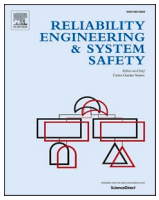
Reliability Engineering and System Safety 217 (2022) 108077

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

journal homepage: www.elsevier.com/locate/ress



A risk science perspective on the discussion concerning Safety I, Safety II and Safety III

Terje Aven

University of Stavanger, Norway

A B S T R A C T

Recently, there has been a discussion in the safety science community concerning the validity of basic approaches to safety, referred to as Safety I, Safety II and Safety III, with Erik Hollnagel and Nancy Leveson in leading roles. The present paper seeks to obtain new knowledge concerning the issues raised, by adding a contemporary risk science perspective to the discussion. Risk is, to a limited degree, addressed in the literature presenting and studying these three approaches; however, as argued in the paper, the concept of risk and risk analysis and management principles and methods are highly relevant and useful for understanding the three safety approaches, deciding on their suitability, and eventually applying them. The paper underlines the importance of an integration of the safety and risk sciences, to further enhance concepts, approaches, principles, methods and models for understanding, assessing, communicating and managing system performance.

1. Introduction

Safety is commonly defined as the absence of accidents and incidents [25]. At a specific point in time, when, for example, walking on thin ice and there is currently an absence of accidents and incidents, you are safe, according to this definition. But, in the next moment, the ice may break and thrust you into the water. This important aspect is not captured by this understanding of safety; nonetheless, safety scientists often refer to it. Authors have pointed to the inadequacy of this definition (e.g., [3]), and it is now often referred to, together with an alternative definition: 'freedom from unacceptable risk' [25]. These two definitions do not, however, express the same idea. When referring to the risk concept, uncertainties are introduced, and – speaking about unacceptable risk – judgments of the magnitude and importance of the risk are also captured. Returning to the ice example, safety in this latter sense is founded in a judgment about the risk related to the ice breaking and its effects for this person. This raises many questions, concerning what risk is and how the magnitude of the risk is determined, and also about how to conclude what is unacceptable risk. To be able to evaluate the soundness and validity of this interpretation of the safety concept, these questions are critical [2,3] but they are not much discussed in the safety science literature. This literature is more concerned about the fact that the attention is placed on unsafe activities or system operations, rather than on safe activities or operations [25]. Safety scientists refer to the former view, having a focus on failures, accidents and losses, as the Safety I perspective. More specifically, the Safety I perspective presumes that things go wrong because of identifiable malfunctions or failures of specific components of the system: technology, procedures, human workers and the organizations in which they are embedded [27]. Furthermore, cause-effect relationships can be established, allowing for a 'fix the problem approach', by identifying the hazards, eliminating them or containing them, and reducing the consequences if a hazardous event should occur. Traditional risk assessment methods, like fault tree analysis, event tree analysis and probabilistic risk assessments (PRAs) (quantitative risk assessments (QRAs)) are seen as key instruments to evaluate the likelihood and importance of the potential scenarios. Following the Safety I approach, a response is needed when something happens or the risk is judged unacceptable, usually by trying to eliminate causes, improve barriers or both [27]. This safety perspective became widespread in the safety critical industries (e.g. nuclear and aviation) between the 1960s and 1980s [27] and is still largely adopted in many industries today, for example oil and gas.

This thinking is in contrast to the Safety II perspective, in which safety is seen as the ability to succeed under varying conditions [25]. Accordingly, safety management should move from ensuring that 'as few things as possible go wrong' to ensuring that 'as many things as possible go right'. Humans are seen as a resource necessary to obtain safety. In line with the Safety II thinking, everyday performance variability provides the adaptations that are necessary to respond to varying conditions and, hence, is the reason why things go right [27]. This perspective is based on a proactive approach, continuously trying to anticipate developments and events. Compared to Safety I, the Safety II perspective requires a different set of methods and techniques, to be able to manage performance variability. Following the ideas and terminology of Hollnagel [[25], p.148], the resilience concept is the sum of Safety I and Safety II, where resilience is understood in this way:

E-mail address: terje.aven@uis.no.

https://doi.org/10.1016/j.ress.2021.108077
Received 2 March 2021; Received in revised form 8 September 2021; Accepted 8 September 2021
Available online 20 September 2021

> A system is resilient if it can adjust its functioning prior to, during, or following events (changes, disturbances, and opportunities), and thereby sustain required operations under both expected and unexpected conditions [26].

Hollnagel [25] also refers to the concept of Safety III but does not establish a specific theory for such a perspective that extends beyond Safety II and resilience. However, other safety scientists have thoroughly discussed the issue, in particular Leveson [35]. In this work, Leveson provides a strong critique of Hollnagel's reasoning, arguing for a Safety III perspective, based on systems theory. According to this perspective, safety is defined as freedom from unacceptable losses (what are unacceptable losses is determined by the system stakeholders). The goal of the related management is to eliminate, mitigate or control hazards, which are the states that can lead to these losses [35]. Accidents are caused by inadequate control of hazards, or more specifically; accidents result from inadequate control or enforcement of safety-related constraints. The focus is on preventing hazards and losses but also on learning from events, accidents, incidents and audits of how the system is performing. The system must be designed to allow humans to be flexible and resilient and to handle unexpected or surprising events [35]. In the following, when referring to Safety III, it is this Leveson perspective that is considered.

This difference in view is noteworthy, and it is natural to ask if it is critical for the understanding, assessment, communication and management of safety, or if it is more about academic quirks, of little relevance for practical safety management. The present paper aims to contribute to this discussion by providing a risk science perspective on the differences between Safety I, II and III. By taking such a perspective, new knowledge is sought, acknowledging that the analysis conducted does not cover all aspects that are relevant for the discussion of Safety I, II and III. Yet, such a risk science perspective is believed to provide important insights, as safety is closely related to risk, and the current discussion in the safety community has, to a limited degree, challenged the more traditional risk concepts, principles and approaches. When referring to key risk science and risk science knowledge in the following, a main reference is documents produced by the Society for Risk Analysis ,51[49–51] and related supporting literature. The SRA documents have been developed by a broad group of senior risk analysts and scientists, with input from members of the society.

To see what such a risk science perspective can add, let us return to the safety concept definition. Safety III, with its formulation 'freedom from unacceptable losses', is based on a similar definition to 'the absence of accidents and incidents' discussed at the beginning of this section. Leveson [35] makes the related comment that

> Complications arise with the introduction of the term 'risk' for measuring safety. Most often, risk is defined as a combination of the severity and likelihood of an unwanted outcome. One problem that arises is that by defining something, such as risk, as only one way to measure it, any alternatives then become impossible including a non-probabilistic measurement. It would be better to define risk as an assessment of safety and then allow different approaches to performing that assessment [35].

However, according to contemporary risk science (e.g., [49]), adding uncertainty to the accidents and losses leads us to the risk concept. It is not based on probabilistic measurement; rather, it acknowledges that different approaches and methods can be used to measure, describe or characterize the risks. As discussed at the beginning of this section, not adding the uncertainty dimension to the undesirable events and consequences leads to both conceptual and measurement problems, as these events and consequences are unknown when looking into the future.

As another example, think about the Safety II and resilience-based approaches highlighting anticipation of events and proactive management. According to Hollnagel [25], the number of intended and acceptable outcomes should be as high as possible, but, as commented by Leveson [35], a statement like 'as high as possible' cannot be meaningfully interpreted without using the risk concept. The same conclusion applies to the 'as low as possible' term used by Hollnagel [25] in relation to Safety I.

