Risk management in human-in-the-loop AI-assisted attention aware systems

Max Nicosia and Per Ola Kristensson
DEPARTMENT OF ENGINEERING, UNIVERSITY OF CAMBRIDGE, CAMBRIDGE, CAMBRIDGESHIRE, UNITED KINGDOM

1 Introduction

AI-assisted systems are becoming more pervasive across a variety of fields. The level of assistance ranges from partial assistance to fully autonomous AI control. In the former, the system assists the operator, so they perform at their highest level under a variety of conditions. In the latter, no input from the operator is necessary. Examples of partial assistance include airport baggage threat detection, which uses machine learning classifiers, and attention aware systems, which employ sensors and AI to monitor operator behavior and assist them where possible. Fully automated AI systems can be found in several sectors, such as aircraft, transport, and medicine.

Any complex system, whether it uses AI or not, is susceptible to risks. Risks need to be assessed and managed. As a consequence, many risk assessment and risk management strategies have been developed over the years. These strategies have been developed to identify, capture, and manage risks in various ways. The reason for a plurality of methods is that no single strategy is capable of capturing and managing all risks in all situations.

For instance, failure mode and effects analysis (FMEA) is one of the first developed risk assessment methods. A failure mode is the manner in which something fails to fulfill its function. The method was originally described in US Armed Forces Military Procedures document MIL-P-1629 [1] in 1949. By the 1960s, NASA [2] was using variants of FMEA in their programs (e.g., Apollo [3], Viking, Voyager, Magellan, Galileo [4], and Skylab [5]) and by the 1970s, it had spread to petroleum exploration [6], the automotive industry [7], wastewater treatment plants [8], and the food industry [9].

Another widely used risk assessment method is fault tree analysis (FTA). FTA was developed in 1962 by Bell Labs to evaluate ballistic launch control systems [10]. Boeing incorporated FTA in their civilian aircraft design process in 1966 [11]. In 1970, the US Federal Aviation Administration (FAA) incorporated FTA into the CFR §25.1309 airworthiness regulations for aircraft. The FAA later extended its use to other areas within the US National

Putting AI in the Critical Loop. https://doi.org/10.1016/B978-0-443-15988-6.00013-3

Airspace Systems [12]. NASA considered using FTA in their Challenger program but decided against it due to the calculations resulting in unacceptably low reliability values. Instead, they favored the continuation of qualitative risk assessment methods, including the previously mentioned FMEA method. This decision proved to be a major oversight after the Challenger accident. As a consequence, NASA reconsidered the importance of using quantitative risk assessment methods and resumed their use, including the use of FTA [13].

Another popular method is event tree analysis (ETA). ETA was developed as an alternative to FTA. Since ETA makes an assumption on systems units either working or failing, it makes the analysis more manageable [14]. As ETA is used to analyze specific events, it allows the identification of all sequences of events and failures that can occur as a result of an originating event and its subsequent events.

FMEA, FTA, and ETA have evolved from the need to address different challenges presented by various types of systems with varying inherent complexities. History has demonstrated that failing to assess and manage risks can lead to catastrophic failures, such as the Challenger accident. However, risk management techniques can be difficult to transfer across domains as each domain has different requirements. This is particularly difficult for systems involving AI, as they have a level of automation that may not be possible to directly control. To tackle this problem, Falco et al. [15] proposed the creation of a regulatory body that would standardize risk management techniques and enforce compliance as a way to improve responsibility and control of AI-assisted systems.

Attention aware systems are a relatively new class of AI-assisted systems, and their risks are currently not well understood. In this work, we want to draw attention to how risk management strategies can be applied to attention aware systems. Our motivation is to avoid history repeating itself.

We first explain what an attention aware system is and what its capabilities are. We then discuss the importance of managing and mitigating risks for such systems. We later discuss various considerations that should be taken into account and give recommendations on how to develop effective risk management strategies. This work is a first step to understanding how to manage risks in human-in-the-loop AI-assisted attention aware systems.

2 Attention aware systems

The main purpose of an attention aware system is to detect operators' focus of attention and manage it in such a way that the system can ensure optimal task operation under a variety of conditions, such as when operators are experiencing fatigue and/or extraordinarily high workloads.

