, DAP-SAF, Brussels, Belgium
oliver.straeter@eurocontrol.int

[2] Deutsche Flugsicherung, Langen, Germany
Joerg.Leonhardt@dfs.de, Diana.Durrett@dfs.de

[3] University of Technology, Munich, Germany
Hartung@lfe.mw.tum.de

Abstract. The European Air Traffic Management system undergoes a dramatic adoption is the next couple of years. Until the year 2020 doubling of traffic, increased automation and increased autonomous operation of aircrafts, and considerable changes of the airspace structure is to be expected. In order to provide a still a safe system in this situation, several concerted activities are initiated which are bound into the ESARR framework (ESARR means European Safety Regulatory Requirements). As part of the ESARR framework, quantitative safety assessment of a planned system change needs to be conducted, which addresses technological, procedural and human contributions to safety performance. However, the power of quantitative risk assessment is depending on the power of the risk assessment methods used. Regarding Human reliability often old fashioned method are used which predetermine misleading safety improvements as they result from a systematical mis-assessment of the human contribution to risk and safety. The resilience principles are used to define a better approach to Human Reliability Assessment (HRA). Note that the paper does not propose to merge Human Reliability and Resilience but to apply the resilience principles to inclusion of Human Reliability Assessment into safety assessments.

1 Introduction

The paper will outline the experiences gained in the current consideration of Human behavior in quantitative safety assessments in the European Air Traffic Management. The experiences, on which this paper is based, were collected in a project for generating a framework for considering human performance in safety assessments.

In the assessment of risk, the conductor is often happy if he has conducted his safety assessment and can come up with a figure describing the expected risk of a system. Having achieved this he is likely to defend his results rather than thinking about the limitations of the risk assessment methods he has used. However, the limitations in the methods obey those areas of risk contribution he has not considered and hence are certainly of future interest, because exactly these contributions receive no countermeasures and therefore likely are the contributions to future accidents.

[1] Note that this paper reflects the opinions of the authors and not necessarily represents the opinion of their affiliation.

In other words, risk assessment is in the dilemma that those contributions to risk not addressed in the assessment are those unmanaged and those potentially leading to future accidents. Most critical is this effect if the Human contribution to risk is wrongly addressed in safety assessments. This is often even worse than not treating the Human contribution to risk at all.

The following chapter will explain this dilemma in detail and the following chapter will then propose a solution by applying the principles of resilience to risk assessment and human reliability in particular.

2 Safety assessment methodologies predetermine the future design of a system

2.1 Classical risk assessment with respect to human contributions

In a classical risk assessment, the Fault Tree / Event Tree approach, Human contributions to risk are assessed by "the human as a system component" with some Human Error Probability attached to it.

The overall reliability (or availability) of the system's risk and effectiveness of its safety functions is then calculated and compared to an expected target level (e.g., calculated as a portion of a TLS – Target Level Safety). It is concluded that the system is safe enough, if the calculated risk is lower than the one set in the expected target level or – vice versa – that the system is not safe enough if the expected target level is exceeded.

Often forgotten in the conclusion are the capabilities of the assessment methods behind the assessment. If the assessment method for instance is not addressing a certain risk contribution, the calculation leads automatically to a lower risk calculated and the system is judged to be safe though it is not.

Obviously, the assessment method, the calculation and the target level cannot be separated from each other in any safety assessment.

In other words, the calculated risk and risk contributions do not necessarily fit with the real risk contributions.

Risk assessments are revealing mitigations for the risk that they have considered in their scope. As a result of this situation risk assessments are in the dilemma that future accidents are also to be seen as a result of missing elements in the risk assessment process. They are a fallacy of risk assessment methodologies. The overconfidence in safety assessment methodologies and misunderstandings in applying them in safety and risk management are determining future accidents.

The classical risk assessment process runs into an arbitrary risk assessment if the limitations of the methods used is not known or not considered in the conclusions from any assessment. This pitfall of the classical accident model can be described starting with epistemic and aleatory uncertainty and then having a look at the accident theory behind.

Epistemic and aleatory uncertainty

In risk assessment, epistemic and aleatory uncertainties are distinguished. The uncertainty between equations or models is known as epistemic uncertainty. Aleatory uncertainty determines those aspects of the system that for some reasons cannot be modeled in the risk model (like actual wind speed for instance).

In a study performed by Theis (2002) it was found that the epistemic uncertainties of a risk model regarding human contributions to risk might easily have the factor of 100, which is much higher than the classical uncertainty due to aleatory uncertainties, which is determined usually with a factor of about 10.

Well known modeling limitations (epistemic uncertainties) in assessing the human risk contributions are (OECD, 2004):

- Lack of consideration of cognitive aspects and decision making under uncertainty
- Organizational influences to risk
- Interdependencies between different human actions
- Human interventions with adverse effects on safety (e.g., due to human-automation interdependencies)

Despite these epistemic uncertainties in human reliability assessment (HRA), quantitative risk assessment can be used and reveals results for the limited scope of the risk model.