The paper is organized as follows. First in Section 2 the methodological approach for the work is described. Then in Section 3, eight important themes are introduced, which show the difference between Safety I, II and III on central topics, including 'the safety concept' and 'anticipation of events and proactive management' briefly discussed above. These themes are discussed in Section 4 in view of risk science knowledge. Finally, Section 5 provides some conclusions.

2. Methodological approach

The present paper is a conceptual paper as discussed in Aven [8], addressing concepts, principles, approaches, methods and models for understanding, assessing, communicating and managing risk and safety. This type of research builds on elements such as: identification (for example, identifying key challenges), revision (for example, changing or modifying a perspective by using an alternative frame of reference), delineation (for example, to focus the study on some aspects and leave others out), summarization (for example, to highlight the key points of a theory), differentiation (for example, to distinguish between alternative definitions and interpretations), integration (for example, to build the analysis on an integrated perspective on risk), by advocating (for example, to argue for a given statement or perspective), and refuting (for example, to rebut a given statement or perspective) [36]. The research is based on creativity, divergent thinking, comparative reasoning, integrative thinking, logic, etc.

Safety I and II were introduced by Hollnagel and Safety III by Leveson, and the present paper focuses its analysis on perspectives and interpretations provided by these authors. Hollnagel and Leveson have made strong contributions to safety science through their work, and the difference in view between these two scientists is therefore important for safety science discussions. However, many other researchers have made contributions to Safety I, II and III, and related topics, providing nuances and enriching the discussion. The paper also seeks to reflect some of these works, in particular recent developments and perspectives. However, the literature on for example Safety II and resilience is huge, and delineation is essential to ensure research focus and clarity on what is covered by the present discussion and what is not. Many researches have contributed to developing Safety I knowledge without referring to Safety I at the time of their work, see discussions in for example Fischhoff [20], Swuste et al. [52], Aven and Ylönen [13], Dekker [18] and Le Coze -[31] [30,31]. As commented by Haavik [22], there are "few or no references to Safety I among those who are given that label".

Safety scientists may refer to the 'absence of accidents and incidents definition' of safety, but operationalize safety differently in methods and in practice. Yet it is important from a safety science point of view to point to the problems of using such a definition. Science requires precision and clarity on fundamental concepts, to ensure rigor and a solid foundation.

The selection of themes in Section 3 was based on an identification of fundamental issues discussed by Leveson [35], and to some extent also issues addressed by Hollnagel [25] for comparing Safety I, II and III. These themes relate to concepts and principles for the understanding, assessment, communication and management of safety. Each issue was reviewed with the purpose of identifying a potential for risk science to add some new insights. A structure for the themes were developed, highlighting some main categories, allowing also the discussion of issues to extend beyond those addressed by Leveson [35] and Hollnagel [25].

As mentioned in the introduction section, the SRA documents and related research provide the reference when referring to risk science and key risk science knowledge. The subjective element in deciding what this knowledge covers is acknowledged. However, building on the comprehensive work by the Society for Risk Analysis, as referred to in

the introduction section, the basis for the analysis is considered rather broad and strong. It is hoped that the conclusions made, as well as the argumentation provided, will stimulate a discussion on how the risk and safety sciences can be better integrated.

One of the reviewers of the present manuscript made the interesting question, how sensitive the analysis and findings of the paper is with respect to this reference. Would a different contemporary account of risk and risk science result in similar conclusions or completely different ones?

A broad discussion of this issue is beyond the scope of the present paper. Many perspectives on risk exist (see e.g. [4]), but not many comprehensive risk science frameworks. Any framework which associates risk with probability would struggle to make sense of Safety II and Safety III, as such frameworks will have many similar features as Safety I. Safety cannot be considered the antonym of risk if risk is defined through probabilities.

3. Eight important themes in the discussion of Systems I, II and III

This section reviews the main differences between Safety I, II and III. The main points are shown in Table 1, highlighting eight central themes, covering basic concepts and principles of safety management. The input on Safety I and II is mainly Hollnagel [25] and on Safety III, Leveson [35].

3.1. Safety concept

For the safety concept theme, reference is made to the introduction section, which pointed to similarities and differences between Safety I, II and III. See also Section 3.3 on variability. There are variations of the safety definitions referred to in Table 1; for example, Provan et al. [44] provide a Safety II definition, considering safety as a system's ability to perform its intended purpose, whilst preventing harm to persons. We observe that, through this definition, the authors combine the Safety II definition of Hollnagel [25], with its emphasis on success, and the Safety I definition's focus on accidents and incidents (harm).

Table 1
Differences between Safety I, II and III, based on Hollnagel [25] and Leveson [35], for eight central themes.

Theme	Safety I	Safety II	Safety III
Safety concept	Absence of accidents and incidents Freedom from unacceptable risks	The ability to succeed under varying conditions	Freedom from unacceptable losses
The risk concept	Events, consequences and associated probabilities	Risk is normally defined as the likelihood that something unwanted can happen	Most often, risk is defined as a combination of the severity and likelihood of an unwanted outcome Risk as an assessment of safety
Variability	Goal to control (limit, reduce) the variability through barriers Commonly expressed through frequentist probabilities and probability models	Inevitable but also useful. Should be monitored and managed Variability (performance variability) is acknowledged as critical for obtaining success and avoiding failures	Design the system so that performance variability is safe, and conflicts between productivity, achieving system goals and safety are eliminated or minimized. Design so that when performance (of operators, hardware, software, managers, etc.) varies outside safe boundaries, safety is still maintained (fault tolerance and fail-safe design)
Causality	Accidents are caused by failures and malfunctions	Emergent outcomes (many accidents) can be understood as arising from surprising combinations of performance variability, where the governing principle is resonance rather than causality	-Accidents are caused by inadequate control over hazards. -Accidents result from inadequate control or enforcement over safety-related constraints Linear causality is not assumed. There is no such thing as a root cause The entire socio-technical system must be designed to prevent hazards; the goal of investigation is to identify why the safety control structure did not prevent the loss.
Models and system characterizations	Assumed to accurately represent the actual system or activity System characterizations: Simple, linear, tractable, complicated systems which allow for decompositions and accurate models based on system components	Accurate models do not exist for intractable systems System characterizations: Intractable, complex, sociotechnical	Models are used by humans to understand complex phenomena. By definition, they leave out factors (otherwise, they would be the thing itself and not useful). For abstractions or models to be useful, they need to include the important factors or factors of interest in understanding the phenomena and leave out the unimportant. The simple linear chain-of-events causality model leaves out too much to be useful in understanding and preventing accidents in complex sociotechnical systems. Alternatives exist based on Systems Theory (STAMP) System characterizations: Linear and more complex sociotechnical systems
Risk assessment	Traditional technical risk analysis methods like FTA, ETA, PRAs (QRAs). Accurately estimate risk using probabilistic-based methods	Not highlighted. Traditional methods not relevant for intractable systems Focus on understanding the conditions where performance variability can become difficult or impossible to monitor and control	Traditional use of risk assessments to identify design flaws and functional glitches, highlighting events, consequences and likelihood
Safety management principles - Anticipation and proactivity	Reactive, respond when something happens or is categorized as an unacceptable risk	Proactive, continuously trying to anticipate developments and events	Concentrates on preventing hazards and losses but does learn from accidents, incidents, and audits of how system is performing
Safety management principles - Learning and improvements	We mainly learn and improve because of failure and mistakes	Learning should be based on frequency rather than severity; thus, weight is given to what goes right, not only failures	We mainly learn and improve because of failures and mistakes

The underlying motivation for the Safety II definitions is that, to understand how accidents occur, it is also necessary to address performance. As Provan et al. [44] state, "the safety-II enables people to dynamically align the pursuit of both safety and effectiveness because there are always multiple conflicting goals, limited resources, and pressures to achieve more".

