the 5th paradox of safety

Erik Hollnagel

Twenty five years ago Jim Reason published a paper on Safety Paradoxes And Safety Culture, which seems to have received surprisingly little attention. In this paper Professor Reason described four paradoxes of safety (a paradox is a seemingly absurd or contradictory statement or proposition which when investigated may prove to be well founded or true.)

1) **First** paradox: That safety is defined and measured more by its absence than its presence.
2) **Second** paradox: That defences, barriers and safeguards not only protect a system, they can also cause its catastrophic breakdown.
3) **Third** paradox: That many organisations seek to limit the variability of human action, primarily to minimise error, but it is this same variability – in the form of timely adjustments to unexpected events – that maintains safety in a dynamic and changing world. (This clearly predates the main idea of both resilience engineering (Hollnagel, Woods & Leveson, 2006) and safety differently (Dekker, 2015)
4) **Fourth** paradox: That an unquestioned belief in the attainability of absolute safety (such as the Zero Accident Vision (ZAV) can seriously impede the achievement of realisable safety goals, while a preoccupation with failure can lead to high reliability.

This paper will discuss each of the four paradoxes in detail and end by proposing a fifth paradox that in many ways is a consequence of the four paradoxes.

First Paradox

The first paradox is essentially about evidence. We have manifest evidence of the ***absence*** of safety, namely the outcomes from accidents, but there is little if any manifest evidence of the ***presence*** of safety, and therefore no way to measure it. This is completely different in the case of quality. Here we know, based on statistical process control (Shewhart, 1931) that quality is present if outputs fall in the six sigma range above the Lower Control Limit (LCL) and below the Upper Control Limit (UCL).There is no similar formal foundation for defining the presence of safety, except the absurd Zero Accident Vision (Zwetsloot et al., 2013); (Sharman, 2014) (Björnberg, et al., 2019), that itself is the subject of the fourth paradox. Quality management, however, also includes a paradox, or at least an *enigma*. Namely that more efforts are spent to understand the occurrence of outcomes below the LCL and above the UCL, than to understand the occurrence of outcomes that are inside the six sigma range, even though these are the outcomes that are both intended and wanted. This is similar to the peculiar lack of interest in safety science and safety management to understand why work goes well, or the presence of safety, which actually is the rule rather than the exception. **The first paradox is therefore true**

Second Paradox

The second paradox is essentially about the limitations of design, not least due to the differences between WAI and WAD. The efficacy of the defences and barriers that are intended to safeguard the system depends on whether the designers have been able correctly to imagine the situations and conditions that could occur, how their safeguards would perform, and not least how people would respond to them.

A further issue with design as a solution, is that design is tantamount to telling stories about the future (Roesler, et al., (2001). Design is about shaping something, an artefact, a work process, a task or job, a machine, or a tool, that does not yet exist, and in particular shape how it should be used to work as intended in a hypothetical future scenario. To do so it is necessary to make assumptions, not only about Work-as-Imagined (WAI). But also about the larger context or ***World-as-imagined***. And telling stories about the future inevitably changes It, including how working conditions will be, and how people respond to the artefacts that have been put in place. Which adjustments or trade-offs are needed to ensure that

the artefacts function despite the way they have been designed. (Hollnagel, 1979). Artefacts can both help protect a system, but can also lead to problems. A simple example is that drivers who do not know how an anti-lock braking system (ABS) works, mistakenly may lift their foot from the brake pedal, when the system engages.

The second paradox is reminiscent of the classical problem known as ***the Ironies of Automation***. Lisanne Bainbridge introduced the term irony in 1989, to describe a solution which could expand a problem it was intended to reduce or even eliminate. The term was first used in the context of automation, but can with equal justification be applied to other issues, such as safety and human factors itself. (Hollnagel and Dekker, 2024).

The general premise for introducing automation is that the human operator is unreliable and inefficient, and therefore should be eliminated from the system.

(Bainbridge, 1989) identified four main problems or ironies of automation.

