Proceedings of the 35th European Safety and Reliability & the 33rd Society for Risk Analysis Europe Conference
Edited by Eirik Bjorheim Abrahamsen, Terje Aven, Frederic Bouder, Roger Flage, Marja Ylönen
 Published by Research Publishing, Singapore.
doi: 10.3850/978-981-94-3281-3_ESREL-SRA-E2025-P0734-cd

Artificial Intelligence and Major Accident Risk: Towards a Framework for Systemic Safety Risk Assessment

Arne Jarl Ringstad

Safety Technology Unit, Equinor ASA, Norway. E-mail ajri@equinor.com

Alexandra Fernandes

Safety Technology Unit, Equinor ASA, Norway. E-mail: almm@equinor.com

Jan Tore Ludvigsen

Safety Technology Unit, Equinor ASA, Norway. E-mail: jtl@equinor.com

Prevention of major accidents is fundamental in all safety critical operations. History shows us that the introduction of new technology may create novel safety risks. Today, fatalities associated with artificial intelligence (AI) are reported in manufacturing, healthcare and transportation. In the energy sector, AI is being increasingly used in an operational context where the potential for major accidents is considerable. The introduction of AI is typically done with a strong focus on information security and civil rights risks, but there is limited systematic focus on major accident scenarios. This is partly because our current methods and approaches are not designed to do so, partly because of the risk focus in the AI domain is on other potential outcomes. Also, AI is a diverse field, with a range of industrial applications, from preventive maintenance and diagnosis, autonomous systems like robots and drones, to decision-making and operator support systems. This paper reports from a project that focused on the inclusion of AI applications in existing approaches to major accident prevention. A core aspect of this work is the safety management concept and AI applications' potential roles in safety barrier management. The result is a risk-based framework that can be used in the development, introduction and application of AI in different contexts, and that can increase the awareness and understanding of AI as a factor in major accident risk assessments. The framework has a systemic focus emphasizing the interplay between technical, operational and organizational factors.

Keywords: Artificial intelligence, Major accident risk, Safety barrier management, Oil and gas

1. Introduction

Safety-critical systems, such as healthcare, transportation, and oil and gas production, refer to contexts where accidents may result in fatalities, serious injuries, significant property damage, or environmental harm (Smith & Simpson, 2020). We argue that the current surge in artificial intelligence (AI) development might generate upsides but also pose new challenges in these environments. Limited understanding of the associated risk landscape and its effects on major accident prevention can complicate AI models' integration in the management of safety critical systems.

1.1. AI and major accident risk

AI has been more often presented in literature as a safety contributor (e.g. Gursel *et al.*, 2025; Perez-Cerrolaza *et al.*, 2024; Tamascelli *et al.*, 2024) rather than a potential risk factor. The literature has analysed how AI models could contribute to increased safety and risk management in critical systems. However, only in recent years has an interest in potential risks of AI itself been present in literature (e.g. Salmon *et al.*, 2024; DNV 2024). These approaches have been stimulated by both the higher levels of deployment of AI-based solutions, and by legal and regulatory efforts to frame the use of such technologies in a rapidly changing background (e.g., European Union, 2024). There is a growing interest on AI risks and AI safety from a technical perspective

(example of creating guardrails for LLMs; Dong *et al.*, 2024), a human-centred perspective (mostly realized through design concepts of trustworthy AI; Kaur *et al.*, 2022), as well as a societal perspective (ethical AI initiatives; Ortega-Bolaños *et al.*, 2024). However, risk management approaches within AI have focused mostly either on cybersecurity or ethical/ moral/societal values views, overlooking safety aspects and risk notions commonly used in safety critical systems.

The current risk approaches tend to analyse AI models in isolation, focusing on validation of the model itself (e.g. Neto *et al.*, 2022), ignoring the context of application and its integration with pre-existing routines – aspects that have been essential in risk management approaches in complex systems.

At the same time, proposed risk management approaches seem to present an over-reliance on humans/users/operators as the ultimate barrier for safe AI deployment with the underlying assumption that human oversight and meaningful control are possible, feasible, and need to be granted in these systems (e.g. Enqvist, 2023; Kyriakou & Otterbachter, 2023).

