Proc. of the European Safety and Reliability Conference (ESREL 2026)
Edited by José C. Matos, Paulo B. Lourenço, Michael Beer, Daniel Oliveira,
Edoardo Patelli, Hélder S. Sousa, Rui A. Silva and Monica Santamaria
 Published by Research Publishing, Singapore.
doi: 10.3850/ESREL2026061419_esrel26-p26253-cd

Beyond Component Failures: A System-Theoretic Hazard Identification for AI-Enabled Systems

Niclas Flehmig

Department of Mechanical and Industrial Engineering, Norwegian University of Science and Technology, Norway. E-mail: niclas.flehmig@ntnu.no

Mary Ann Lundteigen

Department of Engineering Cybernetics, Norwegian University of Science and Technology, Norway.

Shen Yin

Department of Mechanical and Industrial Engineering, Norwegian University of Science and Technology, Norway.

The deployment of Artificial Intelligence (AI) systems into safety-critical systems offers the potential to improve safety, but also introduces novel hazards. Traditional hazard identification methods, focused on component-level failures and chain of failure events, are often insufficient for these complex AI-enabled systems where unsafe component interactions are a primary concern for accidents. Recent research advocates conducting hazard identification for AI-enabled systems with a system-theoretic approach to address unsafe component interactions of system components and traditional component failures. This study adopts a system-theoretic perspective, employing system-theoretic process analysis (STPA) to investigate operational safety challenges in a liquid hydrogen (LH_2) bunkering system that incorporates an AI system for predicting the violation of the lower flammability limit (LFL) of hydrogen. Through our analysis, we identify the hazards and loss scenarios of two different system designs for such an AI-enabled safety-critical system. Furthermore, we derive benefits, limitations and propose enhancements to the STPA methodology itself to better address the characteristics of AI-enabled systems. Our findings contribute to advancing hazard identification methods and risk assessment frameworks, promoting the safe integration of AI systems in safety-critical systems.

Keywords: STPA, AI-enabled systems, safety-critical systems, hazard identification, risk assessment, liquid hydrogen bunkering.

1. Introduction

The integration of AI systems into safety-critical contexts introduces novel hazards arising from both the intrinsic complexity of the AI systems and the intricate system environment. This introduces novel hazards that challenge traditional, component-failure-based hazard identification techniques (Dobbe, 2022). The system-theoretic process analysis (STPA) method provides a top-down framework for this complexity, as it is designed to address interactions among technical systems, human operators, and organizational processes (Leveson, 2012). Despite this potential, the practical application of STPA specifically to AI-enabled safety-critical systems is not well established. This paper addresses this gap by investigating STPA's implementation on two distinct system designs within a real-world case study, evaluating its benefits and limitations and deriving necessary methodological adaptations for this domain.

2. Definitions

2.1. *Safety*

Safety is defined as the absence of unacceptable risk (ISO/IEC Guide 51, 2014). Leveson (1995) adopted this definition to the absence of accidents, i.e., unplanned events leading to unacceptable human, financial, or mission-related loss. This is the underlying definition of safety for STPA.

2.2. *AI systems*

Following the EU AI Act (EU AI Act, 2024), an AI system is defined as a machine-based system exhibiting varying degrees of autonomy in performing tasks that traditionally require human intelligence, such as learning, prediction, recommendation, problem-solving, natural language processing, and decision-making.

2.3. *AI-enabled systems*

An AI-enabled system is a system that incorporates one or more AI systems (Det Norske Veritas, 2023).

3. Research Approach

This work analyzes the application of STPA to AI-enabled systems to derive benefits, limitations, and potential methodological adjustments. The analysis applies STPA to two system designs based on architectural patterns based on the ISO/IEC TR 5469 (ISO/IEC TR 5469, 2024): one directly from the technical report and an extended version (Flehmig et al., 2024). This comparative approach tests the method across different system designs. Both architectural patterns were instantiated within a LH_2 bunkering system context, whose foundational design is based on established maritime and hydrogen safety studies (European Maritime Safety Agency, 2024; Maritime Technologies Forum, 2024; Tamburini et al., 2025). In this context, the AI system functions as part of the overall safety-related system and more specifically as part of a safety instrumented function (SIF). The AI system uses sensor measurements to generate enhanced information for the bunkering control room (BCR). It predicts for a given sample whether a certain hydrogen concentration, the lower flammability limit (LFL), is surpassed after 200 s. This time frame was chosen based on the water systems' response time after a hydrogen release (Alfarizi et al., 2023).

