The Folly of Safety-III

Abstract. The use of the terms Safety-I and Safety-II to characterise two opposite ways to become safe, was met with surprisingly large interest. And, of course also with some skepticism. Even though the terms are very simple and were explicitly defined from the start, there have been some misunderstandings. This paper will specifically look at one of these, namely that Safety-I and Safety-II implies the possibility or existence of a Safety-III, This paper looks at how this happened and tries to explain why Safety-III is an impossible concept.

Introduction

When Safety-I and Safety-II were first described (Hollnagel, 2014), it was as a modest contribution to the slowly growing dissatisfaction with the convential interpretation of safety that grew out of the first workshop on resilience engineering in 2004 as documented by (Hollnagel, Woods & Leveson, 2006). The rhetoric trick to contrast two names if not in a title, then at least in the contents was by no means new. In relation to safety the best known examples are (Dougherty, 1990), which introduced the so-called first and second generation of Human Reliability Assessment (HRA) methods, and a sadly overlooked report (Cook, Woods & Miller, 1998). The idea may possibly be traced back to Charles Dicken's immortal novel " A tale of two cities"published in 1859.

In presenting the concepts of S-I and S-II (Hollnagel, 2014) I explicitly warned against drawing the conclusion that there also would or ever could be a Safety-III.

> "Since Safety-II represents a logical extension of Safety-I, it may well be asked whether there will not some day be a Safety-III? In order to answer that, it is necessary to keep in mind that Safety-I and Safety-II differ in their focus and therefore ultimately in their ontology. The focus of Safety-I is on things that go wrong, and the corresponding efforts are to reduce the number of things that go wrong. The focus of Safety-II is on things that go well, and the corresponding efforts are to increase the number of things that go right.
>
> Safety-II thus represents both a different focus and a different way of looking at what happens and how it happens. Doing so will, of course, require practices that are different from those that are commonly used to day. But a number of these practices already exist, either in principle or in practice, as described in Chapter 8, and can easily be taken into use. It will, of course, also be necessary to develop new methods and techniques that enable us to deal more effectively with that which goes right, and which in particular are able to describe, analyse, and represent the ubiquitous performance adjustments." (Hollnagel, 2014, p.165).

But this warning did not prevent some, e.g., (Leveson, 2020), (Aven, 2022), and (Cooper, 2022), from addressing Safety-III, or perhaps they did not read it. So much for written instructions! And since there, for obvious reasons never was a formal definition of what a Safety-III might be none of the authors had a well-defined target. Of Aven, Leveson, and Cooper only Leveson (2020) attempted to "define" Safety-III, as follows:

> Safety-III: The goal is to eliminate, mitigate, or control hazards, which are the states that can lead to these losses". Leveson's definition of Safety-III was, however, indirectly given in Table 1 (Leveson, 2020, p. 27).

As this definition is a paraphrase of my definition of Safety-I it contains nothing new. (It is really a case of the emperor having no clothes, as Cooper(2022) mistakenly argued.

Both Aven (2022) and Cooper(2022) criticised Safety-II, but neither offered a definition of Safety-III except by reference to Leveson (2020), Cooper even failed to distinguish between resilience engineering, Human & Organizational Performance (HOP), Safety-II and "safety differently" (Dekker, 2015) instead considering them as synonyms, which they clearly are not (le Coze, 2022).

Even though the existence of a Safety-III was never actually proposed, it is not difficult to argue against the very idea of Safety-III. There are, ironically three different types of arguments, a visual argument, a logical argument, and a linguistic argument.

The visual argument against Safety-III

The basic argument can be made like this: Safety-I represents the concern for managing events with unacceptable outcomes (accidents and failures). This follows the first Heinrich axiom which can also be called the Heinrich dogma (as defined by Hollnagel, 2025).

> ”It is widely accepted as true that the cure of a given troublesome condition depends primarily upon knowledge of its cause and the elimination, or at least the mitigation, of that cause”(Heinrich, 1931, p. 31).

(Heinrich, 1959, p. 13) actually proposed a set of ten ‘axioms of industrial safety’, of which the first states:

> “The occurrence of an injury invariably results from a completed sequence of factors – the last one of these being the accident itself. The accident in turn is invariably caused or permitted directly by the unsafe act of a person and/or a mechanical or physical hazard.”