Fig. 6.1 shows an example of a functional architecture of an attention aware system. The dashed rounded rectangle represents the system boundary. Arrows entering the boundary are input signals from other systems and arrows leaving the boundary are output signals from the system. In this particular functional architecture, the functionality of

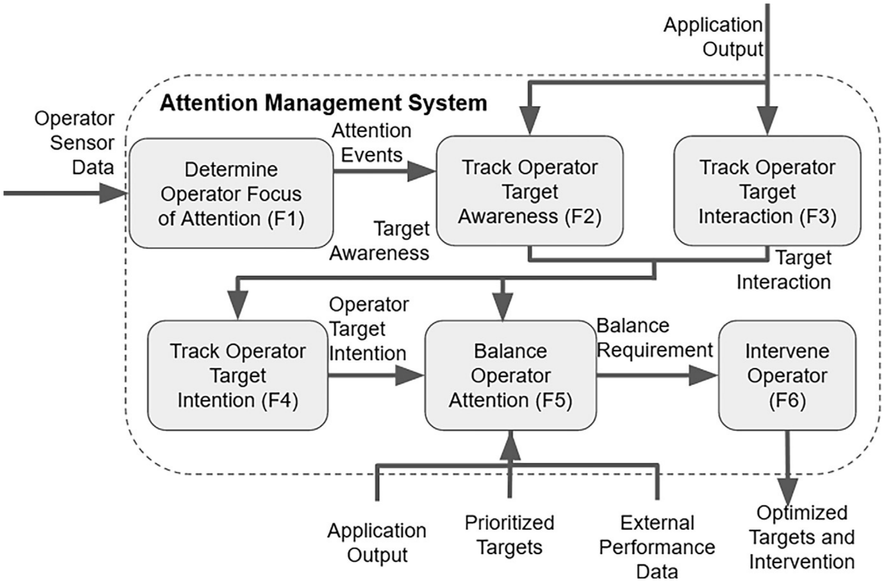
Application Output
Attention Management System
Operator Sensor Data
Attention Events
Determine Operator Focus of Attention (F1)
Track Operator Target Awareness (F2)
Track Operator Target Interaction (F3)
Target Awareness
Target Interaction
Operator Target Intention
Track Operator Target Intention (F4)
Balance Operator Attention (F5)
Balance Requirement
Intervene Operator (F6)
Application Output
Prioritized Targets
External Performance Data
Optimized Targets and Intervention

FIG. 6.1 Functional structure of an attention aware system.

the system has been decomposed into six functional units represented by gray rounded rectangles.

The function `Determine Operator Focus of Attention` parses operator sensor data and generates attention events. The function `Track Operator Target Awareness` combines suitable application output and attention events to detect target awareness. The function `Track Operator Target Interaction` collates target interactions from the output provided by the application. The function `Track Operator Target Intention` collates and analyzes target awareness and target interaction information to estimate potential future operator actions. The function `Balance Operator Attention` combines target awareness, target interaction, and target intention information with the application output, target priorities, and task performance data to build and evaluate the current application and operator state in order to output a balance requirement signal. Finally, the `Intervene Operator` function processes a balance requirement signal by turning it into a sequence of instructions containing the targets and interventions that the application needs to deploy. For a more detailed explanation of these functions, see prior work [16].

From the preceding description it is clear that, depending on the particular implementation and application domain, the complexity of each of the previous functions can vary greatly. This, in turn, will affect the difficulty of analyzing and managing the risks of that particular deployment of the attention management system.

The most complex function in the system is `Balance Operator Attention`. It works by processing signals containing operator and application states to construct an overall operator-application state. It then evaluates this state to detect drops in task performance or any other operator aspect that may require additional attention and infers the source causing a drop in performance. There are several failure modes that can arise in such a function. Two of the most obvious ones are arriving at an incorrect state because signals from previous functions failed in some way, and/or incorrect aggregation and/or assessment of the overall state due to parameterization errors.