3.2. Risk concept

The risk concept is understood in line with the traditional perspective, seeing risk as events (hazards), consequences and likelihood (probability). The risk concept is not commonly discussed as a research topic in safety science; rather, it is focused on what is current thinking and practice. However, Leveson [35] points to a broader perspective on risk when stating that risk is an assessment of safety. The use of the risk concept is addressed in Section 3.6 on risk assessment and further discussed in Section 4.

3.3. Variability

Variability is a key concept in Safety II. It expresses the potential or the propensity to vary. Consider the following simple but illustrative example. Every day, John prepares a hot cup of tea, using the microwave. The warming-up time is commonly 2 or 2.5 min. When it is ready, he quickly takes the cup out of the microwave, sometimes using the handle but most often not. The types of cup vary, as does the amount of water in the cup and the energy level (Wattage). Thus, the variables of interest are time in microwave (X1), using the handle or not (X2), type of cup (X3), volume of water (X4) and Wattage (X5). The activity is nearly always successful, but sometimes the tea is too hot.

Is the activity safe? Following the Safety II nomenclature, it is safe, depending on the ability to succeed under varying conditions. These conditions relate to the variability as reflected by the variables X1, X2, … X5, and potential other variables. For most combinations of X1, X2, … X5, the outcome is a success. The success is ensured if the cup is of a rather standard type, fully filled with water, the warming-up time is no longer than 2.5 min and the cup handle is used. Other combinations could, however, lead to failures, for example types of burning incidents, as will be explained below.

Following Safety II, John makes choices (approximate adjustments) on how long to warm up the water, taking into account the water volume and the Wattage. Detailed analysis could have produced an ideal or optimal choice in this case, but John has not done this.

Suppose a case where John happened by a mistake to specify a three-minute warming-up of the water, instead of 2.5. John commonly takes out the cup without using the handle, but the three-minute warming-up should make him perform an adjustment, using the cup handle to be safe. In this case, John is distracted and, when taking out the cup, he forgets that it was in there for 3 min and not 2.5. He takes the cup, without using the handle; it is very hot, and he quickly draws his hand back, with the result that the hot water spills onto his hand and he is burnt. This combination of variables has led to a failure. For future uses of the microwave, this experience can be useful, to strengthen the ability to succeed under varying conditions, i.e. improve safety.

According to Hollnagel [25], Safety II is achieved because people make what they consider sensible adjustments to cope with current and future situational demands. Finding out what these adjustments are and trying to learn from them constitute key elements of Safety II thinking.

Following Safety I, the aim is to reduce the variability as much as possible, to obtain control. In the above John example, we can think about introducing a procedure specifying the above variables, for example by using a specific type of cup, the same Wattage, a specific warming-up period, and always using the cup handle. Following this procedure, failure will not occur. Assessments of the frequency with which the procedure is not complied with are then conducted, to show that the incident (accident) risk is minor, and that the activity is safe. Estimation of frequentist probabilities and probability models are used for this purpose; refer to the discussion in Section 4.

Safety III highlights the design process of the system: Ensure that the performance variability is safe and that conflicts between productivity, achieving system goals, and safety are eliminated or minimized. For the above microwave example, this means a system specification which balances the need for flexibility to warm the water and ensuring no incidents or accidents. The specification could, for instance, require that the cup handle always be used. To prevent an incident occurring if John does not use the handle, the cup could be equipped with a heat-proof sleeve to protect the hand. In more general terms, Leveson [35] highlights the need for a design so that when performance (of operators, hardware, software, managers, etc.) varies outside safe boundaries, safety – no harm – is still maintained (for example, using fault tolerance and fail-safe design). We will return to this example in Section 4.2, incorporating aspects of uncertainties and risk.

The above example illustrates the variability concept in relation to Safety I, II and III. The system is simple and key features of more complex (intractable) systems (see Section 3.5) - as Safety II and III explicitly aim at addressing - are not shown. Such systems would have been more informative in order to demonstrate the suitability of Safety II and III, but it would be more difficult to explain the basic ideas of the variability concept.

3.4. Causality

Safety I is based on simple cause-effect relationships. Accidents are caused by failures and malfunctions of combinations of system components, as for example represented by fault trees and event trees. It is common to refer to the concept of 'root cause', which is based on the idea that it is possible to find a basic cause that is the root or origin of the problem [23,25]. Following this terminology, the 'root' cause could be identified as, for instance, 'poor quality of the maintenance work' or, in the microwave example, as 'lack of planning of activity' or 'lack of a clear procedure'. However, as discussed by, for example, Hollnagel [23, 25], the concept of root causes is not meaningful. There will always be a need to specify a set of conditions, states and events to explain an accident. It is not sufficient to point to one underlying factor. Safety III also rejects the idea of root causes. The point being made is that the "event that is chosen as the root cause is arbitrary as any of the events in the chain could theoretically be labeled as such or the chain could be followed back farther" [35].

In accordance with Safety II, an observed accident can be explained by referring to a combination of variables (conditions, states, events). That is not the same as making a statement about causality. Often, the combination comes as a surprise; in some cases (for intractable systems, see discussion below), it is also 'emergent', meaning it is difficult or impossible to explain what happens as a result of known processes or developments, for example, not predictable based on the knowledge about the system components [25]. The effects of the combination of variables are governed by functional resonance, which means that the variability of two or more functions can coincide and either dampen or amplify each other to produce an outcome or output variability that is disproportionally large. FRAM (Functional Resonance Analysis Method) is built on this idea [24]. It is a tool to obtain system knowledge, using the concepts of variability, approximate adaptation and functional resonance. Considerable research has been conducted to apply and further develop FRAM, as shown by Patriarca et al. [43]. On the basis of the review by Patriarca et al. [43], it can be stated that FRAM is seen by many safety scientists as a central approach in understanding and modeling complex, dynamic socio-technical systems.

According to Hollnagel [25], the causality credo makes sense in the case when we reason from cause to effect (forwards causality): for example, if a set of components of a system fails, the system fails. However, the opposite type of reasoning is problematic, in the same way as searching for root causes. Hollnagel's key message is that we cannot

go from effect to cause (backwards causality): it is logically invalid. If an undesirable effect occurs, we may, for example, observe that two variables have interacted strongly, and a third variable was in a specific state. This does not, however, allow us to point to a specific cause or a set of causes; see the discussion in Hollnagel [25] concerning the burning batteries of the 787 Dreamliner.

According to Safety III, accidents are caused by inadequate control of hazards (inadequate control or enforcement of safety-related constraints). Understanding causality and developing suitable causality models are considered essential to prevent the accidents and reduce their consequences. The goal of accident investigations is to identify why the safety control structure did not prevent the loss. The entire sociotechnical system must be designed to prevent hazards [35].

In addition to linear causality, causal loops and multiple common (systemic) factors are referred to. When explaining what causality means, Leveson refers to John Stuart Mill (1806–1873), who claims that a cause is a set of *sufficient conditions*: "The cause is the sum total of the conditions positive and negative, taken together, the whole of the contingencies of every description, which being realized, the consequent invariably follows" [39]. Thus, causes (forwards causality, to use Hollnagel's terminology) can be defined in relation to a system analysis, when a set of components failing leads to system failure. Following this view, it cannot, however, be concluded that smoking causes cancer, as smoking does not always lead to cancer. We will discuss the causality concept further in Section 4.2.