1) The first irony was that the more advanced a control system is, the more crucial the contribution of the human operator will be. To replace the operator with automation will therefore be counterproductive.
2) The second irony is was that designer errors can be a major source of operating problems.
3) The third irony was that the designer who tries to eliminate the operator still leaves the operator to do the tasks which the designer cannot think how to automate. This is another consequence of the limitations of WAI.
4) The fourth irony is was that automationwas introduced because it could do the job better than the operators, but since automation cannot actually monitor itself the operator is still needed to monitor that the automation works effectively.

The second paradox is true because predictions never are perfect

Third Paradox

The third paradox is essentially about the traditional view of the human factor. Human factors was launched by (Fitts, 1951), who honestly just called human engineering, with the stated purpose to enhance and improve the human factor, both to ensure a better fit between humans and the machines they had to work with, but also to make humans perform more reliably like the machines or technological artefacts they inevitably were compared to (by means of the famous Fitts' list). Fitts argued that "We have been very much occupied in perfecting the machines and tools which the worker uses in the economic arts. We have hardly attempted to improve the worker himself." (Fitts, 1951, p. iv). To address this issue Fitts proposed a method which he introduced in the following way:

> "We begin with a brief analysis of the essential functions ... We then consider the basic question: Which of these functions should be performed by human operators and which by machine elements?"(Fitts, et al., 1951 p.x)

The Fitts list is today probably better known as the MABA-MABA list after the initials of (Men Are better At … Machines Are Better At), (Dekker & Woods, 2002).The human was from the very start seen as an unreliable, and possibly even error-prone part of human-technology systems. And the variability of human performance was seen as a liability, although we today have learned to see it as an asset, as the third paradox also acknowledges. Human Error is now recognised as a judgement in hindsight (Woods et al., 1994, p. 210), Ernst Mach had argued that much already in 1905, but sine it was part of a philosophical discussion rather than industrial safety, it went by largely unnoticed. Mach wrote that "Erkenntnis und Irrtum fließen aus denselben psychischen Quellen; nur der Erfolg vermag beide zu scheiden" or in English: "Knowledge and error flow from the same mental sources, only success can tell one from the other." (Mach, 1905). In plain language "people do what makes sense to them at the time they do it", because anything else would be impossible and it is only afterwards they find out whether it went well or not. Today we would say that people always do what they believe is the best to do in any situation. If they did not they would be fools. Sometimes, however their understanding of the situation differs from reality, but it is rarely done intentionally. Resilience engineering has fully acknowledged that human performance adjustments are essential for work to go well, it is impossible to describe how work should be done in evert detail due to the unavoidable difference between WAI and WAD (Hollnagel, 2017)

and Braithwaite, Wears & Hollnagel, 2017). The term *adjustments* is preferable to *variability*, because *adjustments* have a positive connotation, which variability has not).

Field of safe travel and field of safe work.

Another situation where performance adjustments are essential is driving in city traffic. Here the driver must constantly look ahead for a *field of safe driving*, not knowing with certainty what other drivers will do (Gibson & Crooks, 1938) described it as follows:

"The visual field of the driver is a rather special sort of field in several respects. It is selective in that the elements of the field which are pertinent to locomotion stand out, are attended to, while non-pertinent elements, such as 'scenery,' normally recede into the background. The most important part of the terrain included in this pertinent field is the road. Within the boundaries of the road lies, according to our hypothesis, an indefinitely bounded field which we will name the field of safe travel. It consists, at any given moment, of the field of possible paths which the car may take unimpeded. Phenomenally it is a sort of tongue protruding forward along the road. Its boundaries are chiefly determined by objects or features of the terrain with a negative 'valence' in perception - in other words obstacles. The field of safe travel itself has a positive 'valence,' more especially along its mid-line. By valences, positive or negative, we refer to the meanings of objects by virtue of which we move toward some of them and away from others. The valences of objects with respect to locomotion may be quite different ones from their valences with respect to eating or esthetic enjoyment when the individual is not simply propelling himself between them. For instance, a hot-dog wagon has a negative valence with respect to locomotion, but a positive one with respect to appetite. The situation Gibson & Crooks considered is illustrated by Figure 1, taken from their joint paper. (The hot-dog wagon was not included in the figure).