In recent years, the awareness towards inherent risks in AI models has increased both in literature and the society in general. Experiences from the deployment of AI systems that demonstrated gross biases (e.g., Barocas & Selbst, 2016; Cobert-Davies *et al.*, 2023) elicited an ethical, legal, and overall societal reflection on how these systems work, are designed, and are deployed.

In 2024, the European Union released the Artificial Intelligence Act, the first regulatory framework specifically designed for AI systems (European Union, 2024). The AI Act defines four risk levels: unacceptable, high risk, limited risk, and minimum risk, and based on this classification there are different expectations on mitigation measures that go from banning systems in the unacceptable risk category to the obligation of informing users that an AI system is in use in the minimum risk assessment. The concept of risk within the EU AI Act is linked to perceived threat to societal and human values and thus mismatched from the concept of risk in critical systems. Although high-risk assessments in the AI Act contemplate for instance, the application of AI for critical infrastructure such as energy production, it potentially addresses a narrow scope of the risk outlook in such contexts.

The Organisation for Economic Cooperation and Development (OECD) has also developed guidelines for managing AI risk (OECD, 2024) and created an (AI-powered) AI Incidents Monitoring tool. OECD's approach is broad, including a classification of AI systems according to five dimensions: people and planet; economic context; data and input; the AI model itself; and the task and output (OECD, 2022). The goal with this framework is mostly to support policymakers, regulators, and legislators to assess risks and opportunities that different types of AI systems present. We argue therefore that this approach, such as the EU AI Act, is mostly suited to address the civil and ethical impacts of AI systems.

The National Institute of Standards and Technology (NIST) in the U.S. Department of Commerce has presented a comprehensive AI Risk management framework (NIST, 2023) meant to support organizations and individuals to achieve trustworthiness in AI systems, reducing potential negative impacts of AI as well as maximizing positive impacts. This framework presents an exhaustive description of the AI systems with playbooks with specific actions on governance, management, mapping and measuring of AI according to both AI actors and topics resulting in a vast matrix with hundreds of entries.

Both the OECD and NIST classification approaches propose a strategy of wide-ranging description of the factors influencing the model and the model use, supporting further analysis or decision-making on the use of AIs, with an underlying assumption that more information will result in better control or at least awareness of the potential risks.

Within the standards and regulatory bodies internationally (and nationally) there have also been recent attempts to analyse the implications of new types of AI models (e.g. ISO & IEC, 2023), with a particular interest towards generative AIs (e.g. NIST, 2024).

In general, current approaches to AI as a potential hazard in major accident scenarios seem to face one of two challenges: While some approaches are simplistic and risk assessment is primarily understood as the testing of a system's predictive abilities in laboratory-like conditions, other approaches are so detailed and comprehensive that ensuring compliance and risk control in practice becomes very challenging, particularly in a context of fast paced technological development. In both cases, risk is understood in an abstract level, whether by focusing on technical specificities and capabilities of the models or focusing on high-level values and principles the models will need to comply to. This results in a disconnection to the specific operational contexts where the AI applications will be used.

1.2. Scope and objectives

In this paper we focus on potential risk factors for AI application within safety critical contexts. We adopt a systemic approach to human-machine interaction, considering technical, operational and organisational factors. Lessons-learned from human-automation collaboration are integrated in the framework. As such, our main objectives are to:

- Develop an approach to AI technology within the context of major accident prevention in the oil and gas industry
- Show how this approach strengthens the focus on relevant aspects of the operational context
- Show how systemic mitigation strategies can be based on this approach

2. Major accident risk and safety barriers

According to Aven (2012) risk can be described as specified consequences (including risk sources, events/scenarios, and effects), a measure of uncertainty for these specified consequences, and the associated knowledge base. The scenarios in focus here are major accidents. Safety science has established a rich conceptual basis and knowledge about systematic approaches to the management of these scenarios.