3.1. *Case study: LH_2 bunkering system*

A LH_2 bunkering system, a process where ships are refueled from a stationary tank was selected as it represents a hazardous process with need for a safety system (Maritime Technologies Forum, 2024; European Maritime Safety Agency, 2024; Tamburini et al., 2025). For the bunkering operation, Alfarizi et al. (2023) proposed an AI system for accident prevention of LH_2 leakage. The AI system predicts whether the LFL will be exceeded within a 200-second horizon. This forecast enables the proactive activation of targeted safety barriers, such as water curtains, sprinkler systems, inert gas injection, or enhanced ventilation, to mitigate the hazard and thereby prevent the need for a full emergency shutdown of the bunkering operation. While the prior work does not detail the integration of such an AI system into an LH_2 bunkering system and its safety-related system, we propose utilizing the AI system's prediction to active safety barriers and thus prevent costly emergency shutdown.

3.2. *Architectural patterns for AI-enabled systems*

The following outlines the two architectural patterns used in this work. The first pattern, used in "System 1", is adopted directly from the referenced technical report. The second pattern, for "System 2", represents an extended architecture synthesized by integrating elements from all patterns described in the same report plus additional components.

3.2.1. *System 1*

The technical report (ISO/IEC TR 5469, 2024) considers three different possible patterns: (I) a safety backup system with non-AI technology and a detection method to switch the control when unsafe behavior by the AI system is detected, (II)

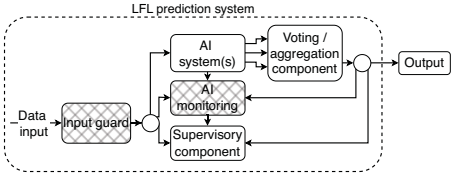

Fig. 1. Architectural pattern of System 2 adapted from Flehmig et al. (2024). The cross-hatched components are the extensions. The non-cross-hatched component represent System 1.

a limiting logic which is determined by a supervisory component to ensure a safe subset of action space, and (III) an AI system redundancy with output voters or aggregators. These patterns can co-exist or are sufficient just by themselves. We adopted architectural pattern (III) and adapted it to our use case, where the AI system functions not as a safety controller, but as a sensor enhancement that provides detailed predictive information to the BCR. Consequently, the limiting logic and the dedicated safety backup system were omitted from the design. This adapted architecture is depicted in Figure 1 using non-cross-hatched components. In this pattern, process data is fed into parallel AI systems, that are differently weighted instances of the same model. For regression tasks, the outputs are aggregated and for classification, a majority vote determines the final system output.

3.2.2. *System 2*

System 2 is an extended version based on the patterns in ISO/IEC TR 5469. As shown in Figure 1 and introduced by Flehmig et al. (2024), it consists of the three pattern suggestions in the technical report and described in Section 3.2.1. Originally, this pattern includes an independent monitor, an input guard and an offline updating component. The monitor tracks inputs, internal states, and outputs for unreliable behavior, alerting the supervisory component of any deviations. The input guard applies rule-based validations to ensure that all inputs adhere to predefined operational bounds, detecting anomalies such as missing values or obvious outliers, and verifies software and hardware dependencies. Any violations are reported to the supervisory component. The offline updating ensures that the performance in terms of reliability of the AI system is according to defined requirements. This would include data collection, re-training of the AI system, testing and verification and finally updating the operating AI system. This is an offline process to ensure reliability and safety.
This extended architectural pattern is rather generic and depends on the actual system to be designed. Therefore we adjusted the pattern to our requirements as mentioned in Section 3.2.1. Furthermore, we excluded the offline updating component because it is out-of-scope for this work.
Process data is routed in parallel to the AI system, the monitoring component, and the supervisory component, the BCR. The BCR operator, informed by the monitor's diagnostics, the raw process data, and the AI system's predictions, manually decides whether to activate safety barriers.