Safety-I in line with the general safety legacy (see Hollnagel, 2024) tries to explain how things go wrong in order to prevent any recurrences.

The safety legacy is the widely and uncritically accepted set of assumptions about how something happens and how events that lead to unacceptable outcomes develop. The safety legacy therefore determines how we perceive and interpret the occurrence of unacceptable outcomes, and how we respond to them -- what we do about them. The natural and practically instinctive response to a sudden unacceptable outcome, which can be harmful, costly or in other ways affect a person's life and activities in an unwanted way, is unsurprisingly to take steps to prevent it, to limit it when it happens, and to prevent it from happening again in the future (yet the latter requires some rather strong assumptions about how the future develops-- what determines future developments and events. In this respect there are clear differences between the three ages of safety defined by (Hale & Hovden, 1998). The focus on work that goes wrong in practice excludes everything else. Safety-II, on the other hand, represents a concern for understanding and managing how events happen regardless of whether the outcomes are acceptable or unacceptable, but especially looks at work that goes well. This is done by trying to understand how work goes well in order both to facilitate the much needed acceptable outcomes and better to dampen or prevent unacceptable outcomes. Safety-I and Safety-II thus do not disagree about the definition of what safety is as there can be only one: Safety is a state where as few acts as possible go wrong, i.e., lead to unacceptable outcomes. Safety-I and Safety-II differ in how a state of safety best can be achieved. Safety-I favours the reduction of acts that go wrong, hence a decremental approach, the ideal being the completely unattainable Zero Accident Vision. Safety-I favours the increase of acts that go well, hence an incremental approach, the ideal being the equally unattainable Visio Centum (meaning that 100% of all acts goes well.

The visual argument is illustrated as follows. If we assume that outcomes follow a normal distribution which they by no means always do, the focus of Safety-I is on rare events with unacceptable consequences as shown in Figure 1.