In the Norwegian legislation related to offshore safety, a major accident means an acute incident such as a major spill, fire or explosion that immediately or subsequently entails multiple serious personal injuries and/or loss of human lives, serious harm to the environment and/or loss of major financial assets (Havtil, 2023).

2.1. Safety barrier management

Safety barriers are established to manage risk and prevent an event from occurring or escalating into an accident or incident with serious, harmful consequences. In Norway, legislation requires oil and gas companies to establish, manage, and maintain safety barriers in a holistic manner (Havtil, 2017).

Figure 1 illustrates Equinor's approach to major accident prevention with safety barriers as a core element. Safety barriers typically consist of combinations of technical- and operational barrier elements required to fulfil a barrier function (Equinor, 2024). A common method in the oil and gas industry has been to use this type of "bow-tie models" as a visualization tool for risk and barrier management. The models are a graphical representation of both proactive measures or barriers to avoid a specific event/ incident (on the left side), and then reactive measures or barriers put in place in the case of the event (reactive measures) to control the incident, mitigate its consequences and/or avoid escalation (e.g. CCPS, 2018). As such, bow-tie models are a good representation of both plausible accident scenarios linked to one pre-defined event, and a good tool to identify control measures the companies can set in place across several levels of safety and security in the context of the existing organisational structure.

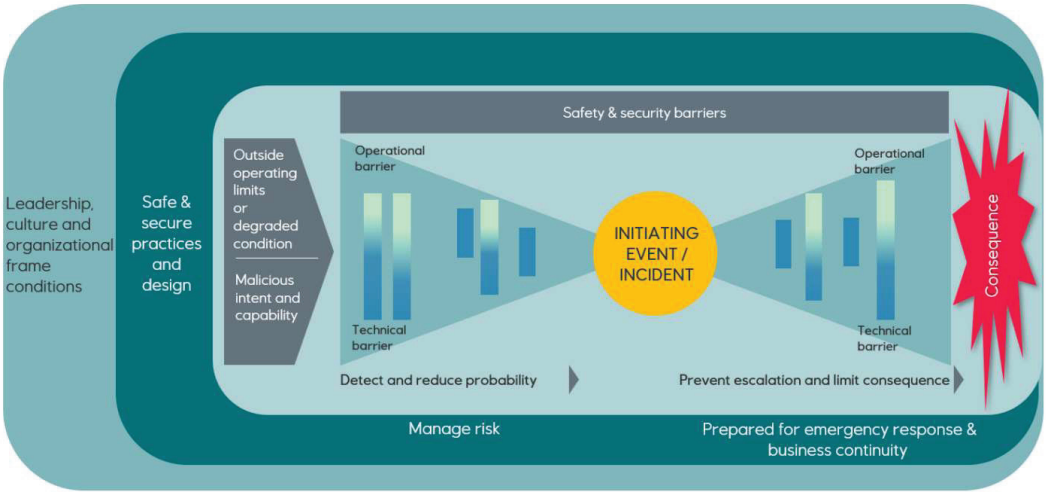

Fig. 1 Equinor's bow-tie model

A technical barrier element is an engineered system, structure, or other design feature which is intended to prevent, detect, control, or mitigate a hazardous event. The technical safety barriers should as far as practicable be independent and resilient to failure in other systems and barriers.

Operational barrier elements are safety-critical tasks performed by an operator, or team of operators. They will normally support activation of one or several technical barrier functions. Safety-critical tasks are typically related to initiation, prevention, detection, response, control, or mitigation of the development of a hazard. Operational barriers depend on and are embedded in the organizational structures.

It is important to note that both technical and operational barrier elements will be defined specifically for each installation/facility and are thus to large degree context dependent.

3. AI and safety barriers

An AI application may have different roles in a barrier management system:

AI can function directly as a technical safety barrier element. An example of this is an AI application that intervenes in a major accident scenario (e.g. by shutting down equipment).

AI may be considered as a performance shaping factor for technical barrier elements. An AI application that collects data from different parts of a facility can monitor the integrity of technical barrier elements and can use pattern recognition to detect anomalies and unforeseen combination of factors that may represent the initiation of a major accident.