3.5. Models and system characterizations

Safety I is based on the idea that models can be established, providing accurate representations of the actual system and activity resulting in an accident. In this way, risk events can be controlled and, to a large extent, avoided; the accident risk is acceptable. Safety II rejects this perspective, arguing that such models cannot be established for many real-life systems (referred to as intractable systems, see below). Safety III acknowledges the importance of making causality models of the system and activities studied but stresses that the traditional linear models (in which adverse outcomes are due to combinations of failures and latent conditions) are not sufficient for understanding complex phenomena – they leave out critical factors. An alternative is promoted: STAMP (System-Theoretic and Processes), which is based on Systems Theory as described by Leveson [33–35]. Some unique aspects of this theory include:

- The system is treated as a whole, not as the sum of its parts ("the whole is more than the sum of its parts").
- A primary concern is emergent properties: properties that are not in the combination of the individual components but 'emerge' when the components interact. Emergent properties can only be treated adequately by taking all their technical and social aspects into account. Safety and security and most other important system properties are emergent.
- Emergent properties arise from relationships among the parts of the system, that is, by how they interact and fit together.
- Systems are viewed as hierarchical structures, for example covering system development reflecting levels such as Congress and legislatures, government regulatory agencies, company management, project management, and manufacturing management (similar to [47]). Each level imposes constraints on the activity of the level beneath it. Safety-related constraints specify relationships between system variables that constitute the nonhazardous system states, for example, the power must never be on when the access door is open [33]. The control processes (including the physical design) that enforce these constraints will limit system behavior to safe changes and adaptations.

STAMP expands the traditional linear causal models to include more complex processes and unsafe interactions among system components, represented by causal loops and multiple common (systemic) factors. STAMP can be applied to complex systems, as it models them top-down rather than bottom-up: all the system details need not be considered, i.e., the system is studied as a whole and not as interacting components [35]. Using this tool, safety is treated as a dynamic control problem rather than a failure prevention problem.

Safety I is suitable for simple, linear, tractable and complicated systems. A system is tractable if the principles of its functioning are known, if descriptions of it are simple, with few details and, most importantly, if it does not change while it is being described ([25], p. 118). A tractable system can be complicated, i.e., it has many components, but we understand and have good knowledge about how they relate to each other and how the system works. It is possible to provide a detailed description of the system.

Safety II focuses on intractable sociotechnical systems. The term 'sociotechnical' relates to the interconnections between the social and technological aspects [13,29,34,35]. A system is intractable if the principles of its functioning are only partly known (or, in extreme cases, completely unknown), if descriptions of it are elaborate with many details, and if systems change before descriptions can be completed ([25], p. 118). A complex system can be intractable; its performance cannot be accurately predicted based on knowing the specific functions and states of the system's individual components [49]. For a complex system, we would not be able to identify all scenarios of interest – the actual scenario occurring will come as a surprise [54]. Safety III considers both linear and more complex sociotechnical systems.

3.6. Risk assessments

Safety I is based on traditional technical risk analysis methods like fault tree analysis (FTA), event tree analysis (ETA), and PRAs (QRAs). A main purpose of the assessments is to accurately estimate the activity risk, by identifying relevant scenarios and calculating associated probabilities. Using the results of the risk assessments, conclusions can be made on whether the risk is acceptable (tolerable) or not.

A main goal of the Safety II literature is to argue that Safety I and particularly traditional risk assessments have strong limitations in capturing important aspects of safety. The risk assessments are not able to properly reflect variabilities, human and organizational aspects, and dependencies between system elements. For intractable systems, risk assessments are incapable of providing accurate estimations and predictions.

Safety II with its focus on accident analysis has been applied in different types of risk assessments, see Hollnagel [24] and Patriarca et al. [43]. Safety III highlights the use of hazard analysis – to understand the system and identifying hazards. It points to the limitations of traditional risk assessment for analysing complex systems.

3.7. Safety management principles - anticipation and proactivity

Following Safety I and Hollnagel [25], the management response is very much reactive, in the sense that it reacts based on historical events and risk calculations to a large extent founded on such events. There are also some proactive elements as discussed in Section 4.3 - a main task of risk assessment is to identify potential events, also new types of events.

Safety II is more proactive in its continuous focus on variability, trying to anticipate developments and future events. Safety III acknowledges the importance of hazard analysis as a proactive tool to anticipate what can happen, not only reflecting historical events. Learning is considered central in both Safety II and Safety III.

3.8. Safety management principles – learning and improvements

Safety I and Safety III are based on the thesis that we mainly learn

and improve because of failures, mistakes and accidents. Safety II, on the other hand, emphasizes learning through what goes right: "It is essential to learn from what happens every day – from performance variability and performance adjustments – because this is the reason why things sometimes go wrong, and because this is the most effective way to improve performance" ([25], p. 163). In general it makes better sense to base learning on small but frequent events than on rare events with severe outcomes ([25], p. 160).

4. A risk science perspective on Safety I, II and III

In this section, the three perspectives, Safety I, II and III, will be discussed in view of current risk science knowledge.

4.1. The safety and risk concepts

Concepts and their definitions are essential scientific building blocks, so also for the risk and safety sciences. As discussed in Section 1, the safety concept, as defined by Safety I, II and III, is problematic. There are three main issues:

(a) The safety concept cannot solely be linked to success – it needs to relate also to undesirable events/consequences.
(b) The safety concept cannot be defined based purely on *absence* or *freedom* from undesirable events/consequences – the severity of these undesirable events/consequences needs to be reflected.
(c) The safety concept cannot be defined without addressing uncertainty – thus also risk ('uncertainty' here refers to epistemic uncertainty: the consequences of the activity considered are uncertain as a result of lack of knowledge).

Issue a) applies to the Safety II definition. 'Ability to succeed' means that there also needs to be a potential for 'no success' or failure in some sense. Without specifying what 'no success' means, it is impossible to make meaningful judgments about safety being large or small. In the microwave example of Section 2, success can be defined as John getting a hot cup of tea. However, 'no success' can be associated with different things, including the water not being at the correct temperature or John being burnt by the hot water. As another example, think about the successful operation of a production plant. If we are to assess whether the safety is high or low, we need to clarify what a non-successful operation means. For example, it would matter a lot if there is a potential for fatal accidents or not.

The safety concept must clarify what the undesirable events/consequences are, to contrast the success. The definition by Provan et al. [44] referred to in Section 3.1 represents an adjustment of the Hollnagel [25] definition, meeting this challenge when considering safety as a system's ability to perform its intended purpose, whilst preventing harm to persons.

Leveson [35] provides some interesting reflections in relation to this discussion. She refers to several quotes from Hollnagel, including:

> 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 right, and the corresponding efforts are to increase the number of things that go right. Hollnagel [[25], p. 179]

> It is more important—or should be more important—that things go right than that things do not go wrong ([25], p. 136).

Leveson [35] questions the rationale for this reasoning. She asks: "Is it more important that passengers enjoy their flight (a thing that can go right) than that planes do not crash (a thing that can go wrong)?" She warns that too strong a weight on what goes right could lead to dangerous practice, as resources are used on aspects that are not important for preventing hazards and accidents or reducing their consequences. Moreover, if things go right, there could be no serious effects, from a short-term perspective, but, under different circumstances, an accident could occur. Leveson asks how that is revealed if the focus is on what is going right.