However, care needs to be taken in generating conclusions from risk models due to the epistemic uncertainties. In the order to take this duty of care into account, disclaimers are well established in nuclear risk assessments for instance. Unfortunately they are not in risk assessments in ATM (Air Traffic Management). The shortfalls of safety assessment methodologies will become unknown shortfalls of the safety performance of a future system.

2.2 The pitfall of predefined safety improvements by wrong risk assessment

Due to unconsidered epistemic uncertainty, conclusions on system design are ill informed about real safety matters.

As an example heavily disputed in ATM (as in nuclear and other industries decades ago as well) automation is seen as a mean to overcome the (seemingly) non-reliance on human performance in standard system functions, justified by the lack of human safety performance in manual control of the standard system functions. However this rationale considers only a limited system scope comparing the automated system with the human in a very particular situation and by only looking at the negative side of human performance (the errors one can make). On the other hand human positive performance (recovery) or additional influences from other parts of the system (e.g., maintenance or management) are neglected. Consequently automation is seen as superior compared to human (e.g., if expressed by errors per flight hour an automated system reveals a 10 E-8

vs. a 10 E-2 for human actions). Despite the fact that this comparison is based more on the restrictions of the safety assessment approach and limited scope rather than on a thorough comparison, as a consequence of the quantitative figures it will be decided that the system function will be automated. As a long-term consequence of this decision, the capabilities to cope with variations of situations are neglected, important risk contributions from the organization are overlooked and unmanaged, and the humans' positive features are lost.

The following figure describes this problem in more detail. Figure 1-A shows the classical thinking. There is an option to automate a certain system function that is done by a Human so far - with a strong argument that herewith the overall system safety becomes much safer. The reason is that the reliability of the automated system is 100 times better than of the human in this particular function. A risk assessment of the functional chain is conducted and proves this generic thinking. However it is already to notice that the expected gain from automation is not factor 100 but only factor 34. As there are three functions determining the overall system, the gain is only 1/3 of what was expected. In a system with *n* different functions the expected gain is only *1/n* of what is expected from the particular gain in one function (assuming equal reliability of all functions). As a general rule, the expected gain is lesser the higher the complexity of the entire system is. However any system designer would decide for Option B as this seems to be the more safe option if only having the limited scope in mind.

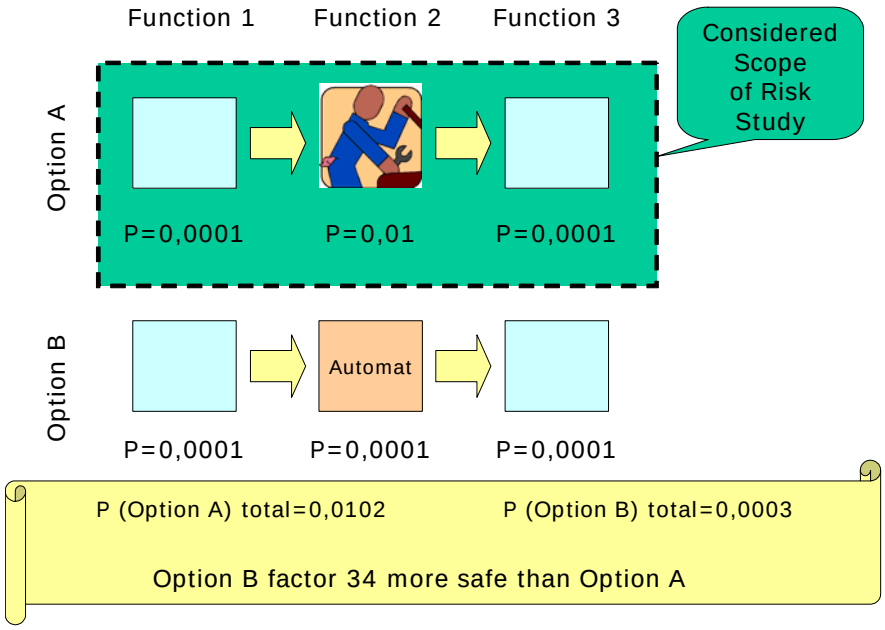
The picture gets completely different if the scope of the study is changed towards all risk contributions determining the overall system reliability. Essential contributions to consider are at least maintenance and recovery. Figure 1-B shows these included into the scene. If, for simplicity reasons, one assumes maintenance errors are at 0,01 and need to be performed every 100th time of system operation (e.g., every 100^{th} day), the probability of failure of the automated system suddenly doubles to 0,0002 (both contributions are in OR relationship). Also the other functions do need some maintenance with the same effect, which was not considered in the original scope of the assessment of Option A and B.