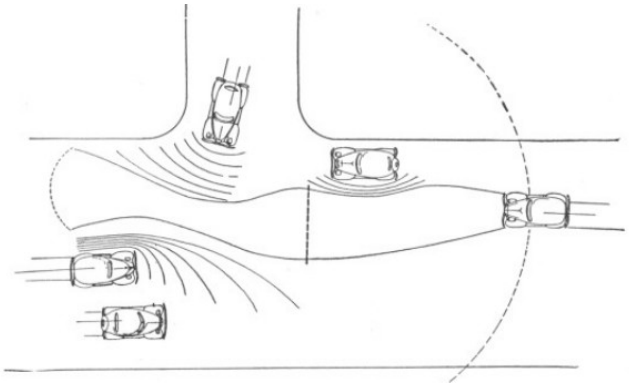

Figure 1: The field of safe travel (Gibson & Crooks, 1938, p. 455)

Fig. 1 is a representation of this field at a specific instant. The field of safe travel, it should be noted, is a spatial field but it is not fixed in physical space. The car is moving and the field moves with the car through space. Its point of reference is not the stationary objects of the environment, but the driver himself. It is not, however, merely a subjective experience of the driver. It exists objectively as the actual field within which the car can safely operate, whether or not the driver is aware of it. It shifts and changes continually, bending and twisting with the road, and also elongating or contracting, widening or narrowing, according as obstacles encroach upon it and limit its boundaries. As will later be evident, other factors may also serve to constrict these boundaries, but for the present we may point out that its depth is limited by frontal obstacles such as other vehicles, policemen and stop lights, and its width by 'marginal' obstacles like curbs, ditches, soft shoulders, walls, parked cars, pedestrians or 'white lines.'

It is now possible to define precisely the operation of steering an automobile. *Steering, according to this hypothesis, is a perceptually governed series of reactions by the driver of such a sort as to keep the car headed into the middle of the field of safe travel"*. (Gibson & Crooks, 1938, p. 454-455).

When driving we must all take into account how others will react to what we do, just as they to some extent must take into account or at least have considered what we will do in response to what they do, This leads to a mutually dependent recursive relation, which quickly becomes very tricky, the reason it works in practice is that there is a set of traffic rules that we all follow, more or less, unless the situation makes it impractical and impossible.

Driving a car is not the same as work except for a very few. The analogue of a Field of Safe Driving (FoSD) can nevertheless be applied to work. Everyone, at the sharp end as well as the blunt end creates a **Field of Safe Work (FoSW**), defined as follows:

> **FoSW** describes work as a perceptual-cognitive process where people continuously assess the range of activities that will achieve their aims and intentions based on the affordances of the workplace (workplace-as-imagined) and their expectations to what others are likely to do. But everyone builds their own field of FoSW, which is why compliance and standardisation do not make much sense, and never will suffice. FoSW is mutually reciprocal and synergistic because everyone tries to imagine how they can successfully do their work, all the time trying to anticipate what others will do. According to this hypothesis, work is a perceptually governed series of reactions by the a person so that the intended work can be carried out as smoothly as possible, without interfering with what others are expected to do or being negatively affected by what others actually do.
>
> The mutual recursive dependency that exists for driving, also exist for daily work, although there are few formal rules in the case of work, mainly the mutual considerations that people in a team show each other. **The third paradox is also true).**

Peformance adjustments as putty.

Since performance adjustments limits predictability, of which the FoST and FoSW both illustrate that. The major source of unpredictability is not in our technological environment but in the people around us, including ourselves. It is clearly preferable if performance adjustments could be avoided. And that is precisely why many organisations seek to limit the variability of human actions as the third paradox describes, modtly without understanding the disadvantages of that. The unattainable ideal is work situations so standardised and rigid, that performance adjustments become unnecessary. Here an interesting analogy is found in the building of walls.