Issue b) relates to all three safety perspectives, I, II and III, but not the alternative definition for I: 'freedom from unacceptable risks', refer to Table 1. As an illustration, consider two activities, 1 and 2. For both activities, the outcomes in a specific period of time are success, an accident with a substantial loss (e.g., one fatality), or an accident with a severe loss (e.g., 100 fatalities). Suppose that the related probabilities for activity 1 are 0.95, 0.05 and 0, respectively, whereas, for activity 2, they are 0.95, 0.01 and 0.04. During a specific time period, no accidents or losses occur, safety is achieved and, according to these safety definitions, the safety is the same for the two activities. Looking forward in time, the safety is unknown, but the probability of no accident or freedom from unacceptable loss is the same. Should that imply that the safety is the same for the two activities? Safety I (absence of accidents and incidents) and Safety III (freedom from unacceptable losses) do not provide clear answers on this question, as they address neither the severity of the accidents and losses nor the probabilities (more generally, the uncertainties). Intuitively, activity 1 is much safer than activity 2, as the most severe consequences are not possible for activity 1, and the probability of activity 2 leading to the extreme outcome is close to the same probability as activity 1 resulting in an accident with a substantial loss.

The example points to the need to also take into account the severity of the undesirable events/consequences, as well as the uncertainties. This leads us to issue (c). Looking into the future, there is uncertainty about 'absence of accidents and incidents' and 'freedom from unacceptable losses'. Hence, using these safety definitions does not allow us to talk about the safety being high or low when considering a future activity, for example the operation of an industrial plant or a journey by plane, as the safety is unknown (uncertain). However, in adding uncertainty to the events and the consequences of the activity considered, we are led to the risk concept, as defined in its most general form [6,10,49]: Risk has two main dimensions, (i) events and consequences of the activity considered, and (ii) associated uncertainties. There are always some consequences labelled undesirable or negative, but the result of the activity could also lead to positive or desirable consequences. Intuitively, the risk concept reflects the potential for an activity to have undesirable consequences.

A distinction is made between the concept of risk and how it is measured or described (characterized). To make a judgment about the risk magnitude, we conduct risk assessments, specify the events/consequences and measure or describe the uncertainties. Probability is the most common approach used for the uncertainty measurements, but it is not the only one available [21], and, for the purpose of adequately characterizing the uncertainties, it is not in general sufficient. This is important for discussions related to safety. To explain, let us return to the above example with two activities, 1 and 2.

Suppose the probabilities for activity 1 are based on an assumption, AS. If this assumption does not hold, the probabilities, 0.95, 0.05 and 0, would not apply. The confidence in the specified probabilities would thus strongly depend on the validity of the assumption AS. Clearly, a conclusion about the activity being safe on the basis of the probabilities alone cannot be justified; we also need to consider the reasonability of the assumption AS. More generally, we need to take into account the strength of the knowledge supporting the probabilities. The activity can be considered safe only if the probabilities related to undesirable consequences are low and the strength of knowledge is strong. Aspects to consider when making judgments about the strength of knowledge include justification of assumptions made, amount of reliable and relevant data/information, agreement among experts and understanding of the phenomena involved [6,10,21]. There should be a special focus on reviewing the knowledge basis, to identify potential surprises: for example, unknown knowns (the analyst team does not have the

knowledge, but others do). What is sufficient for concluding that the probabilities are sufficiently low and the strength of knowledge sufficiently strong depends on the situation considered, reflecting relevant requirements and guidance provided by standards and comparable activities. Other factors, such as costs and risk perception, could also influence these judgments.

The activity is judged safe if the risk is judged sufficiently low. We have established safety as the antonym of risk [2]. This reasoning is valid, conditional on a perspective on risk as here adopted. If risk is expressed through probabilities alone, this link between safety and risk would not apply. A safe activity means that also the risk related to for example potential unknown knowns is judged small. Activities 1 and 2 cannot be considered safe if the probabilities referred to are based on a weak knowledge basis. Weak knowledge could also relate to the potential consequences, for example, opening the door to new and extreme type of outcomes.

Unknown unknowns (events that are completely unknown to the scientific environment) are covered by the risk concept when defined according to (i) and (ii) above. When assessing the risk, the contribution from this type of event can be acknowledged, but not measured beyond qualitative statements like "the risk related to unknown unknowns are considered minuscule because the knowledge of the activity considered is very strong" and "the risk related to unknown unknowns cannot be ignored as the activity is subject to considerable, fundamental uncertainties". Similarly the safety concept interpreted along with (i) and (ii) would reflect unknown unknowns, but what about judgments about the safety being high or low? Clearly the safety cannot be judged high without considering the risk of unknown unknowns to be small. Low judged risk (reflecting also unknown unknowns) means high judged safety. Conversely, high judged safety, means low judged risk (reflecting also unknown unknowns), etc. As safety is linked to risk according to (i) and (ii), the judgments of high or low will follow the same type of logic and rationale.

This discussion also leads us to the alternative Safety I definition: absence of unacceptable risks. When making the judgment that an activity is safe because the risk is sufficiently low, we can interpret this as absence of unacceptable risks. The unacceptability judgment is based on aspects like those referred to above (standards and comparisons with related activities, costs, risk perception issues). Note that we may decide to accept the risk of an activity, even if it is considerable, provided the benefits are high. Then, safety may not be considered to be achieved. What is sufficient for concluding on being safe is a judgment call. Following risk science, it can be based on risk and the absence of unacceptable risk type of judgments.

4.2. Risk assessments, variability, causality, models and system characterizations

An important value of risk assessments is the improved understanding of the system or activity studied. This understanding relates to, for example, the interactions of subsystems and components, including the revelation of common-cause failures, that is, multiple component failures with a common source. Different types of models are developed for this purpose, including system models and probability models. Both types of models reflect variability and causality. Let us first consider system models. Many such models exist, and risk science builds on the total knowledge available on the scientific foundation and practice concerning such models. This means the acknowledgement of different types of models, with their strengths and weaknesses. In general terms, a broad class of such models can be expressed as g(X), where X is a vector of input quantities (variables), and g the model linking X and the output quantity (variable) Y. Such models include fault trees, event trees, load-strength relationships and influence diagrams. A model of this type can be derived for the tea example of Section 3.3. The output Y represents the success and failure of the warming-up process as a function of the input variables, X1, X2, …, X5.

Such models are commonly used in risk assessments; hence, variability in system and system component performance is to a large extent reflected. The models referred to here are simple, linear and causal, and based on decomposition of the system studied into a set of subsystems or components. The models express that if X is the state of components, the system state will be g(X). This leads us to the question about the accuracy of the model. The actual, true output quantity Y could deviate from the model output g(X) – there is a model error g(X) -Y. For simple, linear, tractable and complicated systems, proper modeling makes it possible to control this error, ensuring that the models are sufficient accurate approximations of the phenomena studied. The usefulness of this type of modeling in risk assessment is broadly recognized. The problems arise when analysing intractable and complex systems.

To consider first the intractable systems, as studied by Safety II and Hollnagel [25], the functioning principles of the system are partly unknown, and the system structure changes over time. Hence, it is not possible to develop accurate models of the above type g. The result is that attempts to use risk assessment to accurately estimate risk would fail. The conclusions are the same for complex systems, as, for such systems, performance cannot be accurately predicted based on knowing the specific functions and states of the system's individual components.