Most critical is however the human contribution to recovery. Assuming the human that was present in Option A had the opportunity to recover from unexpected events every 100^{th} time by using his creativity and intuition, suddenly the overall reliability of the system becomes factor 6 more reliable if the Human stays in the system rather than having a fully automated version.

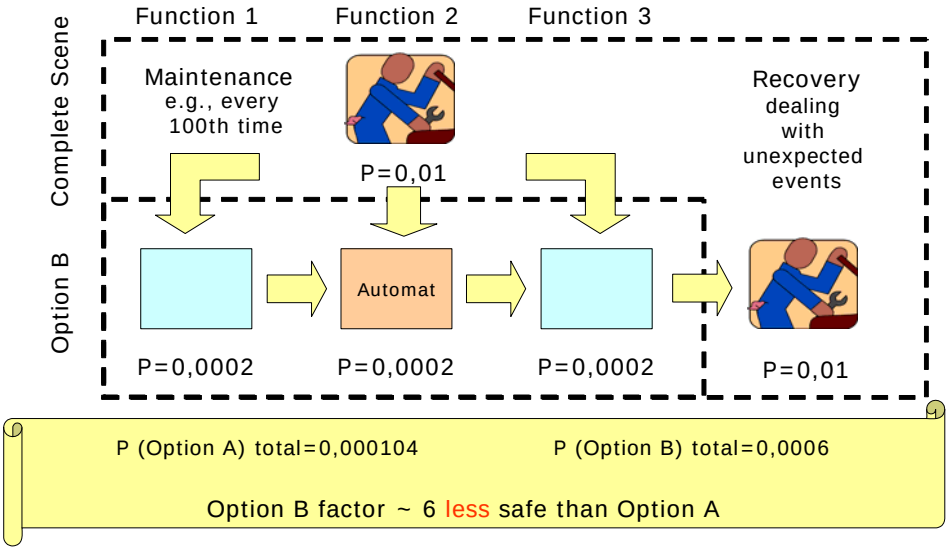
Just by changing the scope of the study, the conclusion turns upside down. The clear scoping before assessment starts is hence an important requirement of any risk assessment.

Even if one decides to combine Option B (full automation) with a human recovery, this recovery might be less effective, as the human is out of the loop and in stand by mode. It is know that this is placing the Human in a less effective condition (Amalberti, 2001). Such a design might then lead to a Human recovery potential of let us say P=0,1 for the given system. Overall the system unavailability of Option B' [fully automated system with human recovery] is then (P= 0,00006) and less than factor 2 better than the Option

A (P = 0,000104). This result is completely covered by the uncertainties in assessments (usually factor 3 to 10) and hence not precise enough to justify a system change.



A) The scope defined by the function to automate



B) The scope defined by the risk contribution to the overall system

Figure 1: The effect of incomplete Scope regarding HRA on the decision for safety functions

Conclusions on system designs need to be carefully reflected on the scope and the validity of the scope in a real setting. The scope determines the outcome and the decisions taken.

Unfortunately most of the risk assessments go for defining the scope by the function to automate and neglect the real scope required. As a consequence of risk assessment, functions are automated and the system becomes not necessarily safer or even unsafe.

2.3 The role of the accident philosophy in determining future accidents

Another aspect is the barrier thinking as the underlying safety philosophy of safety design and risk assessment. In the classical view, barriers are filtering out unwanted behavior. The effectiveness of barriers is then related to the number and size of the "holes" left over in the "Swiss cheese".

The idea of filtering predetermines the measures and means for improving safety. A barrier is most effective if filtering unwanted behavior most effectively. Regarding the Human aspect, this means certain behavior is prevented to occur by system design and best if completely abandoned.

This current thinking of accident development has different "faces" as outlined in Figure 2. The classical barrier concept as depicted in the left side of figure is also reflected in the accident pyramid represented on the right of the figure.