The ideal way to build a wall, is to make sure that every element or stone fits exactly to its neighbour. This can be seen for instance from the Pyramids, Roman bridges, or the walls around Japanese castles, in e.g., in Osaka or Himeji. Or the stone walls built by the Incas (Figure 2.)

Figure 2: An Inca built wall

Figure 3: A rough brick wall

Building a wall as the Incas did is, however, a slow undertaking that requires much work and care. The alternative to use stones that each is carefully shaped to fit precisely with neighbouring stones, is to use prefabricated elements that only fit roughly togetherdo not necessarily fit exactly, such as the bricks we use today, Bricks were invented about 7.000 years ago,in Turkey close to the ancient town of Jericho, which ironically is known today because the walls collapsed. Bricks were at first made of clay that dried in the sun, and replaced by fire bricks about 3.500 years later. These made it much faster to build something, but the disadvantage was that the bricks did not fit perfectly together, so a trade-off between efficiency and thoroughness(Hollnagel, 2009) was required. The lack of a perfect fit was overcome with putty, a fine powder made of white cement mixed with water & other additives to create a solution that filled out the gaps between the rough bricks (Figure 3). After some time the putty will harden an ensure the brick wall remains steady. Today only LEGO bricks fit perfectly together. Just as we no longer can build walls with the precision the Inca people could, neither can we design human-technology systems where the parts fit perfectly together, the efforts of human engineering notwithstanding. In that sense human performance variability is the natural putty that yurns the human-technology system into a whole, and thereby maintains enables safe performance in a dynamic and changing context. ***The third paradox is therefore also true***. In this case the components only have to fit reasonably well to each other, but not perfectly so. The putty will fil any gaps and help to provide the appearance of a whole, especially after some time has passed. It is impossible to fit humans exactly to match the machines, lest we resort to the Procrustean solution, and also impossible to fit the machines exactly to the individual user tailor made technology like tailor made clothes is far too expensive. Instead some kind of putty is. In the case of human factors the putty needed to ensure that the human-technology system works is human agility or performance variability. The deep irony of human factors is that we attempt to remove this putty, because it mistakenly has been seen as a liability rather than the asset it really is. There are interesting historical reasons why this is so. But there are even more compelling reasons to stop doing so.

When WAI is compared to WAD, there is always a small gap or discrepancy. The two never match perfectly (Hollnagel, 2017). So just as when bricks are put together to build a wall or a house, something is needed to fill out and cover the gaps. The alternative would be to ensure that the bricks fit perfectly.

When WAI is compared to WAD, there is always a small gap or discrepancy. The two never match completely (Hollnagel, 2017). So just as when bricks are put together to build a wall or a house, something is needed to fill out and cover the gaps. The alternative would be to ensure that the bricks fit perfectly.

The fact that humans are not machines, and cannot perform a function in exactly the same way multiple times in the way a mechanical clockwork can, and must, was by Paul Fitts seen as a weakness, hence the need of furan engineering. Today we can recognise it as a strength.

The procrustean approach

Since humans as such cannot be designed, at least not yet, the only way to engineer humans was by means of training,in some cases supported by selection of who should be trained. (Such selections are for obvious reasons limited to measurable physical characteristics.

To make sure that the joint human-technology systems, or socio-technical systems as they usually are called today, function as intended, it is necessary that the parts fit together, like the stones in the walls of

the Incas. To describe how this is done (Taylor & Garvey, 1959), resorted to the analogy, of the ***Procrustean*** bed.

The name refers to the legendary ***Procrustes*** (supposedly son of Poseidon), and a robber of Attica, who had an inn on the sacred road between Athens and Eleusis (home of the Eleusian mysteries). The inn had a single iron bed, where Procrustes invited everyone who passed by to spend the night. Procrustes became famous for making the visitors fit the bed, rather than the other way around. If visitors were shorter than the bed, Procrustes would stretch them until they were long enough, and if they were too long he would make them fit the bed by having their feet or parts of their legs cut off. Taylor and Garvey, 1959) used this analogy to criticise training when they wrote that "Two rather different human factor approaches may be distinguished in efforts to optimise the performance of man-machine systems. One seeks to standardise performance by "Making people shorter" , i.e., limit or constrain what they should do, i.e., using less than their full potential by limiting or constraining their performance. The other effort aims at "Making them longer " i.e., extend or stretch human capabilities to meet task demands through additional specialised training.

