From protection to resilience: Changing views on how to achieve safety

Erik Hollnagel

From protection to resilience: Changing views on how to achieve safety

Erik Hollnagel
Ecole des Mines de Paris, CRC, Sophia Antipolis, France

Abstract

Effective safety management requires the ability to learn from the past and to anticipate the future. Yet what we can learn from the past (i.e., accident investigation) and what we can imagine for the future (i.e., risk assessment) depends critically on how we think about it, i.e., the models and methods we have at our disposal. Accident investigations have long been dominated by a search for causes, either as root causes or human errors. Risk assessment has similarly been dominated by static representations such as event and fault trees. In both cases the commonly used models and methods have reached their limits because the reality of our self-created socio-technical environments has become too complex. The alternative is to understand how the variability of human actions is a resource rather than a threat and to define safety as a system's resilience, its ability to adapt and adjust, rather than as the absence of adverse outcomes.

Introduction

Even in the best of all possible worlds, the future is not completely predictable. Events are bound to occur for which we are not prepared, some with positive and some with negative outcomes. Although there are very few situations where things go wrong compared to the very many where things work out fine and where the outcomes are as intended – or at least acceptable under the circumstances – the positive cases tend on the whole to go unnoticed. When the outcome of a task or an activity is acceptable, there is little motivation to look for why that was so; it is simply taken for granted – and even considered normal – that things go right. Conversely, when something goes wrong a relentless hunt for the cause(s) begins, in order to ensure that such an event never happens again.

Unless we are willing to treat adversity with Panglossian optimism we must, of course, find some way to reduce the uncertainty, especially with regard to things that can go wrong. One venue for that is to design processes, systems, and organisations such that hazards are eliminated, or to ensure that the probability of adverse events is reduced to an acceptable level. (A hazard is here defined as an event that can lead to a known loss.) In order to do so, it is necessary that the system can be described in detail and that events develop in a predictable manner. However, since a growing number of socio-technical systems are intractable, it is in practice impossible to achieve an acceptable level of safety by precautionary measures alone, i.e., by eliminating hazards, by preventing unexpected event, or by protecting against unwanted outcomes. Safety by design (analytical safety) must therefore be complemented by safety by management (operational safety). As resilience engineering makes clear, safety is something a system *does*, rather than something a system *has* (Hollnagel, Woods & Leveson, 2006).

Safety by design

The meaning of "safety by design" in this paper is that all possible, or practicable, precautions needed to ensure an acceptable level of safety are taken ahead of time. This could be either when the system is conceived of or designed, when detailed plans for operation are made, or when it is made ready for operation. The engineering parts of the system, namely the technology or hardware, are designed in detail and must perforce be configured, operated, and maintained according to meticulously prepared instructions. In such cases "safety by design" is not only an option but a requirement. Even so, there are limitations to what this can accomplish because technology always is a means to an end rather than an end in itself:

> "Clearly not all accidents can be prevented by design. There are some consequences of technology that cannot reasonably be predicted at the design stage, particularly in new technologies, using new materials and scientific principles. However, once these have led to accidents, there is a clear responsibility for designers to prevent them in future designs."
> Hale, Kirwan, Kjellen (2007, p. 310)

Yet even if the technology is reliable, and even if it functions in a highly predictable manner, there are other problems. For instance, have the safety requirements for the detailed design been adequately defined? Are the selected concepts proven from a safety point of view? Or will will it be possible to implement regulatory, corporate and customer safety requirements within acceptable cost limits? Although these questions address issues that have to do with the use of the system in the wider operational context, they are legitimate for "safety by design" as well.

The non-technical parts of the system, namely the people or the liveware, are far more difficult to design. Indeed, an organisation can never be explicitly designed in the same way that a piece of machinery can, one fundamental reason being that the "components" by their nature are variable and flexible, regardless of whether they are considered individually or collectively. Humans and social systems simply do not function like machines, despite courageous attempts to make them do so through training, interaction design, and automation. For the non-technical parts the alternative is to focus on the operation of the system, and on how it is possible to keep the variability within acceptable limits, i.e., by managing safety.

Safety by management

Safety management has during the last decades of the 20th century become a major issue in itself, with a number of commercial solutions for safety management systems (SMS) on offer. It may therefore be sensible to look beyond the use of SMS as a glib phrase and try to understand what it entails. The crucial point is here the definition of safety, i.e., the definition of what the SMS is supposed to manage.