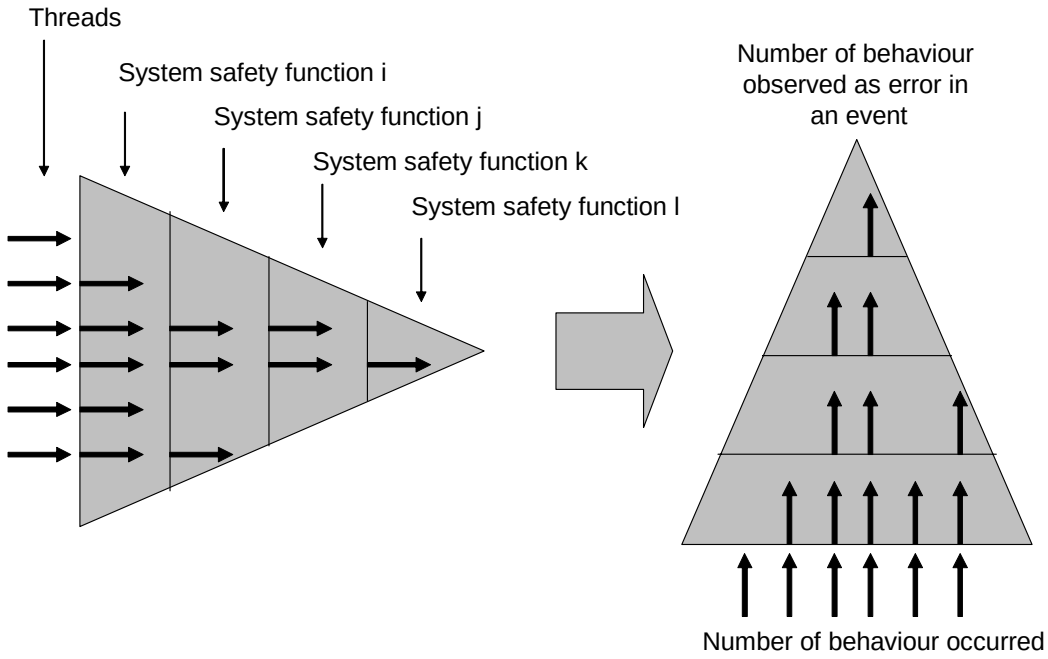


Figure 2: The relationship of accident barrier thinking and the "accident pyramid"

Preventing a particular behavior is rather a way to reduce uncertainty, but not really a mean to increase safety performance.

First a barrier *i* is compensating for deficiencies of the preceding barriers *1..i-1*. However, there could be many different, latent deficiencies that are not addressed by the barrier *i*.

Second, more important is that the way the barrier itself is built defines the human behavior induced. As an example the idea of the operators during the Chernobyl event to actively intervene into the reactor protection system was a result of the way how the barrier "reactor protection system" is built. First the barrier is designed to prevent Humans from intervening but, on the other hand, in certain beyond design bases accidents, they are even requested to intervene. The same non-rationale approach one can find in the TCAS (Tactical Collision Avoidance System) in aviation, which a pilot is supposed to follow; however this system has a false alarm rate of about 8% so that any pilot also has good reasons to be suspicious in following it.

As a result, barriers are only "real" barriers in very particular situation, the so-called design-bases situations but they are not (or less) effective in so-called beyond-design-bases situations. This "double-moral" of barrier thinking always pushes the human in the mode to creatively think of how to potentially overcome a barrier in case of beyond-design-bases situations. The barrier hence defines the active behavior possibly shown by humans under a certain context. If such active behavior leads to unwanted results they are called errors of commission, though the cause for the error of commission is the ill-defined barrier.

This problem holds for all systems where barriers or automated systems are not working with a proper reliability. Accidents hence rather show the deficiencies in the accident theory and risk assessment and than human deficiencies.

3 The use of resilience principles to overcome the pitfalls in risk assessment

3.1 Some prove for a resilient accident theory

The Human Error Probability (HEP) is defined as number of the human errors observed divided by those actions where the same behavior occurred but did not led to an error. If the barrier model was correct, the life of a system safety designer would be relatively easy regarding human contributions to risk. All data required would be available by incident investigation.

The barriers would filter a certain portion of the not required behavior but would also filter, with the same effect, the same behavior if it is not leading to an error. The filter filters behavior, not only errors.

For instance an operator missed to monitor correctly an alarm and this lead to an event the number of errors of this type reported can easily be counted. In the same event (or in other events) an operator might have successfully monitored an alarm and successfully prevented an event. Also this behavior can be counted if reported (i.e., if above the event reporting threshold). Hence events contain unsuccessful as well as successful human interventions. As an example, Table 1 shows the distribution of events from the

German Nuclear in the time of 1965 to 1996 with successful as well as unsuccessful human behavior (Straeter, 1997).

Table 1: Distribution of successful as well as unsuccessful human behavior

	%	Number
Human Performance Events (successful intervention)	0,82	3570
Human Error Events (unsuccessful intervention)	0,18	808
Sum		4378

As one can see from the table, there is a big portion of human well functioning that is not covered by human error analysis. However, such events are normally not investigated because everything was correct. But also the events with human errors contain, if properly analyzed, information about the positive human behavior in events. This was done in the investigation performed in the context of the development of the CAHR method (Straeter 1997/2000). In the investigation of 232 events in total, 98 human errors in 439 sub-events were identified. This equals a ratio of about 22% of unsuccessful to successful human behavior.

Taking these statistics and the accident pyramid assumption into consideration, we would not have any problem of generating human error probabilities. There would be a linear relationship between the number of errors and opportunities as the barriers filter out the behavior with a fixed rate. Figure 3 represents this thinking.