For work, either solution at best only provides a temporary relief. The problem with wither is that people inevitably and involuntarily will revert to their "natural" level of skills and performance whenever something unexpected happens. When it does people will either neglect or forget the constraints ***and therefore do more than they were supposed to do*** (e.g, intervene, manage by exception (Dekker & Woods, 1999) , hence do too much or do it too early), or return to their former level of comfort and competence and forget what they have been taught ***and therefore do less than they were supposed to do*** (e.g,regress to habitual, intuitive or default actions, wrong procedure, hence do too little and do it too late) This will inevitably increase the difference between WAD and WAI, which is the opposite of the intended outcome In either case the predictable and totally involuntary return to the "normal or "natural" way of performance renders the system incapable of functioning as intended. Despite the allure of the machine analogy and the promises of Skinnerian behaviorism, humans cannot be programmed as mindless robots. It was presumably not a problem for Procrustes if people regained their original length during the night. The stretching might well cease to have an effect, but legs would not spontaneously begin to grow. Training has since the early years been an essential solution for human factors engineering as such and the popularity has hardly diminished over time, despite insufficient evidence that it actually works. Training can, of course, be supplemented by recruiting, i.e., selecting people who naturally have the required competencies as already suggested by the second principle of Scientific Management. (Taylor, 1919)

Fourth Paradox

The fourth paradox is that it the absolute pursuit of perfect safety (ZAV) can be counterproductive, whereas more realistic goals can be valuable and useful.The ZAV can be pursued in two completely different manners, either proactively or reactively.

- **Proactive ZAV** requires that it is possible to predict what will happen in the future. This was the firm conviction of the eminent French mathematician Pierre Simon, marquis de Laplace. It may have been an acceptableposition when he published his treatise *Theorie Analytique des probabilitès* in 1820, but it is no longer seen as reasonable Famous philosophers such as Søren Kierkegaard (1813-1855) knew that "all our knowledge is of the oast but our actions are in the future." ecchoeing what Samuel Taylor Coleridge (1722-1834) had written in 1831 "If men could learn from history, what lessons it might teach us. But passion and party blind our eyes, and the light which experience gives us is a lantern on the stern, which shines only on the waves behind us" . Or as a lesser known Danish humorist wrote in the 19th century "It is difficult to make predictions, especially about the future". The proactive approach to the ZAV is therefore doomrd to fail.
- **Reactive ZAV** does not fare much better. It is only possible if everything that potentially could happen already has happened and that the proper lessons have been learned. But that is clearly also impossible. In both cases the problem with ZAV thinking is that it works by preventing that something happens, either *anything* that could happen or *everything* that has happened. In

practice that means introducing barriers and creating obstacles for work, which invariably will make it more difficult for work to go well. It is a universal experience that if something is made difficult or impossible to do, people will find ways around it that easily may introduce a greater risk than the one that ostensibly was eliminated or prevented.

Fifth Paradox

The fifth paradox actually follows from the first paradox. Since we do not know what the presence of safety is, except by a reduced number of accidents, we are unable to increase the presence of safety. That would otherwise be the logical consequence of statements such as Safety Improvement or Increased Safety. We are therefore willy-nilly left with just one option, namely to increase the absence of safety in order to reduce the number of unexpected and unacceptable outcomes, which surely is paradoxical. This kind of indirect solution is necessary because we today almost a century after Heinrich in 1931 kicked off the concern for industrial safety, still do not know what safety is or have any kind of theory of safety to help us! From a linguistic point of view, while we use the comparative form ***safer***, even though we actually just become ***less unsafe***: we can increase the absence of safety but we cannot increase the presence of safety. We do so because we actually do nor know what safety is. There are no theories or models of safety but plenty about accidents, hence about the absence of safety, but none about the presence of safety. **The fifth paradox is therefore as real as the four others**.