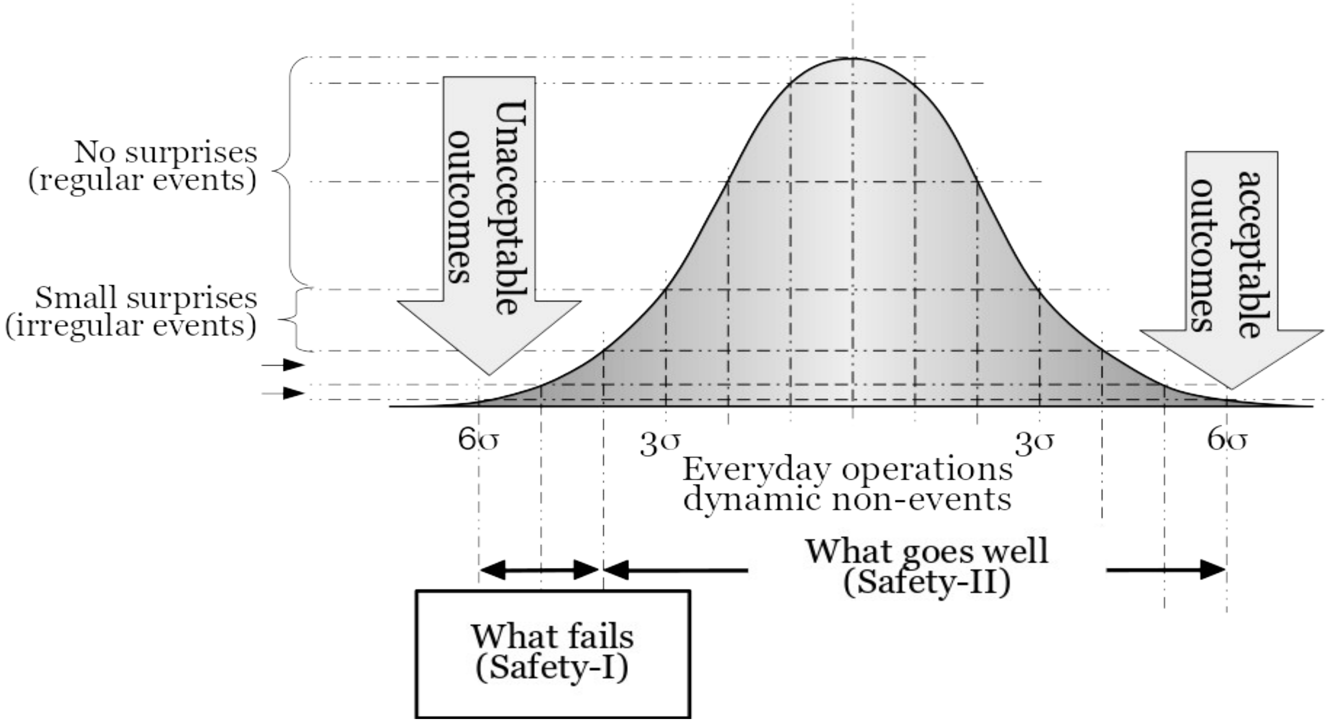

Figure 1: Different foci of Safety-I and safety-II

Safety-I looks at acts that only happen infrequently and are unwanted (unacceptable outcomes of work that does **not** go well hence the low probability outcomes at the left tail of the distribution). Safety-II looks at all events regardless of their outcomes, but in particular at events that occur frequently and lead to the expected outcomes and which therefore by definition go well and are seen as "normal operations" (Hollnagel, Shorrock & Johns, 2021), occupying the middle of the distribution. These also represent the dynamic non-events that Karl Weick (1987) introduced as the foundation of reliable performance. Since Safety-II is concerned with everything that happens (and not just with irregular events that go well or the positive surprises, corresponding to the infrequent outcomes at the right tail of the distribution), there is nothing else to look at. And since Safety-II looks neutrally at all outcomes regardless of whether they are acceptable or unacceptable, there is no other way of looking at them. Safety-II is intentionally biased toward frequent events with acceptable outcomes, but mostly to compensate for these having been traditionally neglected or excluded as being of little or no interest.

It may, of course be counterargued that a Gaussian distribution is inappropriate for the distinction between Safety-I and Safety-II, in line with professor James Reason's safety paradox (Reason, 2000, p. 1) which pointed out that safety is defined more by its absence than by its presence and that safety usually is measured, indirectly by the number of unacceptable outcomes. Safety is therefore not a real-valued random variable for which a continuous probability distribution can be assumed to exist. And even worse when Reason defines safety by its absence it raises the question of whether safety can be modelled or measured at all.

The logical argument against Safety-III

The argument against Safety-III can also be made more formally, as illustrated by Figure 2:

Consider the set of all events, $\boldsymbol{U}$, where the outcome is seen as unacceptable. Then consider the set of all events, $\boldsymbol{A}$, where the outcome is seen as acceptable. Anything that can happen must be a member of the union of the two sets, $\boldsymbol{E} = \boldsymbol{A} \cup \boldsymbol{U}$. Beyond that is only the empty set $\boldsymbol{\emptyset} = \neg\boldsymbol{A} \cup \neg\boldsymbol{U}$

For an engineering view of the safety of socio-technical systems, the focus is limited to **U** and the approach is that of decremental safety or Safety-I. For a systemic or Resilience Engineering view the focus is on **E** and the approach is that of incremental safety or Safety-II.

There could, theoretically, be a study of only ***A***, but it would not make much sense since it would exclude the events with unacceptable outcomes, i.e., ***U***. Established safety practices are effetively limited to the study of and learning from ***U*** (as recommended by Kletz (2001). Since the focus of Safety-II is ***E***, it does by definition include the concerns of Safety-I, although it sees them differently, as useful performance variability rather than as failures (Dekker, 2015; Hollnagel & Dekker 2024). There is therefore no need for a "Safety-III", neither a Safety-IV or any higher order, nor is it logically possible since there cannot be any events that are not members of either ***A*** or ***U***, and therefore of their union ***E***.There is only the empty set **Ø** and that cannot be Safety-III, since there is literally nothing to study and therefore little of value in doing that.