Safety management is, from a practical point of view, a kind of control; indeed, the meaning of 'manage' is 'to exert control over, to direct.' Safety management can therefore be interpreted as the control of the organisational functions and practices that together produce safety. The purposes of an SMS is to ensure that the organisation's "safety processes" develop in the indended direction and that they are not disrupted or hindered by internal or external events and conditions. This reformulation leads to three subsidiary questions. The first is what the organisational "safety processes" really are, i.e., what it is that "produces" safety. The second is how these processes can be controlled in practice, i.e., how their "speed" and "direction" can be changed. And the third is how the outcome or the result can be measured, i.e., what the proper indicators of safety are.

As far as the latter is concerned, it is common practice to associate safety with the "freedom of unacceptable risk," or as ICAO (2006) defines it:

> "Safety is the state in which the risk of harm to persons or of property damage is reduced to, and maintained at or below, an acceptable level through a continuing process of hazard identification and risk management."

By defining safety in this way, a safe state is defined by the absence of something and the measures or indicators are therefore the, hopefully, diminishing numbers of negative events. The established practice is to try to reach this state through a tripartite approach: to eliminate hazards (by design), to prevent initiating events that may lead to adverse outcomes (by constraining operations), and to protect against adverse outcomes (by introducing barriers).

This approach to safety implies a distinction between normal and abnormal operations. The normal operations ensure that the (safety) process goes in the right direction and that it produces what it was supposed to produce or achieve. The abnormal, or off-normal, operations disrupt or disturb the normal operations or otherwise render them ineffective. The main motivation for safety management, as it is commonly practiced, has therefore been to prevent such disruptions or disturbances from taking place. In that sense safety management has been driven by what has happened in the past, hence been mainly reactive.

Theory W

One way of characterising the established approaches to safety and safety management is to propose two idealised positions, expressed as two different theories called Theory W and Theory Z, respectively. According to Theory W, socio-technical systems are safe and efficient because:

- The systems are well-designed and scrupulously maintained.
- The procedures that are provided to operate the systems are complete and correct.
- The operators of the systems, the people at the so-called sharp end, behave as they are expected to, and as they have been trained to.
- Designers can foresee every contingency and therefore provide the systems with appropriate response capabilities.

Theory W describes a world of well-designed, well-tested, and well-behaved systems. It is a characteristic of such systems that there is a high degree of reliability of equipment; that workers and managers are vigilant in their testing, observations, use of procedures, and operations; that staff is well trained; that management is enlightened, and that good operating procedures are in place. The threat to normal performance comes from different types of failures or malfunctions, such as active failures and latent conditions, equipment faults, and human error. Safety can therefore be achieved by constraining performance variability in various ways (Figure 1).

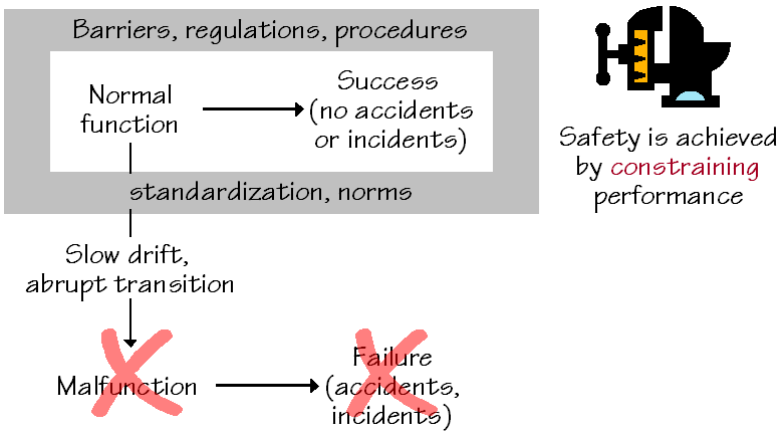

Figure 1. Safety according to Theory W

It follows from the assumptions behind Theory W that performance variability of any kind should be avoided. Technological systems are designed to perform one – or a few functions – very efficiently and with limited variability. Functions that depend on humans, whether as individuals, as groups, or as organisations, are more versatile but also less uniform and the variability is seen as a threat or a disturbance that may end in performance failures. Everything possible is therefore done to prevent this from happening. The change from normal to abnormal operations can happen either abruptly, as when something breaks, or gradually in the form of a drift or slow misalignment (Cook & Rasmussen, 2005). The solution according to Theory W is therefore to constrain performance, for instance, by means of barriers, interlocks, rules, procedures, standardization, interaction design, norms, etc., or even by replacing humans by technology, as in the use of automation. This is supplemented by more wide-ranging activities such as risk management (comprising, e.g., hazard identification, risk assessment, risk mitigation, and risk communication), hazard and incident reporting, safety investigations, safety analyses and safety studies, and safety performance monitoring.