Resilience to the rescue.

The solution is fortunately quite simple, namely to find something that on the one hand contributes to the occurrence of expected and acceptable outcomes, and on the other is characterised by its presence rather than its absence. The solution is, in other words resilience. Resilience was in the beginning defined as follows:

"a resilient system is defined by its ability effectively to adjust its functioning prior to or following changes and disturbances so that it can continue its functioning after a disruption or a major mishap, and in the presence of continuous stresses." (Hollnagel, 2008, p. x).

Resilience has later been defined more explicitly, in the following manner:

"Resilience is defined as the intrinsic ability of a system to adjust its functioning prior to, during, or following changes and disturbances, so that it can sustain required operations under both expected and unexpected conditions." (Hollnagel, 2011, p. 275). will also contribute to safety because it means there will be fewer if any accidents, defined as unexpected events with unacceptable outcomes. Although resilience is defined by its presence, it is still a social construct (Searle,1992), hence cannot be measured nor is it measurable as such. Resilience is a performance characteristic (something a system does) rather than a uniquely defined quality (something a system has). But even if we cannot measure or manage resilience as such, we can measure and manage the potentials for resilient performance. (Hollnagel, 2016), Hollnagel, Licu & Leonhardt. 2021) and (Hollnagel, 2025). The potentials for resilient performance are today known as the foursystemic potentials

Are there paradoxes of resilience?

Since safety carried with it the five paradoxes, it is necessary to ask whether resilience avoids that. The answer is fortunately affirmative.

- Resilience avoids the first paradox, because it is defined by its presence rather than its absence, and is measured or assessed by means of the four potentials.
- Resilience also avoids the second paradox, because it does not depend on a combination of defences, barriers and safeguards measure or assessed by the strength of the four potentials, Figure 4: The potential to respond the potential to monitor, the potential to learn, and the potential to anticipate (Hollnagel, 2016), Hollnagel, Licu & Leonhardt. 2021) and (Hollnagel, 2025)
- Resilience also avoids the third paradox, because it recognises that human performance adjustments and variability are strengths that should ve facilitated rather than limited.
- **Third** paradox: Resilience engineering does not attempt to limit the variability of human action But rather recognises its unique value

- Resilience also avoids the fourth paradox, because it pursues a visio centum (Hollnagel, 2025) rather than a ZAV.
- Resilience finally avoids the fifth paradox, because the aim is to establish and maintain the **presence** of resilience.