Another complex function is `Intervene Operator`. This function uses the information in the balance requirement signal to address the sources of poor performance identified by the previous component. To achieve this, it devises an intervention strategy that assists the operator in returning to an acceptable level of operating performance. An intervention strategy consists of coordinated instructions where, for example, the saliency of certain information presented on a display is manipulated in such a way that the operator's focus of attention is directed either toward it or away from it. However, the mechanisms used to influence the operator introduce their own failure modes. For example, saliency changes can appear confusing to the operator or the behavior prompted by the system can be misunderstood by the operator.

Another important aspect consists of the objectives behind interventions. As previously stated, the primary purpose of the system is to manage the operator's focus of attention such that the operator can maintain an appropriate task performance level. However, the system can leverage the operator's focus of attention in a variety of ways depending on the behavior the system wishes to induce in the operator. For example, the system can intervene with the operator to ensure that they balance their actions between all tasks. As a result of stress, fatigue, or misunderstandings, the operator can neglect some of their duties in favor of others. This can lead to failures in monitoring important information or periodically carrying out less frequent tasks.

Another objective of interventions can be to ensure that operators use their time or actions more effectively. For example, the system can help the operator focus on higher-priority targets or tasks. However, for this to be possible, the system needs to understand the tasks to potentially identify more optimal ways of performing them. In other cases, poor task performance may be caused by the operator misunderstanding or misinterpreting information. This can happen due to, for example, the state of the operator, such as being under stress or experiencing fatigue, or due to cognitive overload. In such situations, the system needs to either drive the operator's focus of attention toward relevant information or lower the amount of disruption caused by the information being presented to the operator. Depending on the circumstances, choosing one mechanism over another can lead to additional failure modes.

Finally, the system can manage attention to preempt potential errors by precluding the operator from performing certain actions by estimating the intention of the operator. However, such detection mechanisms need to be reliable within the application domain or they can lead to further unexpected operator behavior and potential failure modes.

An attention management system can be seen as a type of human-machine teaming in which the AI tries to affect operator behavior in order to influence their task performance.

Moreover, as the system is continuously monitoring the operator and application state, it can use this state information to adapt its level of influence accordingly. This adaptation capability means that the system can vary its level of automation during operation. Within the 10-point system automation model of Parasuraman et al. [17], an attention management system alternates between levels 1, 2, and 3 during operation, where level 1 provides no assistance, level 2 makes the system provide all possible alternative actions/decisions to the operator, and level 3 reduces the number of decisions for the operator. However, while level 3 automation reduces the number of decisions for the operator by changing the saliency of certain targets, it still permits the operator to remain in full control as all decision options continue to be visible and accessible to the operator at all times. Ensuring that the system manages the switch between the automation levels is a strength of the system's capabilities, but also a source of risk.

3 Risk management of attention aware systems

Risk management is carried out to ensure risks are understood and under control in order to prevent risks from materializing into incidents. We define risk as the likelihood of undesirable behavior in the joint human-machine system multiplied by the impact of such undesirable behavior happening.

Hubbard [18] defines risk management as: "The identification, assessment, and prioritization of risks followed by coordinated and economical application of resources to minimize, monitor, and control the probability and/or impact of unfortunate events." For reference, risk analysis is the systematic identification of potential sources of harm—hazards—and their risks. Risk management is the entire process of identifying hazards and subsequently analyzing and controlling risks. In other words, risk management encompasses hazard identification, risk estimation, risk evaluation, risk control, and risk monitoring.

Before a risk management strategy can be developed, it is first necessary to perform a risk analysis. The first step in risk analysis is to define the system boundary around the joint human-machine system. We will only consider risks within the boundary. In practice, deciding on a system boundary requires careful reflection, as it is vital any that relevant factor to eventual risk is captured within the system boundary. Having set the system boundary, the system within the boundary can be mapped out so that it is fully understood. This can, for example, involve decomposing the system's overall function into function structures and constructing diagrams indicating the flow of signals and other information.