Safety for underspecified systems

Theory W corresponds to systems where the principles of functioning are clear and where there is good predictability. This was not an unreasonable assumption at the time when the currently established methods were developed, roughly between 1965 and 1985, but it is much less reasonable today. There are two main reasons for that, first that all systems of interest are more or less intractable, and second that performance variability is inevitable.

Intractable Systems

In order for a system to be tractable, four conditions must be fulfilled: (1) that the principles of functioning are known, (2) that a description does not contain too many details, (3) that a description can be made relatively quickly, and (4) that the system does not change while the description is being made. The last condition is the most important one, and is in a way a synthesis of the three preceding.

Theory W clearly requires that systems are tractable. Many of the present day systems of major interest for industrial safety are unfortunately intractable rather than tractable. This means that the principles of functioning are only partly known, that the description is elaborate and contains many details, that it takes a long time to make, and that the system therefore change while the description is made. In consequence of that it will be impossible to provide a complete description or specification of the system. Intractable systems are underspecified in the sense that details may be missing or unavailable (e.g., Clarke, 2000). But if a system is underspecified it is clearly not possible to provide precise procedures or instructions. The people working in the system, be it at the sharp end or at the blunt end, must therefore be able to apply the available prescriptions and procedures to conditions and situations that differ from what was assumed. In other words, it is necessary that people are able to vary or adapt what they do to ensure that the system functions as required and achieves its operational goals. Performance variability – whether it is called improvisation, adaptation, efficiency-thoroughness trade-off, sacrificing decisions, or creativity – is therefore necessary, hence an asset rather than a threat.

The inevitability of performance variability

While machines and technological artefacts are designed, built, and maintained so that they can produce a near constant performance – at least until they fail and must be replaced – the same is not the case for humans and for organisations. There are many reasons why human performance never can be constant or machine-like. One is that physiological functions (muscles, nerve cells, sensory organs, etc.) are subject to fatigue, saturation, and accommodation, and that they regularly require a period of rest or reconstitution. Many basic psychological functions, such as attention or vigilance, are also limited with regard to how long they can be maintained at a constant level. A second reason is that humans seem to have an innate tendency to vary what they do, often to avoid monotony or to find an easier way to accomplish a task. It is this ingenuity and creativity that is at the heart of adaptability and of the ability to overcome constraints and underspecification. A third reason is socially induced variability, for instance in the sense that others will have expectations – or even informal norms – about how much and how little effort is acceptable in a given situation. A different source of variability is the organisational culture, specifically the safety culture. A fourth reason is that performance depends on the state of the organisation and environment, which may vary in terms of demands to work, resources available, etc. A fifth and final reason is the variability due to ambient working conditions, for instance temperature, climate, humidity, noise, etc.

Performance variability is for these reasons inevitable, whether on the level of the individual or the organisation. This is fortunately not a disadvantage since performance variability is necessary to overcome the underspecification of large socio-technical systems. But it is a problem if safety models and methods fail to recognise that.

Theory Z

Since most systems of interest today are intractable, it is impossible to provide a complete description of them or to specify what an operator should do even for many normally occurring situations. This is recognised by Theory Z, according to which socio-technical systems are safe and efficient for the following reasons:

- Humans learn to overcome the inevitable shortcomings, such as design flaws and functional glitches.
- Humans can adjust their performance to meet the actual demands of a situation.
- Humans can interpret procedures and apply them to suit actual conditions.
- Humans can detect when somethings fails or goes wrong, and can in many cases correct for it as well.

In Theory Z, performance variability is both normal and necessary and is the source of positive and negative outcomes – successes and failures – alike. Failures can consequently not be prevented by eliminating performance variability, i.e., safety cannot be managed by constraints. The solution is instead to identify the situations where normal performance variability may combine to create unwanted effects and to monitor continuously how the system functions in order to intervene and "dampen" performance variability that threatens to get out of control, cf. Figure 2. Conversely, performance variability should be accentuated or amplified when it can lead to successful outcomes. Thus rather than looking for ways in which something can fail or malfunction, we should try to understand the characteristics of normal performance variability, and specifically how internal and external factors may affect the size and nature of the variability.