To meet this challenge, alternative models are needed. Two examples are FRAM and STAMP, as referred to in Section 3. Risk science welcomes all models that can be used to improve the understanding of systems and activities, as the standard ones based on linear causal modeling are not suitable for intractable and complex systems. As is clear from Leveson [35], there is discussion concerning what approaches and models are best suited to this purpose, as there is for all concepts, principles, methods and models within a scientific field. It is beyond the scope of the present paper to perform an evaluation of the suitability of FRAM and STAMP for system analysis and modeling, but it is important to note that uncertainties are not explicitly included. To apply these models in a risk assessment, uncertainties need to be taken into account. The issue is discussed by Bjerga et al. [16] and Bjørnsen et al. [17]. As discussed in Section 4.1, safety cannot be meaningfully discussed without incorporating uncertainties and risk, which motivates a stronger integration of safety science and risk science knowledge on analysing the performance and failures of different types of systems.

Risk assessment of intractable and complex systems is not about accurately estimating risk but about understanding risk – what events can occur, how and what are the consequences – and about describing and characterizing uncertainties and knowledge [6,51]. With such a perspective, risk assessments can always be conducted and produce relevant information for decision makers and other stakeholders.

The causal chains and event modeling approach, using models of the type g(X) referred to above, have been shown to work for a number of industries and settings. The approach has limitations in accurately describing the world, but the suitability of a method and model also has to be judged with reference to its ability to simplify the world. All models are wrong, but they can still be helpful, to use a well-known phrase.

The degree to which systems are, in fact, intractable or complex can be discussed. There could be intractable or complex aspects of a system or activity, but the main functions are still approximately linear/complicated. Within many industries (e.g., nuclear, process and oil & gas), it is possible to list at an overall level what types of undesirable events (hazards, threats) can occur, but how these will occur is not always straightforward. If we consider the accidents and near-accidents occurring, most systems and activities will in fact be described as complicated – there are many sub-systems and components, but we understand and have good knowledge about how they interact. This does not mean, however, that surprises do not occur, but it is often because the knowledge of the assessors in the relevant case is weak or wrong, and not as a result of fundamental, scientific deficiencies in the knowledge of the relevant processes and phenomena [12].

Probability models are another category of models used in risk

assessment. Similar to g(X), such models are functions of probability models of system components' characteristics. In the simplest form, they can be written as h(p) for a function h, where p is a vector of frequentist probabilities. Referring to the microwave example, a probability model can be established, producing a frequentist probability that John is able to successfully warm up the tea, as a function of the frequentist probability distributions for the variables, X1, X2, …, X5. A probability model is a set of frequentist probabilities, expressing fractions of time in which the relevant events will occur if the situations could be repeated over and over again infinitely. Thus, probability models and frequentist probabilities need to be justified as any other models. For unique situations, such models and probabilities do not exist. The frequentist probabilities are, in general, unknown and uncertain and need to be estimated. Statistical inference, both traditional and Bayesian, is a well-established tool to conduct such estimations and express uncertainties. The strengths and limitations of these approaches in risk contexts have been thoroughly discussed in risk science; see for example Aven [11] [10,11].

The probability models in risk assessments are based on linear causal models and are continuously improved, incorporating human, operational and organizational factors. An example is Mohaghegh et al. [40], who present a 'hybrid' approach to studying dynamic effects of organizational factors on risk for complex socio-technical systems. The approach integrates system dynamics, Bayesian belief networks, fault trees and event sequence diagrams.

For intractable and complex systems, a decomposition model approach would not work, but studying the system as a whole and not as interacting components, also probabilistic analysis could be informative. Frequentist probabilities could be justified in some cases, but knowledge-based (subjective, judgmental) probabilities can always be specified, as they express the analysts' judgments – the degree of belief – that the events will occur, or that the statements are true. If the assessor expresses that the probability is at least 0.95, it means that the assessor's uncertainty and degree of belief is comparable to randomly drawing a red ball out of an urn containing 100 balls, of which 95 or more are red. This probability is conditional on some knowledge K, and this knowledge needs to accompany the probabilities, together with judgments of the strength of the knowledge, as discussed in Section 4.1. Hence, risk can always be assessed, described and characterized.

Thus, there is potential for Safety II to be further developed by incorporating aspects of uncertainties and risk in the modeling and analysis of the safety, not for the purpose of obtaining accurate risk estimates but to improve the understanding of the safety (and the risks), supporting decision makers and other stakeholders in their handling of the safety (and the risks). Some papers addressing the issue were mentioned above [16,17], but this is considered an important research challenge requiring much more work.

Safety III has the same potential. Uncertainties and risk are not integral aspects of Safety III and STAMP. The importance of likelihood judgments is acknowledged in Safety III, but probabilistic analysis is not always considered meaningful, as often a basis for the numbers is lacking. Using imprecise probabilities and strength of knowledge judgments represents a possible way of extending the system modeling and analysis.

Finally, in this Section 4.2, a comment on the causality concept. Philosophers have discussed the concept for centuries. As mentioned in Section 3.4, Leveson [35] understands a cause as events and conditions which are sufficient for the accident to occur. From such a definition, as mentioned in Section 3.4, we cannot conclude that smoking causes cancer, as smoking does not always lead to cancer. Safety research/science is, however, very much concerned with drawing conclusions of this type. People need to know whether, for example, eating a specific food or participating in a specific activity is safe. The question is whether a specific dose or exposure is dangerous. When this cannot be concluded deterministically, we are led to risk judgments. The question then becomes whether the dose or exposure leads to an unacceptable risk. But what does 'lead to' mean? It is not sufficient to just require a correlation between the input and the output, as the correlation could be explained by other factors (confounding variables). One solution is to define causality in this way [10]: We can say that, for B to cause A (for example, smoking to cause lung cancer), at a minimum, B must precede A, the two must covary (vary together), and no competing explanation can better explain the correlation between A and B. Hence, the issue of causality will always be dependent on argumentation and justification, leaving it open to a possible better explanation in the future.

4.3. Safety and risk management

The resilience engineering and management of Safety II focuses on how the system (organization) functions as a whole, and four abilities are commonly highlighted [25,28]: how it responds to events (changes, disturbances, and opportunities), how it monitors what is happening (including its own performance), how it anticipates risk events and opportunities, and how it learns.

For the present discussion in this paper, this raises several questions:

- What is the link between resilience management and risk management (and risk science)?
- To what extent does resilience management build on risk science and risk management knowledge?
- How can resilience management be enhanced by incorporating such knowledge?

The term 'resilience management' is used here instead of the more commonly applied term 'resilience engineering', as the discussion extends beyond engineering contexts.

These issues have been partly discussed in Aven [7,9]. Here, we summarize and extend some key points made in these two papers.

First, according to contemporary risk science, resilience and resilience management are key elements of risk management [9,48,51]. The central risk science argument for highlighting resilience is that events (changes, disturbances, and opportunities) occur, and this risk needs to be properly handled, to avoid serious consequences. Strengthening the resilience means reduced risks. Of special importance are potential surprises and the unforeseen. The risk assessments and the related management have limitations, as the knowledge that these assessments are built on could be more or less strong and even wrong.