Having set the system boundary and mapped out the system, it is possible to identify hazards and estimate risks. However, this is not always straightforward, as we discuss later on. Another challenge is quantifying risks in terms of their impact and their frequency of occurrence. Assessing their impact can be difficult, as it may not always be known what other undesirable behavior may manifest as a result of a single risk materializing into an undesirable event. Assessing individual risks is important, as it is critical to be able to understand the level of risk in subsystems and the system as a whole in order to determine an overall acceptable level of risk.

Having identified all hazards and estimated the risks, these risks now need to be managed—monitored and controlled—throughout the operating life of the joint human-machine system. Risk management methods provide established processes for achieving this and they may be qualitative or quantitative. When developing a risk management strategy for an AI-assisted joint human-machine system, both types of methodologies may be required. In general, systematic approaches are required, as risk assessment depends on human judgment and humans have a tendency to underestimate or overlook risks.

Attention aware systems are AI-assisted joint human-machine systems that involve multiple stakeholders. They depend on multiple external systems and require extensive parameterization to function correctly. Since each specific deployment has its own associated specific domain risks, the risk of experiencing undesirable behavior is very high.

In the event of undesirable behavior, the repercussions can affect multiple stakeholders and it may be difficult to attribute responsibility. Let us consider the following simple scenario: the system reduces the saliency of a piece of information that needs to be considered by the operator during a specific task and, as a result of this, the operator fails to perceive the information and does not complete the task in a timely manner. In such a situation, it is not obvious where to attribute responsibility or determine the root cause. The information was always visible to the operator, and while the system increased the saliency of alternative information, or reduced the relevant information's saliency, these actions ultimately amounted to the system incorrectly managing the operator's focus of attention. Nevertheless, throughout this process, the operator still remained in full control, having access to all relevant information.

To ameliorate such scenarios, we will draw attention to some strategies that can be considered when managing the risks of such systems and highlight potential specific risks that may arise in deployment.

4 Risk management considerations

Our purpose is to draw attention to some considerations when developing a risk management strategy for attention management systems. As such, we focus mainly on the potential sources of risks, the stakeholders involved, and the importance of evaluating the effectiveness of the risk management strategy.

As explained in Section 2, an AI-assisted attention aware system is in between a fully manual system (no automation) and a fully autonomous system (full automation). A key difference compared to a fully autonomous system is that operators remain in full control of operation (i.e., there is no action automation). An AI-assisted attention aware system manipulates the presentation of information the operator receives without direct operator control.

Successful operation of an AI-assisted attention aware system will depend primarily on three factors: (1) domain context including task complexity; (2) the joint human-machine system's parameterization; and (3) factors relating to the individual operator, or team of operators, such as skill level, experience, and behavior.

A complication that is unique to such a human-in-the-loop system is that the AI-assisted functionality can possibly generate actions that are unthinkable or otherwise confusing to human operators, such as hiding relevant information from operators because the AI-assistive function disagrees with the operator on its relevance. We briefly mentioned this when we gave an example of a failure scenario in Section 3.

While the risk of any AI-assisted attention aware system is coupled to a particular system configuration and domain, we have identified four general sources of risk that are likely to require substantial attention when risk managing such systems. These are (1) system failures, (2) incorrect parameterization, (3) operator characteristics, and (4) domain context characteristics.

Some risks can arise individually or out of interactions between the stakeholders involved. The stakeholders of such a system include the operators, the people responsible for parameterizing the system, any entity that is directly or indirectly controlled by the operator through the system, the organization providing the system, and the organization managing and employing the operator.

System failures are all failures that involve the system not operating as intended despite being correctly parameterized. This can involve both software failures, such as program bugs, and hardware failures, such as loose cables or sensor errors. As described in Section 2, the system depends on four incoming signals: (1) the external performance data, (2) the output of the application, (3) the prioritized targets, and (4) the operator sensor data. Each of these signals can fail or operate at different reliability levels. Any risk management strategy needs to consider the levels of possible operation and their impact on overall system operability.

Parameterization errors can lead to undesirable behavior as a result of the system being incorrectly configured. For example, depending on the metrics the system is monitoring and the thresholds it uses as the basis for intervention deployment decisions, the system may react in completely different ways under the same circumstances. Similarly, the mechanisms set up for manipulating target saliencies will need to match the context and requirements of the application domain. For example, simple color mismatches can result in catastrophic failure situations with operators misunderstanding the importance of the information they are presented with.