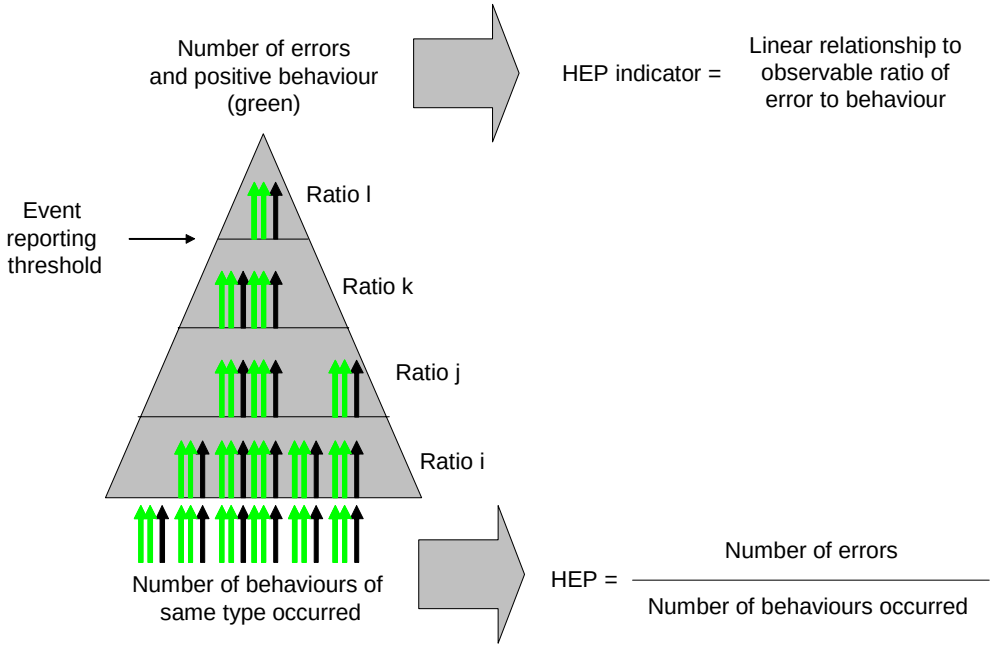


Figure 3: The relationship of accident barrier thinking and the quantification of Human error probabilities

Though incident data contains positive information about human performance and can be used as an indicator for Human Reliability, the same investigation of the 232 events could not find any linear relationship. Indeed a probabilistic relationship was found between the observed ratio of errors to performance and HEPs as depicted in Figure 5.

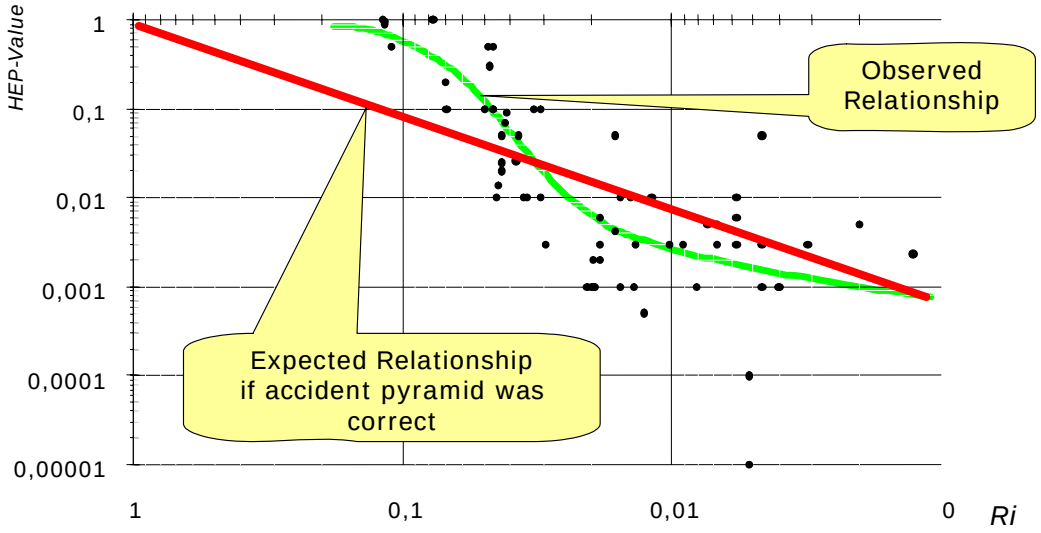


Figure 5: Observed ratio of errors to performance and HEP in the development of the CAHR method

Based on similar experiences many disputes have been taken place on the sense and non-sense of the calculation of the Human Error Probability (Hollnagel, 1999; Amalberti, 2001; Straeter, 2005).