3.3. *STPA*

STPA is a hazard identification technique designed to address system complexity by using a system-theoretic approach and treating safety as a control problem. It posits that accidents can arise not only from component failure, but from unsafe or missing control actions within complex interactions (Rausand and Haugen, 2020; Leveson, 2012). A brief overview of its key steps follows, where the comprehensive guidance is available in the STPA handbook (Leveson and Thomas, 2018).

- **Identify purpose of the analysis**: The analysis begins by defining losses, and hazards that could lead to a loss. These hazards are defined within a system boundary that delineates the scope of the designer's or engineer's control.
- **Model the control structure**: The next step models the system's functional relationships using feedback control loops within the established boundary. Key components are the controller that is an algorithm, human operator, or an organization and the controlled process that is another controller, a process or a component controlled by the controller. Their interaction is defined by control actions, constraints enforced by the controller, and feedback, responses from the controlled process.
- **Identify unsafe control actions**: With the control structure defined, the next step identifies unsafe control actions (UCAs): control actions that, under a specific context and worst-case conditions, would lead to a hazard. An UCA is characterized by five elements: the source controller, the type of flaw, e.g., action is provided or not provided, the specific control action, the triggering context, and its link to a hazard.

- **Identify loss scenarios**: The final step identifies the causal factors behind UCAs, such as inadequate requirements, design errors, incorrect feedback, or component failures. The resulting analysis informs additional safety requirements, identifies mitigation, drives improvements, and provides specific safety recommendations.

4. STPA of System 1 and System 2

This section presents the results of the STPA performed for both system designs, following the established methodology step by step and concluding with a brief comparative discussion. A foundational assumption of this analysis is the reliable, safe, and independent operation of the ESD system, which represents a SIF (SIF-I). This function is implemented by a conventional controller that automatically triggers an ESD if the hydrogen concentration at any monitored sensor location exceeds the LFL. Consequently, this subsystem and its associated control actions and feedback loops are excluded from the subsequent detailed hazard analysis, even though they reside within the broader safety-related system boundary. Furthermore, the analysis is scoped to the deployed system state. Causal factors originating from prior lifecycle phases, such as development, training, or testing of the AI system, are therefore not considered within this assessment.

4.1. *Identify purpose of analysis*

The first STPA step yielded identical results for both designs, as losses are defined at a high level and are largely independent of specific component differences. Overall, five losses were identified for the AI-enabled LH_2 bunkering system, involving four distinct stakeholders: on-site operator, ship company, fuel station company, and AI system developer. The fifth loss $L5$ has no specific stakeholder assigned, as its impact is global. The system boundaries are identical for both designs: a LH_2 bunkering system, as described in Section 3.2, incorporating an AI system for predicting concentration of hydrogen (SIF-II) and an ESD system (SIF-I). The sole system design difference is the monitoring and input guard component in System 2. For the system-level hazards, however, this distinction yields no variation as shown in Table 2.

Table 1. The losses identified across the two system designs.

ID	Loss
L1	loss of life
L2	loss of or damage to equipment
L3	loss of fueling mission
L4	loss of reputation
L5	environmental loss

Table 2. The hazards identified across the two system designs.

ID	Associated losses	Hazard
H1	L1, L2, L3, L4, L5	LH_2 bunkering system releases hydrogen above the LFL.
H2	L3, L4	LH_2 bunkering system does not complete the bunkering process.

4.2. *Model control structure*

A hierarchical control structure was created for each system design based on the architectural patterns from Section 3.2 and information from Alfarizi et al. (2023); Tamburini et al. (2025); Maritime Technologies Forum (2024); European

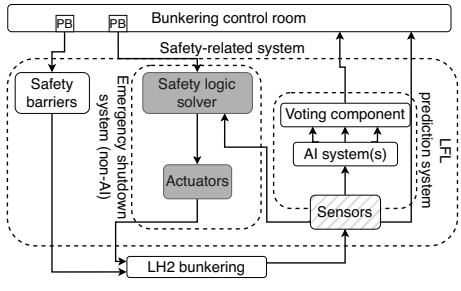

Fig. 2. The hierarchical control structure of System 1. ”PB” stands for ”push button”.