Operator characteristics are another general source of risk. The operator can be out-of-phase with the system, which can result in undesirable behavior as a result of the operator misunderstanding the system's interventions or the system misunderstanding the operator's actions. The former situation can happen if the operator is unaware of the specifics of an intervention, despite the system being correctly parameterized. The latter situation can arise due to a variety of reasons, including parameterization errors, system failures, or unexpected or unknown operator behavior.

Domain context characteristics are another source of risk, as such characteristics can cause the system to react in an unexpected manner to unaccounted inputs and thus result in the system generating undefined behavior. Examples of risks associated with domain context characteristics include targets changing their visual appearance (e.g., if a new type of mine is encountered by a submarine demining vehicle), unexpected task complexities, such as a sudden rapid explosion of targets on a display, or a sudden unexpected situation, such as an unexpected incident situation in an air-traffic-control system.

Further, interactions between stakeholders can cause many undesirable outcomes. These may range from miscommunication to an inability to maintain proper records about what parameterizations are required for deployment. In Section 5, we discuss some diagrammatic methods of representing various aspects of information circulation and communication to help identify any associated risks.

Finally, once risk assessments have been carried out and mitigation strategies have been developed, it is important to monitor and evaluate the efficacy of risk management. Since the behavior of AI-assisted attention aware systems depends on both specific deployment situations and operators, new risks may arise from such system dynamics. It is also possible that emergent behavior caused by interventions may give rise to new unanticipated risks. As such, it is necessary to continuously monitor and evaluate risks throughout the lifetime of the system.

5 Risk management approaches

Risk management consists of carrying out the following tasks: (1) identifying the system boundary, (2) mapping the system, (3) identifying hazards, (4) assessing risks, and (5) devising suitable strategies to manage/mitigate risks.

However, not all of these tasks can always be carried out in this order. For example, sometimes the system boundary is not clear until all relevant components and subsystems have been identified. Therefore, it is necessary to first understand the system and all its components and subsystems. This process is known as system mapping.

Diagrams are very helpful for system mapping. Such diagrams can, for example, capture the people involved and their functions, the information flowing between people, and the system or the organization that the system is embedded within. Organizational diagrams are useful for determining the people involved, their roles, and their relationships with other people. Organizational diagrams can also help with understanding the scope and determining the boundaries of all involved entities. Information diagrams can be used to capture the relationships between documents in the system. Communication diagrams reveal how information flows between stakeholders and other entities in the system. The latter can be of particular importance if system parameterization changes need to be requested or acted upon. These diagrams also provide context for the system operator's role in the entire communication structure of the organization.

Once the system has been mapped out it is possible to assess risks. There are many well-established methods for risk assessment. In this work, we focus on three methods, which were briefly introduced earlier.

FMEA is a method for identifying potential failure modes in products or processes and devising corrective measures to address the resulting associated risks. A failure mode is the manner in which a system, mechanism, or component fails to fulfill its function. Once a failure mode has been identified, its effect, severity, and cause can be determined. FMEA involves identifying the issues related to failure modes, ranking them by their importance, and devising corrective measures for issues with serious concerns. In addition to identifying failure modes, FMEA helps with establishing failure rates and root causes of known failures. The advantage of using FMEA is that it is good at systematically cataloging all possible sources of failure. As such, an FMEA is ideal for collecting information and exchanging it between teams. This helps with the early identification of potential sources of failure. FMEA is not designed to demonstrate how robust a system is to multiple failures, or failures due to external events. Another weakness FMEA shares with many other risk assessment methods is that an FMEA can become too large to be effectively maintained and understood.