Summarizing, the classical HEP thinking is also a result of the accident philosophy but not necessarily the right approach to quantification of human contributions to risk. Important to note is that the barrier thinking is not only reflected in the structure of risk assessment but also in the quantification.

For Human Reliability Assessment the thinking in a classical definition of the Human Error Probability (HEP) needs to be reconsidered. Though the classical definition makes perfectly sense in the thinking of the classical risk assessment approach, the empirical results suggest to overcome this classical thinking. The conclusion would be to think about the sense and (non)-sense of the barrier thinking.

3.2 A proof of (non)-sense of the barrier thinking

Treating the human as what he is, a part of the system performing a task with some uncertainty, not behaving deterministically (as a designer wants), basically means he/she is showing a distributed, uncertain behavior instead of a stable behavior that can be blocked by barriers.

This uncertain behavior can be understood as a cybernetic problem of general nature and there are sciences having dealt with such problems already.

Physics went through a similar renewal of understanding in the times of the relativity theory about 100 years ago. Physics knows the effect of such uncertain behavior in quantum physics as interference. The effect investigated was simple. As presented in Figure 6, the behavior of an electron changes after passing a barrier. Expected would be to have no existence of electrons in the shadow area, but indeed there are some and these are building an interference with other electrons. Obviously the electron has two types of nature, is behaves like a particle until it reaches the barrier but then behaves like a wave after having passed it.

Any electron needs to be seen as a particle but also at the same time as a wave. This is called the wave/particle dualism. Whether we assume an electron to be a particle or a wave is just reflecting our limitation to understand the physical problem completely rather than a real difference.

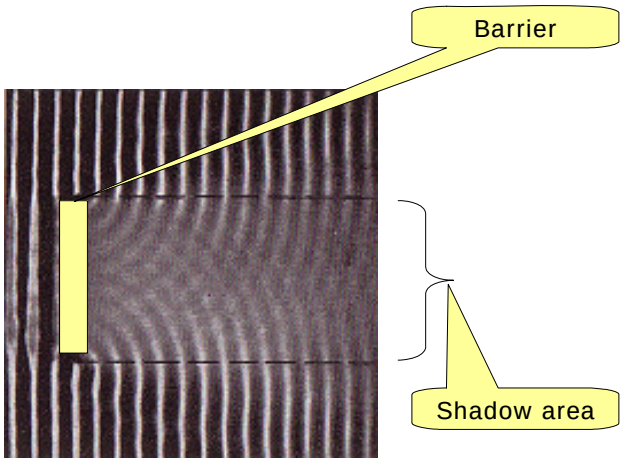


Figure 6: Wave/particle dualism.

The effect outlined in the figure can, without any further consideration, be transferred to the effects observed in accidents. A barrier does not lead to the expected constraint one might like to have with a barrier but to a distribution of behavior after the barrier passed.

Based on these considerations from Physics, the fallacy in human thinking that barriers can block human behavior can be seen in the same way. We expect the human to behave in exactly the same manner all the time but must see that the behavior is also actively changed by the barrier.

The result of applying the cybernetic view to the human behavior is represented in Figure 7.

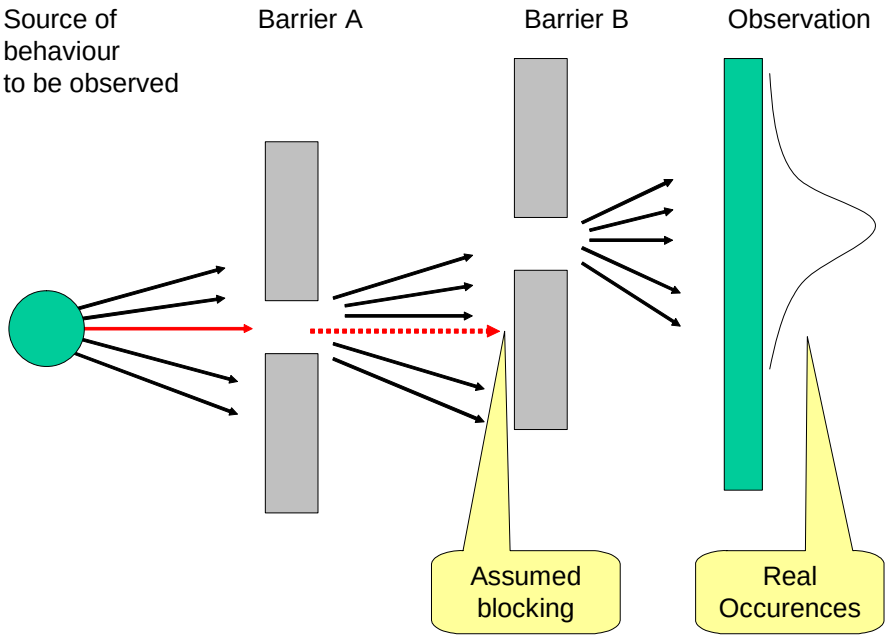


Figure 7: The fallacy in human thinking that barriers block.