Maritime Safety Agency (2024). Both systems incorporate the following core components: the LH_2 bunkering process, the LFL prediction system (LFL-PS) with the AI system, and the independent ESD system. Both systems access the hydrogen concentration sensors. The LFL-PS additionally collects data from supplementary sensors, such as ambient temperature, pressure, wind speed, and wind direction. The systems differ, however, in their feedback pathways. In System 1, Figure 2, the AI system communicates its predictions directly to the BCR. In System 2, Figure 3, the BCR receives not only the AI system prediction but also supplementary diagnostic information regarding the AI system's internal behavior and input validity. When deviations or unreliable states are detected by the monitoring components, this information is likewise relayed to the BCR. In both systems, a positive prediction of high hydrogen concentration prompts the BCR to activate predefined safety barriers. System 2, however, provides the operator with enhanced situational awareness by additionally reporting on the AI system's operational integrity.

4.3. *Identify unsafe control actions*

We identified three UCAs for both systems. UCA-1 describes when the BCR falsely triggers the ESD and causing an incomplete bunkering process (H2). The other UCAs are detailed in Table 3. The source of these UCAs is the BCR. The controller action is binary where "1" is the default setting, while "0" triggers a safety barrier.

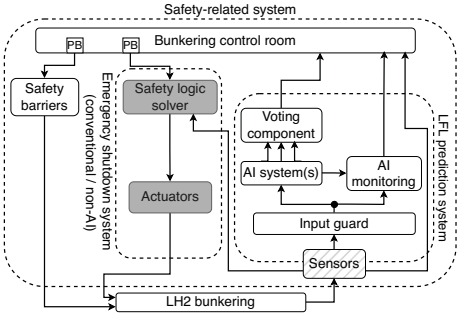

Fig. 3. The hierarchical control structure of System 2.

Table 3. UCAs for both systems origin from the BCR.

not providing causes hazard	providing causes hazard	too early, too late
UCA-2: BCR does not trigger a safety barrier before LFL is surpassed. (H2)	N/A	UCA-3: BCR triggers a safety barrier too late. (H2)

4.4. *Identify loss scenarios*

Finally, we identified the causal loss scenarios that lead to UCAs. Inadequate process models are the central causal factor. For System 1, four loss scenarios were derived: two for UCA-2, and one each for the others. The causal factors are: out-of-distribution data, misclassification, missing data, insufficient trust in AI system, deprecated dependencies, and hardware failures. For instance, one UCA-2 scenario involves an upcoming LFL violation where the BCR fails to trigger a safety barrier due to incorrect information from the AI system originating from out-of-distribution data or misclassification of in-distribution data.

System 2 yielded five loss scenarios, one for UCA-1, two for UCA-2 and UCA-3 and the following set of causal factors: out-of-distribution data, misclassification, insufficient trust in AI system, and hardware failures. A loss scenarios for UCA-3 is the BCR triggers a safety barrier too late. This delay can result from a situation where the AI system provides a correct prediction, but the monitoring component falsely reports unreliable behavior. Consequently, the BCR operator hesitates to act. After several time steps, the monitoring correctly indicates reliable AI behavior, prompting the operator to finally activate the barrier. However, because the safety barrier requires a specific activation and response time to be effective, the delayed deployment renders it insufficient to prevent the hazardous condition, ultimately leading to an unavoidable ESD (H2). This flawed monitoring can originate from misclassification errors within the monitoring.

5. Benefits, limitations and potential adjustments

This section outlines STPA's benefits and limitations for AI-enabled systems and proposes methodological adjustments. STPA was developed to address the hazard identification of complex, software-intensive systems where traditional, linear failure models are insufficient (Leveson, 2012). Given that AI-enabled systems exhibit at least comparable complexity, STPA's foundational approach is applicable and beneficial. However, fundamental differences between conventional software and AI systems mean the technique is not directly translatable and requires adaption. While complexity is a shared characteristic, key distinctions include:

- Software is rule-based and explicitly defined by developers, while AI systems are data-driven and self-organizing through learning.
- AI systems are often non-deterministic in their internal processes and outputs, whereas traditional software is typically deterministic.