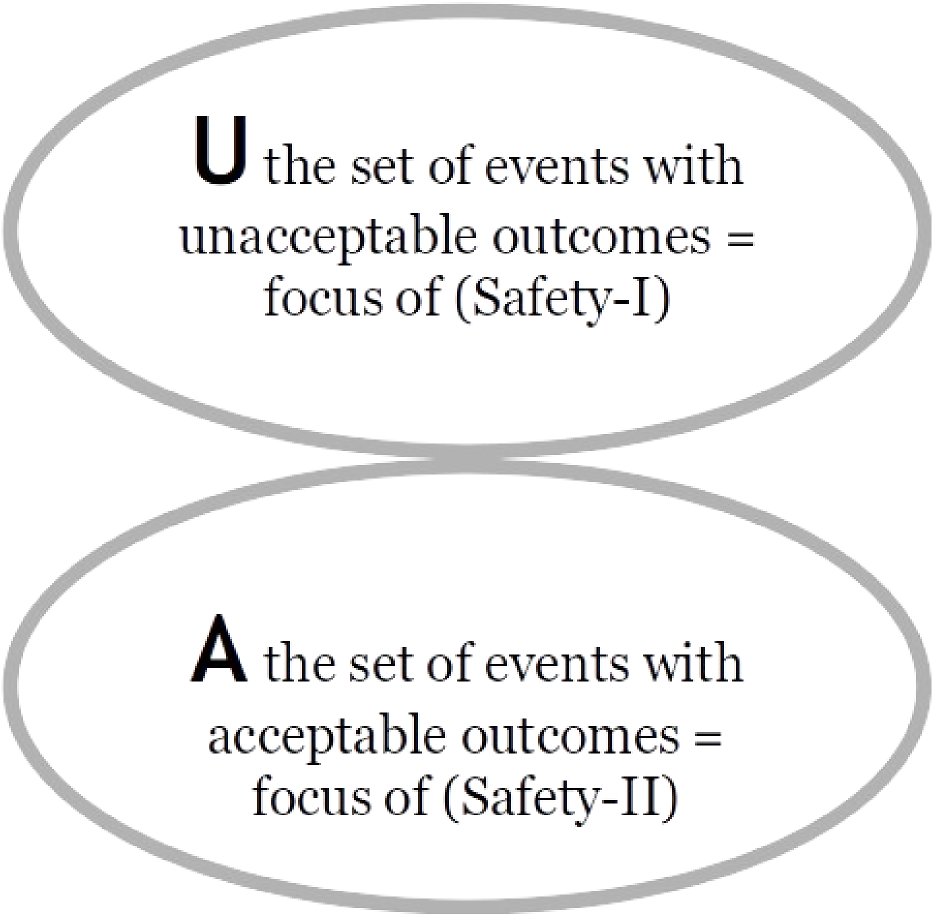

Figure 2: Venn diagram illustrating why Safety-III is logically impossible

The linguistic argument against Safety-III

If the logical arguments fails, third option is the simple semantic trick of renaming the two approaches. And this is easily done. The efforts of Safety-I, by definition aim to reduce the number of events that lead to unacceptable outcomes, this is a *decremental approach*. and resources used toward that purpose are clearly a **cost**, the efforts of Safety-II is to increase the number of events that lead to acceptable outcomes and resources used to achieve that are therefore not a cost but an **investment**. The former can be called decremental safety, because the purpose is gradually to reduce something (it must be gradual because large changes might potentially disrupt the everyday operations of concern and lead to serious unanticipated outcomes (Merton, 1936), Safety-II instead aims to increase the number of events that lead to acceptable outcomes, this is an *incremental approach* (it is again necessary to do it incrementally to avoid a possible disruption of established work patterns and routines, thatgo against the purpose, this can be called incremental safety, because the purpose is gradually to increase something, the names decremental and incremental safety (Hollnagel, 2026) also remove the juxtaposition of Safety-I and Safety-II, and the temptation to postulate a Safety-II therefore disappears. The only logical alternative to either a decremental or incremental approach is to keep the status quo or laissez-fare strategy, which means to tacitly accept that a number of events will lead to unacceptable outcomes (including possible injuries and deaths) which, even if their number was small would be politically and ethically unacceptable, for an organisation as well as for regulators or the public. The third option to keep the status quo, is neither very attractive, nor does it make much sense, since it logically involves spending efforts on something where the outcome cannot be measured. Apart from being financially unacceptable it also flies in the face of the regulator's paradox (Weinberg & Weinberg, 1979). Even if a (very) small number of accidents might be economically affordable for a large organisation that lives by the ALARP principle(Jones-Lee & Aven, 2011), it can never be morally acceptable.

The new names also introduce two different forms of safety culture a decremental safety culture and an incremental safety culture (Hollnagel, 2026), that each are better defined than the vanilla flavour safety culture that created more problems than it has solved since it was proposed some 20 years ago (van Nunen et al, 2018).