The distribution of the real observation is fitting exactly to the probabilistic curve represented in Figure 5 above and explains the observation made regarding the probabilistic curve found during the development of the CAHR method.

This cybernetic perspective concludes that the classical barrier thinking and HEP calculation is not fitting to empirical findings as well as cybernetic considerations on human behavior.

The approach to optimize the barriers' safety performance by filtering may lead to the effect that even small deviations might break the system because the need for human variability is increased and this variability increases the potential of a break through succeeding barriers.

The effect is that the assessment procedure is leading to a superficial or ostensible feeling of being safe (because the risk model tells so). A beyond barrier focused assessment is needed in which the risk assessment is not underestimating the real risk because it is only looking at the risk related to the barrier's scope.

3.2 A resilient HRA approach

In the light of the cybernetic observations and the wave particle dualism presented in the previous section, the resilience thinking can be understood as bringing an old well established experience (the dualism of behavior) into accident thinking; and this makes sense by all means, from the physical, mathematical and psychological perspective.

A resilient approach for HRA starts with an optimization regarding specific system states under full consideration of deviating scenarios and by taking the variation of hu-

man performance in its context into account. As outlined in Figure 8, this variation is essential to have a proper assessment of the human contribution to risk.

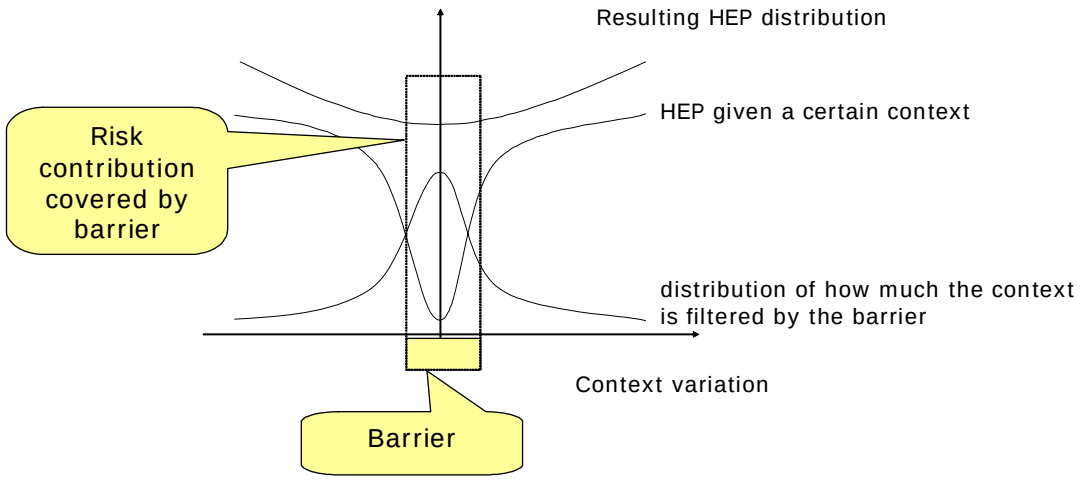


Figure 8: The fallacy in human thinking that barriers result into unconsidered higher risk contribution of the humans under a given barrier.

As a general rule the barrier is designed regarding a certain nominal behavior but human performance deviates based on the context, not only on the barriers scope. The resulting distribution for those elements of the HEP distribution breaking through a barrier is likely a result of the outer areas of the overall distribution, which is more context driven. Though the context is less likely, the HEP for the given context is much higher as for the nominal case and the resulting overall HEP distribution is even higher for the outer areas of the distribution.

For designing a system optimally according to all contextual conditions, the human behavior needs to be simulated according to all conditions in a systematical way and to achieve an assessment that is going beyond a barrier focused assessment. Uncertainties in human behavior, in contextual conditions and in failure conditions should be modeled as well.

Such a simulation needs first of all a more sophisticated safety assessment approach going beyond Fault Tree / Event Tree. Known solutions are Bayesian networks or dynamic risk assessment approaches. Such approaches consider the variation in Human performance rather than delivering single point values for human performance in a nominal case.

In actual use for showing ESARR 4 compliance are Bayesian networks as used by DFS. Therefore the work on Human Reliability was performed with DFS as this is the approach in practical use going beyond classical assessment. More research based approaches are dynamic risk models. Such an approach was undertaken in a project for dynamic risk modeling (DRM) at Eurocontrol (see Leva et al., 2006).

The DRM research project was aimed at developing a simulation approach able to provide a quantitative analysis of some critical activities of Air Traffic Control (ATC) operators considering the organizational context in which they take place and the main cognitive processes underneath. The process was able to provide a trial application in a specific case study in the ATM context.