FTA is used to trace a failure path to identify all the events that lead to said failure. Fault trees are normally represented graphically, with each event connected through logic gates. Tracing events also allows for identifying the components involved in the failure. Once the components and the events that lead to the failure have been identified, it is possible to develop a strategy to prevent it from reoccurring. FTA is a top-down approach that allows the mapping of the dependencies of each event and the calculation of the probabilities of specific failures, provided that the probabilities of the events involved are known. An advantage of FTA is that it allows analyzing the effects of initiating faults, which is the opposite of FMEA. Additionally, FTA allows analyzing multiple complex failures taking place at the same time. An FTA will consider external events, which is useful for assessing how robust a system is against single and multiple failures. A disadvantage, however, is that FTA is not suitable for finding all possible initiating faults.

ETA is used to determine the probability of a specific event occurring based on the probability of the chronological sequence of events leading up to it. ETA is useful for determining the effect that a particular failure can have on the overall system. The approach is inductive and follows a bottom-up approach. An advantage of ETA is that it enables the assessment of multiple simultaneous functions in both failure and success states. This is useful, as events do not need to be anticipated as they are only the starting point. An ETA is also useful for identifying single sources of failures and system paths that can lead to failure. Additionally, an ETA is suitable for modeling complex systems, as it can visualize cause-and-effect relationships. Finally, an ETA allows tracing faults across boundaries of subsystems. A disadvantage of ETA is that it always starts with one initiating event at a time, which must be identified before commencing the analysis. Another disadvantage is that partial successes and failures are not accounted for.

As with any new technology, AI-assisted attention aware systems present new challenges in terms of configuration and deployment. Organizations deploying such systems need to make sure that they can manage evolving requirements and potential new sources of undesirable outcomes. Most importantly, they need to ensure that appropriate mechanisms are in place to mitigate possible unexpected undesirable outcomes.

To ensure that risk management strategies are effective, they need to be continuously evaluated by systematic monitoring and evaluation of risk at all levels in the system. Exactly how this is to be carried out is dependent on the precise application domain. Nevertheless, ensuring reliable accounting of all registered failures and undesirable events, as well as continuous monitoring of risk rates in the system, is critical to fully understanding the level of risk in the system and the efficacy of any active risk management strategies. Additionally, monitoring the effects of mitigation procedures can also assist system evaluation. Finally, ensuring that the cost and effort of each strategy are yielding a comparable risk reduction benefit is of utmost importance for ensuring organizations optimize their use of available resources.

6 Discussion and conclusions

Applying risk management strategies to AI-assisted attention aware systems provides many benefits. First, it allows early detection of potential human-machine system failures and the creation of appropriate mitigation and monitoring strategies to eliminate, reduce, and track the effects of such failures. Mitigation strategies can range from improving the design of a system to increasing redundancy, which simply incorporates a verification mechanism that prevents miscommunications or errors.

A second benefit is that risk management strategies can provide assurance that some minimum level of safety has been considered. This is of particular benefit for organizations as they want to ensure that their products are reliable and safe and their reputations maintained. Moreover, in some domains, implementing appropriate risk management strategies is part of the obligations necessary for regulatory compliance or adherence to ethical standards.

As explained in this chapter, AI-assisted attention aware systems are human-in-the-loop systems with multiple sources of failure. In addition, they are often meant to be used in safety-critical operations. They present challenges similar to those presented by fully autonomous AI systems, as well as additional risks caused by the reliance on human operators and the resultant human-AI interaction. As with any AI system designed for decision-making, the adaptive nature of the system can make it difficult to identify the root cause of a fault or undesired outcome. For example, the operator can fail to perceive a target due to a system intervening in a way that ends up obfuscating the target. While such a target is always visible to the operator, the system intervention actively diverts the operator's attention away from the target.

We also highlight that identifying risk in AI-assisted attention aware systems can be challenging. This is because it is not always evident what the source of a particular failure or undesirable event is, or how frequently it occurs. We advocate a systematic approach and, as a first step, suggest three widely used risk assessment strategies: FMEA, FTA, and ETA. These methodologies provide both quantitative and qualitative analysis mechanisms. FMEA is suitable for devising corrective measures for all possible faults. ETA is useful for identifying what other failures can occur as a result of specific undesirable events or failures. FTA focuses on identifying the factors that lead to certain faults.