AI may also work as performance shaping factors for safety critical tasks – or operational barrier elements. An AI application that supports operators in the performance of safety critical tasks can improve the integrity of operational barrier elements. For instance, an application that diagnosis the status in a facility may aid operators in a situation with alarm flooding in the control room.

However, the introduction of new technology in a safety barrier system will also introduce potential pitfalls, including:

Technical factors. These factors are linked to known risks regarding the AI model itself. For instance, predictive error has been a well explored factor linked to the discrepancies between the model's predictions and actual outcomes (e.g. Zhang *et al.*, 2022); temporal

degradation and model drift are other examples referring to potential gradual decline of model performance, as input data deviates from training data with time (Vela *et al.*, 2022); training data sets and their limitations are another source of risk (e.g. Mohammed *et al.,* 2024) given unrepresentative or incomplete training data (for instance, in predictive monitoring there is high availability of "normal operation" data and low availability of "abnormal or incident" data, which will affect model performance for detection of non-ordinary patterns). On the other hand, development options such as the choice of the model, or choice of reward functions can lead to suboptimal performance or unintended behaviour of the model, reducing its reliability and predictability (e.g. Dayal *et al.*, 2022).

Interaction factors. These factors refer to the quality and efficiency of the human-AI interaction that will be crucial for overall system performance. As mentioned earlier, human oversight, is a central assumption in most risk management approaches today for critical application of AI models (Kyriakou & Otterbachter, 2023). Loss of human oversight will result in lack of adequate monitoring/control of the systems performance leading to errors or unintended behaviour (e.g. Sterz *et al.*, 2024), linked to lack of oversight, the notion of human out of the loop can be referenced to in situations where human's capacity of intervention is hindered when they are excluded from the core control tasks (e.g. Gómez-Carmona *et al.*, 2024). Other well documented risks in interaction relate to for instance over-reliance on AI systems (e.g. Klingbeil *et al.*, 2024), or low trust in the systems leading to its rejection (Afroogh *et al.*, 2024). Another relevant effect is the analysis of the combined performance of human-AI systems and which benchmark it can be measured against (for instance, performance assessment of model alone, human alone, or human and model together; Vaccaro *et al.*, 2024). The types of tasks that the model can take over are also of relevance due to its implications on the engagement of the human in the overall control process (e.g. Deranty & Corbin, 2024).

Organisational factors. Can refer to for instance, poor work design, and implications for role definition as integration of AI models could create confusion about expectations and responsibilities in concrete tasks (e.g. Schlicht *et al.*, 2021). The lack of definition of redundancy measures and work design related to the AI integration is a risk that might result in lower operational resilience and efficiency. Training of the operators, both on how to use the AI systems and supporting the creation of mental models on its functioning is relevant, as well as the potential risk for de-skilling of operators in tasks that are taken over by the AI systems (Morandini, *et al.*, 2023). Regarding leadership and governance there is a risk of loss of accountability due to the use of AI systems (Papagiannidis *et al.*, 2025); simultaneously lack of intentional planning of aspects such as resource allocation for maintenance of the AI models throughout its life cycle could increase the risks of technical factors affecting performance negatively. A central risk is strategic misalignment, whereby organizational goals need to be efficiently integrated in the AI models to avoid long-term disconnect between the organizations' core goals and expectations and the objectives pursued by the AI models – in order for AI to bring value it needs to result from a well-defined need of the organisation and/or be a means to achieve the organisation's goals (e.g. Christian, 2020).

Summing up, it seems clear that introducing AI applications to a barrier system may have both upsides and downsides on barrier integrity. It is therefore critical to maintain a balanced view. To quote Havtil (2025): AI is *also* a risk factor.

4. Discussion and conclusion

The approach outlined in this paper has several features that address challenges discussed in the introduction:

- The safety (major accident related) assessment of AI technologies is performed in accordance with established principles and concepts developed in safety science. Thus, the assessments are not performed from scratch but rather included as one more element in standard safety approaches.