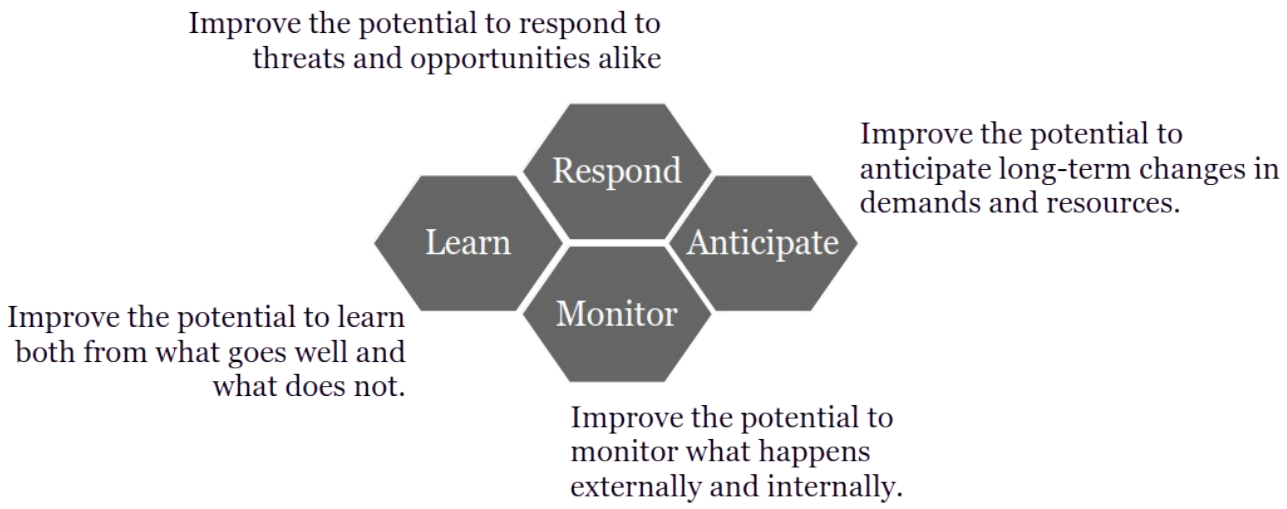

Figure 4: The four systemic potentials for resilient performance

Conclusion

The persistent problem is that unexpected events with unacceptable outcomes (usually called accidents happen, and look likely to continue to happen despite every effort to the contrary. Based on the discussion of the five pardoxes here, it must be concluded that safety as such is not the solution, and neither is safety culture. The obvious question is whether anything can be done to improve this condition. Because something must be done, because the present state is detrimental to our societies and ways of living.

Tthe solution can neither be the defeatism of Murphy's law "Anything that can go wrong, will go wrong" (Bloch, 2003) nor the resignation and quiet acceptance of the "normal accidents" that (Perrow, 1984) introduced. Although both have nominal psycgological value neither can make any practical difference. Or have any concrete impact.

References

Bainbridge, L. (1983). Ironies of automation *Astomatica*, *19*(6) pp. 775 779,

Björnberg, K. E., et al. (2019). *The vision zero handbook*. Cham, Switzerland: Springer International Publishing

Bloch, A. (2003). Murphy's law. Harmondsworth UK: Penguin.

Braithwaite, J., Wears, R. L. & Hollnagel, E. (Eds.)(2017). *Resilient health care, Volume 3: Reconciling Work-as-Imagined and Work-as-Done*. Boca Raton, FL: CRC Press / Taylor & Francis.

Dekker, S. W. A., & Woods, D. D. (2002). MABA-MABA or abracadabra? Progress on human—automation co-ordination. Cognition, Technology ¢» Work, 4, 240-244.

Dekker, S. W., & Woods, D. D. (1999). To intervene or not to intervene: The dilemma of management by exception. Cognition, Technology & Work, 1, 86-96.

Fitts, P. M. (1951). *Human Engineering for an Effective Air Navigation and Traffic Control System*. Washington, DC: National Research Council.

Fitts, P. M. et al. (1951). Human engineering for an effective air-navigation and traffic-control system. Washington D. C: National Research Council.

Hollnagel, E. (1979). *Predictability, Design, and Evolution. NKA/KRU-P2(79)* Roskilde, Denmark: Risø National Laboratories

Hollnagel, E.(2009). *The ETTO principle: Why things that go right sometimes go wrong*. Farnham, UK: Ashgate.

Hollnagel. E (2017) Can we ever imagine how work is done? *HindSight 25* 10-13

Perrow, C. (1984). *Normal Accidents: living with high-risk technologies*. Princeton, NJ: Princeton University Press.

Roesler, A.,et al., (2001). Design is telling stories about the future.Columbus, OH: Cognitive Systems Engineering Laboratory, Ohio State University,.

Sharman, A. (2014). *From accidents to zero: A practical guide to improving your workplace safety culture*.

Shewhart, W. A. (1931). *The Economic Control of Quality of Manufactured Product*. New York, NY: D. van

Taylor, F V., & Garvey, W. D. (1959). The limitations of a 'Procrustean' approach to the optimization of man-machine systems. *Ergonomics*, *2*(2), 187-194.

Taylor F. W. (1919). *The principles of scientific management*. New York: Harper & Brothers Publishers

Zwetsloot, G. I. J. M., et al., (2013). The case for research into the zero accident vision. *Safety Science*, *58*, 41-48.