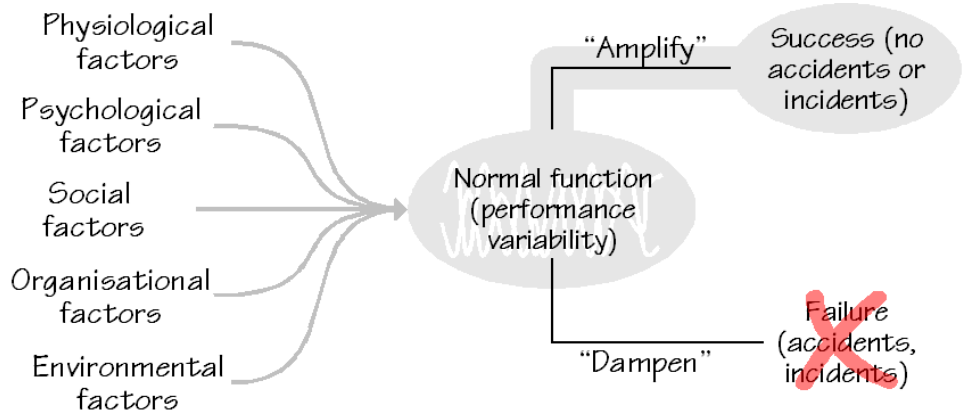

Figure 2: Safety according to Theory Z

Safety by management: Resilience engineering

The difference between Theory W and Theory Z corresponds to the difference between a reactive and a proactive approach to safety management. If performance variability is both normal and necessary, safety must be achieved by managing performance variability rather than by constraining it. This is consistent with resilience engineering, which is based on the following principles (Hollnagel, Woods, & Leveson, 2006).

- Performance conditions are always underspecified, as argued above, and individuals and organisations must therefore adjust their performance to the current conditions. Because resources and time are finite, such adjustments will inevitably be approximate.
- For tractable systems, most adverse events can be attributed to a breakdown or malfunctioning of components and normal system functions. For intractable systems, most adverse events cannot. They are instead best understood as the result of unexpected combinations of normal performance variability or as the converse of the adaptations necessary to cope with real-world complexity.
- Effective safety management cannot be based on hindsight, nor rely on error tabulation and the calculation of failure probabilities. Safety management cannot be only reactive but must also be proactive. Resilience Engineering looks for ways to enhance the ability of organisations to create processes that are robust yet flexible, to monitor and revise risk models, and to use resources proactively in the face of disruptions or ongoing production and economic pressures.

A resilient system is defined by its ability to adjust its functioning prior to, during, or following changes and disturbances so that it can go on working even after a major mishap or in the presence of continuous stress. A resilient system accepts a constant sense of unease and remains sensitive to the possibility of failure (Hollnagel, Nemeth & Dekker, 2008). The quality of resilience can be defined more precisely by pointing to four essential qualities or abilities that a system or an organisation must have, cf. Figure 3.

- A resilient system must be able to respond to regular and irregular threats in a robust, yet flexible, manner. It is not enough to have a ready-made set of responses at hand, since actual situations often do not match the expected situations – the only possible exceptions being routine normal operation.

The organisation must be able to apply the prepared response such that it matches the current conditions both in terms of needs and in terms of resources. In terms of the three types of threats proposed by Westrum (2006), this is the ability to deal with regular threats. The responses enables the organisation to cope with the *actual*.

- A resilient system must be able flexibly to monitor what is going on, including its own performance. The flexibility means that the basis for monitoring must be assessed from time to time, to avoid being trapped by routine and habits. The monitoring enables the system to cope with that which is, or could become, *critical* in the near term.
- A resilient system must be able to anticipate disruptions, pressures, and their consequences. This means the ability to look beyond the current situation and the near future, and to consider what may happened in the medium- to long-term. In terms of the three types of threats proposed by Westrum (op. cit.), this is the ability to deal with the irregular threats, possibly even the unexampled events. The anticipation enables the system to cope with the *potential*.
- Finally, a resilient system must be able to learn from experience. This sounds rather straightforward, but a concrete solution requires consideration of which data to learn from, when to learn, and how the learning should show itself in the organisation – as changes to procedures, changes to roles and functions, or changes to the organisation itself. The learning enables the organisation to cope with the *factual*.