  This means that risk mitigation can be performed in accordance with principles already in effect. The strategy is – including AI applications or not – "…to establish and

maintain barriers so that the risk faced at any given time can be handled by preventing an undesirable incident from occurring or by limiting the consequences should such an incident occur." (Havtil, 2013, p. 1).

- The assessment is sensitive to operational conditions. The approach to safety barriers described above is installation/operation specific. This lessens the criticism raised against laboratory testing of AI-applications and/or assessments with very general and comprehensive compliance criteria.
- The approach is systemic. Technical, human and organisational factors are directly included in the description of technical and operational safety barrier elements.
- The approach encompasses both potential upsides and downsides of the introduction of AI in a safety critical system.
- The integrity of safety barriers is dependent on a number of factors outside the barriers per se (see Figure 1), 1: organisational capabilities/conditions: including operational capacity, competence and management focus, and 2: design, practice and risk management AI technology will also influence these factors, and it would be possible to extend the current approach outlined here to include these additional factors.

As mentioned, there are aspects of AI technology that are challenging in a safety management context. Perhaps the most important is the dynamic (and sometimes unpredictable) nature of some applications. While in traditional automation paradigms, the software was designed for one task and would maintain the benchmark/ qualified performance level and characteristics, this will not be the case with AI models which are able to learn, adapt, and self-improve both in expected and positive ways, but also in potentially negative or unexpected fashion, impacting the overall system performance. This is a general feature of some AI applications, so it is beyond this paper to address it. However, it does underline the need for controlled introduction of AI in a safety related context, and the need for a technology qualification process that is life cycle based.

The holistic and systemic features of the proposed approach have similarities to current conceptions of resilience engineering (Patriarca, 2021). We believe this can be a fruitful approach as AI technology is entrenching work in high-risk organisations at an increasing rate.

4.1 Conclusion

This paper has described the fundamental building blocks for managing major accident risk associated with the use of AI in high-risk contexts. More work is needed to specify practical tools and methods for applying this approach in practice. However, by including AI applications in current analysis tools and methods it is possible to utilise existing safety practices to control major accident risk also in this case.

5. References

Afroogh, S., Akbari, A., Malone, E. et al. Trust in AI: progress, challenges, and future directions. *Humanit Soc Sci Commun 11*, 1568 (2024). https://doi.org/10.1057/s41599-024-04044-8

Aven, T. (2012). The risk concept—Historical and recent development trends. *Reliability Engineering & System Safety, 99*, 33–44. https://doi.org/10.1016/j.ress.2011.11.006

Barocas, S. & Selbst, A. D. (2016). Big Data's Disparate Impact. *104 California Law Review* 671 http://dx.doi.org/10.2139/ssrn.2477899

CCPS - Center for Chemical Process Safety Energy Institute (EI) (2018). *Bow Ties in Risk Management*. Hoboken, N.J.: John Wiley & Sons. ISBN 9781119490388

Christian, B. (2020). *The alignment problem: Machine learning and human values*. W. W. Norton & Company.

Corbett-Davies, S., Gaebler, J. D., Nilforoshan, H., Shroff, R., & Goel, S. (2023). The measure and mismeasure of fairness. *The Journal of Machine Learning Research*, 24(1), 14730-14846. 1808.00023

Dayal, A., Cenkeramaddi, L. R., & Jha, A. (2022). Reward criteria impact on the performance of reinforcement learning agent for autonomous navigation. *Applied Soft Computing, 126*, 109241. https://doi.org/10.1016/j.asoc.2022.109241

Deranty, JP., & Corbin, T. (2024). Artificial intelligence and work: a critical review of recent research from the social sciences. *AI & Soc 39*,

675–691 https://doi.org/10.1007/s00146-022-01496-x

DNV (2024) Kunnskapsoversikt knyttet til forsvarlig bruk av kunstig intelligens i petroleumssektoren. DNV Report 2024-1519, Rev. 1. https://www.havtil.no/contentassets/ef58508a2e4641aebba4091811795020/dnv-2024-1519-kunnskapsoversikt-forsvarlig-bruk-ki-petroleumssektoren-rev-1.pdf