5.1. *Benefits*

STPA offers distinct benefits for analyzing AI-enabled systems:

- It captures complex component interactions and socio-technical relationships that linear techniques often miss, making it uniquely suitable (Dobbe, 2022). For instance, STPA captures complex interactions in System 2, such as those between BCR, AI system and AI monitoring, revealing how the added safety layer introduces new failure modes. Consequently, System 2 yielded a greater number of loss scenarios than System 1, demonstrating STPA's value in analyzing the inherent trade-off between safety complexity and emergent risk.
- STPA offers a distinct perspective on hazard identification and can reveal additional hazards compared to traditional techniques (Rausand and Haugen, 2020). For example, by explicitly modeling component interactions, particularly in socio-technical systems such as those enabled by AI, STPA can uncover emergent hazards that must be addressed for system safety. These may include issues like insufficient operator trust in the AI system, which could stem from inadequate personnel training or a deficient organizational safety culture.
- It enables the design of safe system from the start, improving safety and reducing lifecycle costs. For instance, the identification of loss scenarios and casual factors can be the starting point of design iteration that try to mitigate or prevent these.

5.2. *Limitations*

STPA's limitations for AI-enabled systems stem from both general methodological constraints and AI-specific challenges. A well-known general limitations is its procedural complexity, which often requires guidance or external experts (Koessler and Schuett, 2023). The limitations identified are:

- While established hazard identification methods from safety-critical industries provide a foundational starting point for risk assessment in AI-enabled systems (Koessler and Schuett, 2023), they are not inherently designed for this context and require careful adaptation to be fully effective. The adoption rather than replacement of these methods offers the advantage of leveraging the existing familiarity and expertise of safety professionals with these proven tools.
- Second, STPA, while designed for software-intensive systems, is not specifically tailored for those incorporating AI systems. The methodology remains functional at a high level but lacks the requisite granular language for formulating detailed loss scenarios. The nature of interactions with AI systems extends beyond the binary "correct/incorrect" framework typical of STPA, as AI systems' outputs can be probabilistic and ambiguous. Similarly, a focus on traditional sensor-based feedback is insufficient for detailed AI-specific hazard identification.
- Third, the STPA construct of the operator's process model that is crucial for ensuring safe control actions, is fundamentally challenged by AI systems. The complexity and opacity of AI decision-support can impede operator under-

standing, a situation exacerbated by the current limitations of meaningful explainability techniques.

5.3. *Adjustments*

STPA is already a valuable hazard identification technique for AI-enabled systems due to its system-theoretic perspective and ability to model complex interactions (Koessler and Schuett, 2023; Dobbe, 2022). To address the aforementioned limitations and improve diffusion of the technique, we propose two enhancements:

- A qualitative risk assessment to compare different system designs.
- Adapt the loss scenario identification methodology with AI-specific guidance and language.

5.3.1. *Qualitative risk assessment to enable comparability*

Our case study reveals a methodological limitation in STPA: the lack of an integrated framework for comparing different system designs, whether between two AI-enabled systems or against a purely conventional, non-AI system. This issue is not new, prior work by Kim et al. (2021) attempted to address it by introducing a risk priority number approach to quantify and rank hazards. However, within the STPA context, quantitatively scoring often elusive causal factors can be arbitrary and of limited meaning. A structured qualitative comparison could be far more practical and insightful. This could be achieved using a predefined checklist or a comparative risk matrix that assesses key design attributes, such as safety layer depth, system complexity, transparency, and operational resilience which enables a clear, side-by-side evaluation of design alternatives.