Conclusions

The folly of Safety-III is not proposing Safety-III as a concept, which I never did anyway, except as a hypothetical question, followed by a warning not to do it. The folly is arguing against Safety-III as if it had ever been proposed, suggested, or defined. The folly is to rally against something that is just a product of stereotyped and oversimplified reasoning.The folly is to give in to the practically atavistic reaction of inferring a sequence (I-II-III), where there clearly is none and where none was ever intended. In this case the only harm was the folly of proposing Safety-III. But the same reaction is often seen when people "find" causal dependencies while investigating accidents (what the great philosopher Nietzsche aptly called the cause creating drive).

> "(t)o trace something unknown back to something known is alleviating, soothing, gratifying and gives moreover a feeling of power. Danger, disquiet, anxiety attend the unknown – the first instinct is to eliminate these distressing states. First principle: some explanation is better than none … The cause creating drive is thus conditioned and excited by the feeling of fear." Nietzsche (1997, org. 1887, Chapter 5).

I am admittedly a psychologist and not an engineer but, and I am therefore not surprised to meet folly every now and then (Tuchman, 1984) (Morozow, 2013). But I am bewildered when it is dressed up to look like an academic argument as in Leveson (2020).

The inevitable conclusion is that Safety-III not only is an as yet undefined concept but that it also is utterly meaningless and unnecessary with no value for neither current nor future safety practices.

References

Aven, T. (2022) "A risk science perspective on the discussion concerning Safety I, Safety II and Safety III." *Reliability Engineering & System Safety 217* (2022): 108077. Part I.

Cook, R. I., Woods, D. D, & Miller, C. A. (1998) *A Tale of Two Stories: Contrasting Views of Patient Safety. Report from a Workshop on Assembling the Scientific Basis for Progress on Patient Safety*. Chicago, IL: National Patient Safety Foundation. National Health Care Safety Council of the National Patient Safety Foundation at the AMA.

Cooper, M. D. (2022). The emperor has no clothes: A critique of Safety-III. *Safety science*, *152*, 105047.

Dekker, S. 2015. *Safety Differently. Human Factors for a new era*. Boca Raton, FL: CRC Press.

Dougherty, E. M. Jr. (1990). Human reliability analysis - Where shouldst thou turn? *Reliability Engineering and System Safety*, *29*(3), 283-299.

Heinrich,H. W. (1931). *Industrial Accident Prevention*. New York: Mcgraw-Hill Insurance Series.

Hollnagel, E. (2014). *Safety-I and Safety-II: The past and future of safety management*. Farnham, UK: Ashgate.

Hollnagel, E. (2025) *From Safety to Safely: Principles and practice of Systemic Performance Management*. Abingdon, Oxon, UK: Routledge.

Hollnagel, E. (2026) *Decremental and incremental safety cultures: Safety-I and Safety-II revisited*. Bocan Raton, FL. CRC Press.

Hollnagel, E., & Dekker, S. W. A. (2024). The ironies of 'human factors'. *Theoretical Issues in Ergonomics Science*, 1-11.

Jones-Lee, M., & Aven, T. (2011). ALARP—What does it really mean? *Reliability Engineering & System Safety*, *96*(8), 877-882.

Kletz, G. A. (2001). *Learning from Accidents* (3rd Edition.) Routledge.

Nietzsche, F. (1977, org. 1878). *Twilight of the Idols. Or, How to Philosophize with the Hammer*. Indianapolis/Cambridge:Hackett Publishing Company, Inc.

le Coze, J. C. (2022). The 'new view' of human error. Origins, ambiguities, successes and critiques. *Safety science*, *154*, 105853.

Leveson, N. G. (2020). Leveson, N. (2020). *Safety III: A systems approach to safety and resilience*. MIT Engineering Systems Lab Retrieved December 14 2024 from<Sunnyday. mit. edu/safety-3.Pdf.

Merton, R. K. (1936). The unanticipated consequences of purposive social action. *American sociological review,* 1(6), 894-904.

Morozow, E (2013). *To save everything, click here: the folly of technological solutionism*. New York: The Perseus Books Group.

Tuchman, B. (1984). *March of Folly*.New York: Knopf.

Van Nunen, K.et al. (2018). Bibliometric analysis of safety culture research. *Safety science, 108*, 248-258.

Weinberg, G.M. & Weinberg, D. (1979). *On the design of stable systems*. New York: Wiley.