Dong, Y., Mu, R., Jin, G., Qi, Y., Hu, J., Zhao, X., Meng, J., Ruan, W., & Huang, X. (2024). Building guardrails for large language models. Proceedings of the 41st International Conference on Machine Learning (ICML 2024). https://doi.org/10.48550/arXiv.2402.01822

Enqvist, L. (2023). 'Human oversight' in the EU artificial intelligence act: what, when and by whom? *Law, Innovation and Technology*, *15*(2), 508–535. https://doi.org/10.1080/17579961.2023.2245683

Equinor (2024). *Framework for major accident prevention*. Unpublished presentation.

European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending certain Union legislative acts (Artificial Intelligence Act). Official Journal of the European Union, L 168, 1–60. http://data.europa.eu/eli/reg/2024/1689/oj

Gómez-Carmona, O., Casado-Mansilla, D., López-de-Ipiña, D., & García-Zubia, J. (2024). Human-in-the-loop machine learning: Reconceptualizing the role of the user in interactive approaches. *Internet of Things, 25,* 101048. https://doi.org/10.1016/j.iot.2023.101048

Gursel, E., Madadi, M., Coble, J. B., Agarwal, V., Yadav, V., Boring, R. L., & Khojandi, A. (2025). The role of AI in detecting and mitigating human errors in safety-critical industries: A review. *Reliability Engineering & System Safety, 256*, 110682. https://doi.org/10.1016/j.ress.2024.110682.

Havtil (2013). Principles for barrier management in the petroleum industry. https://www.havtil.no/contentassets/11851dc03a84473e8299a2d80e656356/principles-for-barrier-management-in-the-petroleum-industry_2013.pdf.

Havtil (2017). Principles for barrier management in the petroleum industry *BARRIER MEMORANDUM 2017*. https://www.havtil.no/contentassets/43fc402b97e64a7cbabdf91c64b349cb/barriers-memorandum-2017-eng.pdf

Havtil (2023). The management regulations. https://www.havtil.no/en/regulations/all-acts/?forskrift=611

Havtil (2025). Artificial intelligence is also a risk factor https://www.havtil.no/en/explore-technical-subjects2/main-issue-2024/

International Organization for Standardization (ISO), & International Electrotechnical Commission (IEC). (2023). ISO/IEC 42001:2023 - Information technology — Artificial intelligence — Governance and management of AI systems. https://www.iso.org/standard/81230.html

Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2022). Trustworthy Artificial Intelligence: A Review. *ACM Computing Surveys, 55(2),* Article 39, 38 pages. https://doi.org/10.1145/3491209

Klingbeil, A., Grützner, C., & Schreck, P. (2024). Trust and reliance on AI: An experimental study on the extent and costs of overreliance on AI. *Computers in Human Behavior, 160*, 108352. https://doi.org/10.1016/j.chb.2024.108352

Kyriakou, K., Otterbacher, J. In humans, we trust. *Discov Artif Intell 3,* 44 (2023). https://doi.org/10.1007/s44163-023-00092-2

Mohammed, S., Budach, L., Feuerpfeil, M., Ihde, N., Nathansen, A., Noack, N., Patzlaff, H., Naumann, F., & Harmouch, H. (2024). The effects of data quality on machine learning performance on tabular data. arXiv preprint arXiv:2207.14529. https://doi.org/10.48550/arXiv.2207.14529

Morandini, Sofia & Fraboni, Federico & De Angelis, Marco & Puzzo, Gabriele & Giusino, Davide & Pietrantoni, Luca. (2023). The Impact of Artificial Intelligence on Workers' Skills: Upskilling and Reskilling in Organisations. *Informing Science. 26*. 39-68. 10.28945/5078.