This approach within the field of HRA (Human Reliability Analysis) is able to interact with standard risk assessment methodologies in order to "foresee" the possible criticalities arising from human performance in the ATC working contexts. Indeed, the simulator that has been used (named PROCOS), tries to integrate the quantification capabilities of human reliability assessment methods with a cognitive evaluation of the operator, by means of a "semi static approach".

The pilot study was aimed at providing an overview of possible opportunities related to the use of a cognitive simulator within CONOPS and evaluating the potential use of HERA Predict in combination with PROCOS for concept evaluation (e.g., by analyzing the contributing factors to human error observed in incidents, or by making use of experiences of approaches developed in other industries like the CAHR method).

The approach allows dealing with classical barriers and with uncertainties of human aspects and contextual influences in a homogeneous approach. The probabilistic quantification of the CAHR approach was used to conduct the quantitative risk assessment.

The simulation of human behavior at a range of uncertainty and uncertainty of the failure scenarios takes into account that behavior may differ probabilistically not deterministically. The probabilistic task modeling rather than deterministic task modeling allows overall a Human behavior modeling going beyond the modeling of the task required according to a certain barrier.

4 Conclusions

HRA needs to steer rather than adapt technical barrier thinking. Classical Fault Tree / Event Tree thinking, another derivate of the accident barrier philosophy needs to be substituted by simulation in order to achieve this.

The paper suggests conducting safety assessments according to the principle of resilience. This implies to have a holistic system view and to perform safety assessment with full considerations of the limitations of the quantitative safety assessment approach chosen, which also includes using more advanced safety assessment approaches (like dynamic risk modeling) with full consideration of the uncertainties in the safety assessment. This also proposes a resilient approach to considering human behavior consisting of a safety assessment starting the safety analyses with modeling the human behavior rather than the technical behavior of the system, with fully aligning assessment and operational experience, including all levels human involvement in the system and with not treating human behavior as an appendix to a technical safety analysis.

The thoughts generated in this paper also can give hints to include resilience into classical safety assessments, namely:

- The scope of application of the risk assessment method to the suggested safety-improvement needs to be clearly described.
- No risk target can be seen independent from the methods used for assessing risk.
- No severity classification can be seen independent from the methods used for assessing risk.
- No risk assessment can be used for deriving conclusions without stating clearly the limitation of the approach.
- Any quantitative figure used in system safety design needs to have uncertainties if it should be of any value.

The paper made clear that the use of certain safety assessment methods (like the classical fault tree / event tree approach) as currently transferred to ATM and a recommended practice for conducting quantitative safety assessments is of less value. While this approach works well for systems with low interdependency of the system elements and with little human involvement, it is also known since decades that this approach has severe limitations in respect to interdependent systems with many human involvements. Unfortunately the ATM system is of latter kind, which implies that the safety assessment approach does not fit the safety features to be modeled in the system. On the other hand, recommendations for system design will be drawn from quantitative safety assessments, like automating certain system functions for instance (even if not appropriate). Applying fault tree / event tree in an unconsidered manner regarding the pitfalls consequently predetermines how a future system will be designed and how resilient the future system will be in respect of safety, in particular regarding interdependencies and human performance.

The safety assessment methodologies for potential future systems need to have a resilient safety assessment approach instead.

References

Amalberti, R. (2001). The paradoxes of almost totally safe transportation systems. Safety Science 37: 109-126.

Hollnagel, E. (1999) Looking For Errors Of Omission And Commission Or The Hunting Of The Snark Revisited. Reliability Engineering and System Safety. Elsevier.

Leva, C., De Ambroggi, M., Grippa, D., De Garis, R., Trucco, P. & Sträter, O. (2006) Quantitative analysis of ATM safety issues by means of Dynamic Risk Modelling (DRM). Eurocontrol Safety R&D Seminar. Barcelona.

OECD-CSNI (2004) Technical Opinion Papers No. 4 – Human Reliability Analysis in Probabilistic Safety Aseessment for Nuclear Power Plants. OECD NEA No. 5068. OECD-NEA. Paris. (ISBN 92-64-02157-4)

Sträter, O. (1997) Beurteilung der menschlichen Zuverlässigkeit auf der Basis von Betriebserfahrung. GRS-138. GRS. Köln/Germany. (ISBN 3-923875-95-9)

Sträter, O. (2000) Evaluation of Human Reliability on the Basis of Operational Experience. GRS-170. GRS. Köln/Germany. (ISBN 3-931995-37-2)

Sträter, O. (2005) Cognition and safety - An Integrated Approach to Systems Design and Performance Assessment. Ashgate. Aldershot. (ISBN 0754643255)

Theis, I. (2002) Das Steer-by-Wire System im Kraftfahrzeug – Analyse der menschlichen Zuverlässigkeit. Shaker. München.