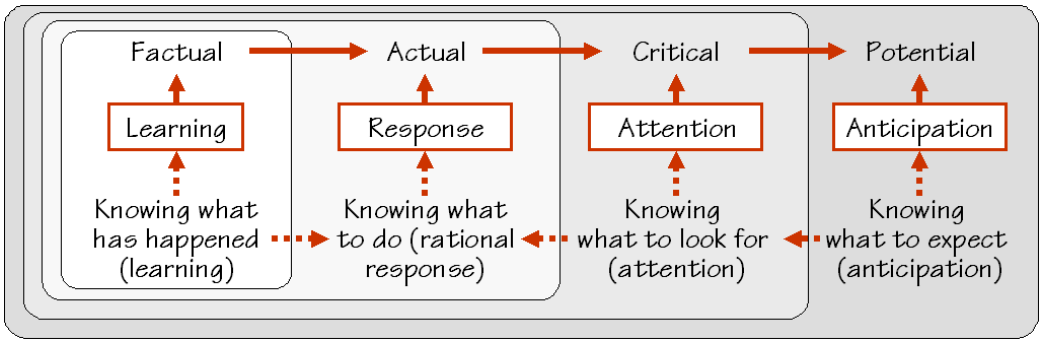

Figure 3: Principles of resilience engineering

Conclusions

Effective safety management requires the ability to learn from the past and to anticipate the future. Yet what we can learn from the past (i.e., accident investigation) and what we can imagine for the future (i.e., risk assessment) depends critically on how we think about it and on the models and methods we have at our disposal. Accident investigations have long been dominated by a search for causes, either as root causes or human errors. Risk assessment has similarly been dominated by representations of linear combinations of a small number of events, such as event and fault trees. In both cases the underlying assumption is that systems – and events – are tractable.

It makes sense that models and method would be just about adequate for the typical type of problems at the time they were developed. Indeed, there would be little reason to develop a method that was more complex or more powerful than required, not least because it would be difficult to imagine what that should comprise. New models and methods are developed because existing models and methods sooner or later encounter problems for which they are inefficient or inadequate. This, in turn, happens because the socio-technical systems where accidents happen continue to develop and to become more complex and more tightly coupled. The inevitable result is that any method after a while becomes underpowered because the nature of the problems change, although it may have been perfectly adequate for the problems it was developed for in the first place.

The risks that dominate in present day systems have a different aetiology than the risks that dominated even one or two decades ago. This has two important ramifications. The first is that it is more difficult to understand present day risks – at least until an accident has happened. It is harder to understand the "mechanisms," because risks can arise from non-linear interactions among normal performance

variability as well as from consequences of failures and malfunctions. And because of that it is also more difficult to think of ways to reduce or eliminate the risks, hence to manage safety.

The second ramification is that many of the established risk assessment and accident investigation methods are inadequate for tightly coupled, intractable systems. This dilemma was made clear when Perrow (1984) proposed that accidents could be seen as normal, in contrast to risk assessment and accident investigation methods that naturally focus on that which is abnormal or dysfunctional. The lesson to be learnt from that is that we must continue to evaluate critically the methods that are at our disposal. The fact that a method has worked in the past is no guarantee that it will also work in the future. Indeed, by the time a method has become adopted as a standard it is almost certain to be outdated. The ways that socio-technical systems develop means that risks can emerge in different ways, and that existing methods therefore sooner or later will need to be complemented with more powerful approaches. What these will be, no one can say for certain.

References

Clarke, S. G. (2000). Safety culture: Underspecified and overrated? *International Journal of Management Reviews*, *2*(1), 65-90.

Cook, R., & Rasmussen, J. (2005). "Going solid": a model of system dynamics and consequences for patient safety. *Quality and Safety in Health Care*, *14*, 130–134.

Hale, A., Kirwan, B. & Kjellen, U. (2007). Safe by design: Where are we now? *Safety Science*, *45*(1-2), 305-327.

Hollnagel, E., Woods, D. D. & Leveson, N. (2006) (Eds.). *Resilience engineering: Concepts and precepts*. Aldershot, UK: Ahsgate.

Hollnagel, E., Nemeth, C. P. & Dekker, S. (2008). *Remaining sensitive to the possibility of failure (Resilience Engineering Perspectives Volume 1)*. Aldershot, UK: Ashgate.

ICAO (2006). *Safety management manual* (DOC 9859 AN/460). Montreal, Canada: International Civil Aviation Organization.

Perrow, C. (1984). *Normal accidents: Living with high risk technologies*. New York: Basic Books, Inc.

Westrum, R. (2006). A typology of resilience situations. In E. Hollnagel, D. D. Woods & N. Leveson (Eds.). *Resilience engineering: Concepts and precepts*. Aldershot, UK: Ahsgate.