National Institute of Standards and Technology (NIST). (2023) *Artificial Intelligence Risk Management Framework* (AI RMF 1.0). United States Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1

National Institute of Standards and Technology (NIST). (2024) *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile* (NIST AI 600-1). United States Department of Commerce. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

Neto, A. V. S., Camargo, J. B., Almeida, J. R., & Cugnasca, P. S. (2022). Safety assurance of artificial intelligence-based systems: A systematic literature review on the state of the art and guidelines for future work. *IEEE Access, 10*, 130733–130770. https://doi.org/10.1109/ACCESS.2022.3229233

Organization for Economical Cooperation and Development (OECD). (2022). The OECD Framework for the Classification of AI systems.

https://wp.oecd.ai/app/uploads/2022/02/Classification-2-pager-1.pdf

Organization for Economical Cooperation and Development (OECD). (2024). *Assessing potential future artificial intelligence risks, benefits, and policy imperatives.* OECD Artificial Intelligence Papers, No. 27. DSTI/CDEP/AIGO(2023)13/FINAL

Ortega-Bolaños, R., Bernal-Salcedo, J., Germán Ortiz, M. et al. Applying the ethics of AI: a systematic review of tools for developing and assessing AI-based systems. *Artif Intell Rev 57*, 110 (2024). https://doi.org/10.1007/s10462-024-10740-3

Patriarca, R. (2021). Resilience Engineering for Sociotechnical Safety Management. In M. Ungar (ed.) *Multisystemic Resilience*. Oxford University Press. https://doi.org/10.1093/oso/9780190095888.003.0025

Papagiannidis, E., Mikalef, P., & Conboy, K. (2025). Responsible artificial intelligence governance: A review and research framework. *The Journal of Strategic Information Systems, 34(2),* 101885. https://doi.org/10.1016/j.jsis.2024.101885

Perez-Cerrolaza, J., Abella, J., Borg, M., Donzella, C., Cerquides, J., Cazorla, F. J., Englund, C., Tauber, M., Nikolakopoulos, G., & Flores, J. L. (2024). Artificial intelligence for safety-critical systems in industrial and transportation domains: A survey. *ACM Computing Surveys, 56(7),* Article 176, 40 pages. https://doi.org/10.1145/3626314

Salmon, P. M., King, B. J., Elstak, I., McLean, S., & Read, G. J. M. (2024). Tomorrow's demons: A scoping review of the risks associated with emerging technologies. *Ergonomics.* https://doi.org/10.1080/00140139.2024.2416554

Schlicht, Larissa & Melzer, Marlen & Roesler, Ulrike & Voß, Stefan & Vock, Silvia. (2021). An Integrative and Transdisciplinary Approach for a Human-Centered Design of AI-Based Work Systems. 10.1115/IMECE2021-71261.

Smith, D. J., & Simpson, K. G. L. (2020). The safety critical systems handbook: A straightforward guide to functional safety: IEC 61508 (2010 edition), *IEC 61511 (2016 edition)* and related guidance (4th ed.). Butterworth-Heinemann.

Sterz, S., Baum, K., Biewer, S., Hermanns, H., Lauber-Rönsberg, A., Meinel, P., & Langer, M. (2024). On the quest for effectiveness in human oversight: Interdisciplinary perspectives. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '24). https://doi.org/10.48550/arXiv.2404.04059

Tamascelli, N., Campari, A., Parhizkar, T., & Paltrinieri, N. (2024). Artificial intelligence for safety and reliability: A descriptive, bibliometric and interpretative review on machine learning. *Journal of Loss Prevention in the Process Industries, 90*, 105343. https://doi.org/10.1016/j.jlp.2024.105343

Vaccaro, M., Almaatouq, A. & Malone, T. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nat Hum Behav 8*, 2293–2303 (2024). https://doi.org/10.1038/s41562-024-02024-1

Vela, D., Sharp, A., Zhang, R. et al. Temporal quality degradation in AI models. *Sci Rep 12*, 11654 (2022). https://doi.org/10.1038/s41598-022-15245-z

Zhang, X., Chan, F. T. S., Yan, C., & Bose, I. (2022). Towards risk-aware artificial intelligence and machine learning systems: An overview. *Decision Support Systems, 159,* 113800. https://doi.org/10.1016/j.dss.2022.113800