Risk management knowledge is also needed for resilience management. To illustrate, consider the example of ventilators used to help the breathing process for patients in relation to the COVID-19 pandemic in 2020. Clearly, a large number of available ventilators would increase the resilience in such a situation and, consequently, reduce the vulnerabilities. There is, however, a cost associated with the ventilators, and, before the occurrence of the pandemic, these costs would need to be balanced against risk. Thus, investments in resilience cannot be seen in isolation from risk considerations. With hindsight, it is easy to draw the conclusion that more ventilators should have been available when needed, but there are many 'demands' for reduced vulnerabilities and improved resilience, and the resources are limited. Prioritizing is a management and political task. Risk assessments and risk science in general provide support for this type of prioritization [12].

Practice shows that resilience management and risk management are often separate activities, with minimal interactions on scientific and professional matters [9]. This is an unfortunate situation, as the problems addressed need both risk and resilience-based thinking and methods. To define and communicate this management and science, terms like risk & resilience management and science could be used [9].

Safety I is associated by Hollnagel [25] with a traditional, 'narrow' perspective on risk, characterized by historical data and probabilities, where control is sought by identifying all risks and ensuring that the related accident scenario probabilities are sufficiently small. The approach is suitable for well-established systems, with limited

uncertainties, but not for other types of systems. Current practice is not well described by Safety I, as commented by Leveson [35]. Robustness and resilience are given due attention. Yet, it can be argued that current practice and related standards fail to fully capture the message brought forward by Safety II on resilience management. A main obstacle is the risk perspective. To allow for and encourage resilience-based thinking, one must take a different view on risk assessments than is traditionally used. A shift is needed from accurate risk estimation to improving the understanding of risk. There will always be qualitative aspects to report beyond quantification. Potential surprises may occur relative to the knowledge reflected by the risk assessments. Systems may be considered non-complex, overlooking critical aspects of the system. Acknowledging these limitations of risk assessment means giving weight to resilience-based thinking and analysis, as risk control cannot be ensured by risk assessments and their follow-up.

Contemporary risk science has acknowledged this [10,48,51]. Considerable research and development have been carried out in recent years to enhance the risk assessments and the risk management, to better take into account the knowledge dimension, particularly surprises and the unforeseen (black swans) (see e.g. [4,41,42]). In this way, there is a potential for stronger interactions between safety science and risk science.

Although the safety concept of Safety III is not built on uncertainties and risk, risk assessment is seen as an integral part of the safety control structures [35]. Using contemporary risk science, there is a potential to strengthen the role of risk assessments in relation also to Safety III and STAMP, as outlined in Bjerga et al. [16].

Hollnagel refers to Safety I as reactive and Safety II as proactive (Section 3.7). The reasoning is that Safety I reacts to failures and accidents, whereas Safety II focuses on adjusting performance to keep things working. The picture is however more nuanced. As commented by Leveson [35], Safety I builds on risk assessments which are proactive instruments. In for example probabilistic risk assessments (quantitative risk assessments), a main goal is to identify *potential* failures or deviations, which may not have been based on historical events.

Both Safety II and Safety III are based on a systems approach. Risk management and risk science recognize the importance of having such a view to system understanding and analysis and, hence, for understanding and characterizing risk in relation to safety applications. Risk management and risk science here build on what is current scientific knowledge within safety science and research on system modeling and analysis.

In relation to Safety II and resilience-based perspectives, the issue of what 'success' and recovery (sustaining required operations) mean can be discussed – should the definitions include aspects of system improvements? The common perspective is that system performance improvements over time could be a goal, but such a goal should not be an aspect of the safety and resilience concepts as such. In line with this thinking, it is, for example, meaningful to combine resilience and antifragile-based policies, as the antifragility concept acknowledges and stimulates stressors to improve the system over time [5,37,46,53]. The basic idea of the antifragility concept is that we must accept, even appreciate, some level of stressors, failures and mistakes, in order to obtain better performance over a longer time horizon. Just as our bodies and minds need stressors to be in top shape and improve, so do other systems and activities [5]. However, some authors explicitly incorporate aspects of improvement in the definitions; for example, Meerow et al. [38] define urban resilience in this way:

> Urban resilience refers to the ability of an urban system-and all its constituent socio-ecological and socio-technical networks across temporal and spatial scales-to maintain or rapidly return to desired functions in the face of a disturbance, to adapt to change, and to quickly transform systems that limit current or future adaptive capacity [38].

The improvement aspect is linked to the system transformation process strengthening the adaptive capacity.

Hollnagel and Leveson discuss what we learn most from: success or failure. Safety science provides important knowledge on the issue, but so do many other fields, including quality management, systems theory and general learning theory (e.g., [1,15,19,25]). It is outside the scope of the present paper to discuss the issue, but the present author finds the evidence supporting the view that most learning derives from mistakes to be strong. Who has not personally experienced the importance of failures for future development and progress? The importance of failure is also reflected by the antifragility concept discussed above. This does not mean that focus should be on failures only. As highlighted above, risk science welcomes alternative approaches and the safety science research highlighting success adds important knowledge to how to best manage safety and risk.

5. Conclusion

Risk is, to a limited degree, addressed in the literature concerning Safety I, II and III. The present paper argues that there is a potential for further development of these safety perspectives, by integrating risk science knowledge concerning concepts, principles, approaches and methods. The present paper has pointed to some of these. Table 2 summarizes the main conclusions. A key point is that safety cannot be meaningfully defined, assessed and managed without taking into account risk. When safety science research discusses risk, it is usually by reference to traditional probabilistic perspectives which are not in line with contemporary risk science knowledge. Clearly safety science can be strengthened by building on this knowledge. In contrast to the traditional probabilistic perspectives, modern risk science concepts and principles provide a well suitable framework for understanding, characterizing, communicating and handling safety for all types of applications as due considerations are given to uncertainty, potential surprises and robustness/resilience. Risk assessments where the aim is improved risk understanding in line with current risk perspectives, can provide useful decision support also for intractable and complex systems. There is a potential for incorporating uncertainty and knowledge considerations in Safety II and III types of models, leading to enhanced understanding of risk and performance. The safety and risk fields and sciences are closely related, and the future development of each of them depends on contributions from the other. In recent years we have seen a considerable number of contributions stressing the importance of strengthening the foundations of both safety science (e.g. [18,32,45]) and risk science [6,14,49–51], with strong calls also for integration of theories and practices. The present paper demonstrates the need and

Table 2
Summary of main conclusions.

Theme	Conclusions
The safety and risk concepts	Safety cannot be meaningfully defined without relating the concept to risk. Safety should be considered the antonym of risk or, if used in the meaning "safety is achieved", as absence of unacceptable risks
Risk assessment and modeling	Safety II and Safety III provide system models that supplement linear causal models, and these models are important for assessing intractable and complex systems. These models can be enhanced by incorporating uncertainty and knowledge considerations. Risk assessments, where the aim is improved risk understanding, can provide useful decision support, also for intractable and complex systems
Safety and risk management	Contemporary risk management and risk science build on resilience-based thinking and methods to meet future events: both expected and surprising events. Risk and resilience management need to be better integrated, to ensure and improve performance and avoid undesirable events.

importance for this type of work and efforts.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The author is grateful to two anonymous reviewers for their useful comments and suggestions to earlier versions of this paper.