To help the discovery of risks in AI-assisted attention aware systems, Section 4 presents four general sources of risks. These are (1) system failures, (2) incorrect parameterization, (3) operator characteristics, and (4) domain context characteristics. The first source involves failures in software or hardware. The second source encompasses configuration errors that fundamentally affect overall system behavior. The third source accounts for the operator failing to understand the system's interventions or the system not understanding the operator's behavior. The fourth source relates to the joint human-machine system failing to manage unexpected changes or other aspects of the application domain itself.

We predict AI-assisted attention aware systems will be increasingly critical as human-in-the-loop systems become increasingly sophisticated and prominent in application domains ranging from security and air-traffic control to manufacturing and healthcare. However, such systems also introduce additional complexities, which give rise to additional risk. This chapter is a first step toward tailored risk management approaches for such systems.

References

[1] USDD, MIL-P-1629-Procedures for Performing a Failure Mode Effect and Critical analysis, 1949.

[2] R.A. Neal, Modes of failure analysis summary for the Nerva B-2 reactor, Astronuclear Laboratory Westinghouse Electric Corporation, 1962. Technical Report WANL-TNR-042.

[3] Office of Manned Space Flight, Apollo Program, Apollo Reliability and Quality Assurance Office, Procedure for Failure Mode, Effects and Criticality Analysis (FMECA), NASA, Washington, DC, 1966.

[4] NASA, Failure Modes, Effects and Criticality Analysis (FMECA), Practice no. PD-AP-1307 (1999).

[5] NASA, Experimenters' Reference Based Upon Skylab Experiment Management, Technical Memorandum (TM) NASA-TM-X-72397 (1974).

[6] M.K. Dyer, D.G. Little, E.G. Hoard, A.C. Taylor, R. Campbell, Applicability of NASA Contract Quality Management and Failure Mode Effect Analysis Procedures to the USGS Outer Continental Shelf Oil and Gas Lease Management Program, National Aeronautics and Space Administration, 1972.

[7] K. Matsumoto, T. Matsumoto, Y. Goto, Reliability analysis of catalytic converter as an automotive emission control system, SAE Trans. 84 (1975) 728–738.

[8] C.W. Mallory, Application of Selected Industrial Engineering Techniques to Wastewater Treatment Plants, vol. 1, US Government Printing Office, 1973.

[9] W.H. Sperber, R.F. Stier, Happy 50th birthday to HACCP: retrospective and prospective, Food Saf. Mag. (2009) 42–46.

[10] C.A. Ericson, C. Ll, Fault tree analysis, System Safety Conference, Orlando, Florida, vol. 1, 1999, pp. 1–9.

[11] A.F. Hixenbaugh, Fault tree for safety, Boeing Co Seattle WA Support Systems Engineering, 1968. Technical Report.

[12] Federal Aviation Authority, System Safety Handbook, 2000.

[13] M. Stamatelatos, W. Vesely, J. Dugan, J. Fragola, J. Minarick, J. Railsback, Fault Tree Handbook With Aerospace Applications, National Aeronautics and Space Administration, Washington, DC, 2002.

[14] P.L. Clemens, R.J. Simmons, System Safety and Risk Management: NIOSH Instructional Module, US Department of Health and Human Services, 1998.

[15] G. Falco, B. Shneiderman, J. Badger, R. Carrier, A. Dahbura, D. Danks, M. Eling, A. Goodloe, J. Gupta, C. Hart, et al., Governing AI safety through independent audits, Nat. Mach. Intell. 3 (7) (2021) 566–571.

[16] M. Nicosia, P.O. Kristensson, Design principles for AI-assisted attention aware systems in human-in-the-loop safety critical applications, in: Engineering Artificially Intelligent Systems, Springer, 2021, pp. 230–246.

[17] R. Parasuraman, T.B. Sheridan, C.D. Wickens, A model for types and levels of human interaction with automation, IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. 30 (3) (2000) 286–297.

[18] D.W. Hubbard, Healthy skepticism for risk management, in: The Failure of Risk Management, Chapter 1, John Wiley & Sons Ltd, 2020, pp. 1–19, https://doi.org/10.1002/9781119521914.ch1.