References

[1] Ackoff RL. Ackoff's best: his classic writings on management. New York: John Wiley & Sons; 1999.
[2] Aven T. Safety is the antonym of risk for some perspectives of risk. Saf Sci 2009;47:925–30.
[3] Aven T. What is safety science? Saf Sci 2014;67:15–20.
[4] Aven T. Risk, surprises and black swans. New York: Routledge; 2014.
[5] Aven T. The concept of antifragility and its implications for the practice of risk analysis. Risk Anal 2015;35(3):476–83.
[6] Aven T. Risk assessment and risk management: review of recent advances on their foundation. Eur J Oper Res 2016;25:1–13. Open Access.
[7] Aven T. How some types of risk assessments can support resilience analysis and management. Reliab Eng Syst Saf 2017;167:536–43.
[8] Aven T. Reflections on the use of conceptual research in risk analysis. Risk Anal 2018;38(11):2423–5.
[9] Aven T. The call for a shift from risk to resilience: What does it mean? Risk Anal 2019;39(6):1196–203.
[10] Aven T. The science of risk analysis. New York: Routledge; 2020.
[11] Aven T. Bayesian analysis: critical issues related to its scope and boundaries in a risk context. Reliab Eng Syst Saf 2020;204:107209.
[12] Aven T, Thekdi S. Risk science: an introduction. New York: Routledge; 2021.
[13] Aven T, Ylönen M. A risk interpretation of sociotechnical safety perspectives. Reliab Eng Syst Saf 2018;175:13–8.
[14] Aven T, Zio E. Globalization and global risk: How risk analysis needs to be enhanced to be effective in confronting current threats. Reliab Eng Syst Saf 2021;205:107270.
[15] Bergman B, Klefsjö B. Quality. 2nd ed. Lund, Sweden: Studentlitteratur; 2003.
[16] Bjerga T, Aven T, Zio E. Uncertainty treatment in risk analysis of complex systems: the cases of STAMP and FRAM. Reliab Eng Syst Saf 2016;156:203–9.
[17] Bjørnsen K, Jensen A, Aven T. Using qualitative types of risk assessments in conjunction with FRAM to strengthen the resilience of systems. J Risk Res 2018;23(2):153–66.
[18] Dekker S. Foundations of safety science. Boca Raton, FL: CRC Press, Taylor & Francis group; 2019.
[19] Deming WE. The new economics. 2nd ed. Cambridge, MA: MIT CAES; 2000.
[20] Fischhoff B. Risk perception and communication unplugged: twenty years of process. Risk Anal 1995;15:137–45.
[21] Flage R, Aven T, Baraldi P, Zio E. Concerns, challenges and directions of development for the issue of representing uncertainty in risk assessment. Risk Anal 2014;34(7):1196–207.
[22] Haavik TK. Debates and politics in safety science. Reliab Eng Syst Saf 2021;210:107547.
[23] Hollnagel E. Barriers and accident prevention. Aldershot, UK: Ashgate; 2004.
[24] Hollnagel E. The functional resonance analysis method for modeling complex socio-technical systems. Farnham: Ashgate; 2012.
[25] Hollnagel E. Safety-I and safety-II. London: CRC-Press; 2014.
[26] E. Hollnagel (2016) Resilience engineering. https://erikhollnagel.com/ideas/resilience-engineering.html. Accessed March 2, 2021.
[27] Hollnagel E, Wears RL, Braithwaite J. From safety-I to safety-II: a white paper. The resilient health care net. Australia: University of Southern Denmark, University of Florida, USA, and Macquarie University; 2015. Published simultaneously by the, https://www.england.nhs.uk/signuptosafety/wp-content/uploads/sites/16/2015/10/safety-1-safety-2-whte-papr.pdf.
[28] Hollnagel E, Woods D, Leveson N. Resilience engineering: concepts and precepts. UK: Ashgate; 2006.
[29] Kleiner BM, Hettinger LJ, Dejoy DM, Huang YH, Love PED. Sociotechnical attributes of safe and unsafety work systems. Ergonomics 2015;58(4):635–49. Apr3https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4566878/#cit0022.
[30] Le Coze JC. Vive la diversité! high reliability organisation (HRO) AND resilience engineering (RE). Saf Sci 2019;117:469–78.
[31] Le Coze JC. Safety science research. evolution, challenges and new directions. Boca Raton, FL: CRC Press, Taylor & Francis group; 2019.
[32] Le Coze JC, Pettersen K, Reiman T. The foundations of safety science. Saf Sci 2014;67:1–70.
[33] Leveson N. A new accident model for engineering safer systems. Saf Sci 2004;42:237–70.
[34] Leveson N. Engineering a safer world: systems thinking applied to safety. Cambridge, MA: The MIT Press; 2016.
[35] Leveson N. Safety I–II, resilience and antifragility engineering: a debate explained through an accident occurring on a mobile elevating work platform. Int J Occup Saf Ergon 2020;25(1):66–75.
[36] MacInnis DJ. A framework for conceptual contributions in marketing. J Mark 2011;75(4):136–54.
[37] Martinetti A, Chatzimichailidou MM, Maida L, van Dongen L. Safety I–II, resilience and antifragility engineering: a debate explained through an accident occurring on a mobile elevating work platform. Int J Occup Saf Ergon 2019;25(1):66–75.
[38] Meerow S, Newell JP, Stults M. Defining urban resilience: a review. Landsc Urban Plan 2016;147:38–49.
[39] Mill JS. A system of logic, ratiocinative, and inductive: being a connective view of the principle of evidence, and methods of scientific inquiry. London: J.W. Parker; 1843.
[40] Mohaghegh Z, Kazemi R, Mosleh A. Incorporating organizational factors into probabilistic risk assessment (PRA) of complex socio-technical systems: a hybrid technique formalization. Reliab Eng Syst Saf 2009;94:1000–18.
[41] Paté-Cornell E, Cox, Jr A. Improving risk management: from lame excuses to principles practice. Risk Anal 2014;34(7):1228–39.
[42] Paté-Cornell ME. On black swans and perfect storms: risk analysis and management when statistics are not enough. Risk Anal 2012;32(11):1823–33.
[43] Patriarca R, Gravio GD, Woltjer R, Costantino F, Praetorius G, Ferreira P, Hollnagel E. Framing the FRAM: a literature review on the functional resonance analysis method. Saf Sci 2020;129:104827.
[44] Provan DJ, Woods DD, Dekker SWA, Rae AJ. Safety II professionals: how resilience engineering can transform safety practice. Reliab Eng Syst Saf 2020;195:106740.
[45] Rae A, Provan D, Aboelssaad H, Alexander R. A manifesto for reality-based safety science. Saf Sci 2020;126:104654.
[46] Ramezani J, Camarinha-Matos LM. Approaches for resilience and antifragility in collaborative business ecosystems. Technol Forecast Social Chang 2020;151:119846.
[47] Rasmussen J. Risk management in a dynamic society: a modeling problem. Saf Sci 1997;27(2/3):183–213.
[48] Renn O. Risk governance: coping with uncertainty in a complex world. London: Earthscan; 2008.
[49] SRA (2015) Glossary Society for Risk Analysis. https://www.sra.org/resources. Accessed February 23, 2021.
[50] SRA (2017) Core subjects of risk analysis. Society for Risk Analysis. https://www.sra.org/resources. Accessed February 23, 2021.
[51] SRA (2017) Risk analysis: fundamental principles. Society for Risk Analysis. https://www.sra.org/resources. Accessed February 23, 2021.
[52] Swuste P, Gulijk CV, Zwaard W, Oostendorp Y. Occupational safety theories, models and metaphors in the three decades since World War II, in the united states, britain and the netherlands: a literature review. Saf Sci 2012;62:16–27.
[53] Taleb NN. Anti fragile. London: Penguin; 2012.
[54] Turner B, Pidgeon N. Man-made disasters. 2nd ed. London: Butterworth-Heinemann; 1997.