5.3.2. *Translating loss scenario concept for AI-enabled systems*

During the identification of loss scenarios, we observed that STPA's conventional framework is inadequate for analyzing AI-enabled systems. To apply STPA effectively, its guidelines must be translated into AI-specific terms. For instance, in System 2, outputs from the AI system and its independent monitor can be inconsistent or contradictory, leading to flawed process beliefs by the BCR operator. Additionally, AI outputs are often probabilistic and non-deterministic, making the identification of loss scenarios and their casual factors more complex. This translation is essential, as it yields more detailed loss scenarios and may reveal novel causal factors that can be addressed to improve system safety. STPA's traditional causes of inadequate feedback, such as sensor faults, do not account for AI-specific failure modes, including deceptive outputs (Hubinger et al., 2024), misleading or incomprehensible monitoring (Miller et al., 2017), or emergent behavior (Hendrycks, 2024). Formally integrating these novel causes is necessary to accurately describe and subsequently mitigate AI-specific hazards.

6. Discussion

This paper employed a case study approach to evaluate STPA for AI-enabled systems and derive methodological adjustments. We demonstrated that STPA provides a valuable perspective for hazard identification in this domain by analyzing component interactions and socio-technical dynamics, aspects often neglected by traditional techniques (Koessler and Schuett, 2023; Rismani et al., 2024). Its adoption as complementary analysis is therefore meaningful. To improve its utility and diffusion, we proposed two enhancements:

- A qualitative risk assessment to enable system design comparison.
- AI-specific guidance for loss scenario identification, translating STPA's concepts to address novel AI-specific failure modes.

Our analysis, while conducted according to the STPA handbook (Leveson and Thomas, 2018) and reviewed by experts, is a high-level application and can be further refined. The LH_2 bunkering case study is grounded in real-world system designs (Maritime Technologies Forum, 2024; European Maritime Safety Agency, 2024), but the AI system integration is conceptual, based on our expertise. Implementing the proposed adjustments requires further precision and validation. The risk assessment must carefully incorporate system-theoretic principles. Similarly, developing

comprehensive AI-specific guidance demands interdisciplinary work between AI safety expertise, safety engineers, and AI experts.

7. Conclusion

In this work, we applied STPA to two distinct AI-enabled system designs within a safety-critical LH_2 bunkering context. The analysis indicated STPA's utility in identifying loss scenarios inherent to the incorporation of AI systems while revealing methodological limitations. Our primary contribution is the proposal of two targeted enhancements, that is a qualitative risk assessment and AI-specific methodology for loss scenarios identification, aimed at improving the technique's practicality and accelerating its diffusion for AI-enabled system risk analysis. Future work should focus on refining and validating these proposals: defining a concrete framework for STPA risk assessment and developing comprehensive, translatable guidelines for AI-specific loss scenarios identification.

Acknowledgement

This work was carried out as a part of SUBPRO-Zero, a Research Centre at the Norwegian University of Science and Technology. The authors gratefully acknowledge the project support from SUBPRO-Zero, which is financed by major industry partners and NTNU.

References

Alfarizi, M. G., F. Ustolin, J. Vatn, S. Yin, and N. Paltrinieri (2023). Towards accident prevention on liquid hydrogen: A data-driven approach for releases prediction. *Reliability Engineering & System Safety 236*, 109276.

Det Norske Veritas (2023). Assurance of AI-enabled systems.

Dobbe, R. I. J. (2022). System Safety and Artificial Intelligence.

EU AI Act (2024). European Union Artificial Intelligence Act.

European Maritime Safety Agency (2024). Hazard identification of generic hydrogen fuel systems.

Flehmig, N., M. A. Lundteigen, and S. Yin (2024). Implementing Artificial Intelligence in Safety-Critical Systems during Operation: Challenges and Extended Framework for a Quality Assurance Process. In *IECON 2024 - 50th Annual Conference of the IEEE Industrial Electronics Society*, pp. 1–8. IEEE.

Hendrycks, D. (2024). *Introduction to AI Safety, Ethics, and Society*. Taylor & Francis.

Hubinger, E., C. Denison, J. Mu, M. Lambert, M. Tong, M. MacDiarmid, T. Lanham, D. M. Ziegler, T. Maxwell, N. Cheng, et al. (2024). Sleeper agents: Training deceptive llms that persist through safety training.

ISO/IEC Guide 51 (2014). Safety aspects - Guidelines for their inclusion in standards.

ISO/IEC TR 5469 (2024). Artificial intelligence — Functional safety and AI systems.

Kim, H., M. A. Lundteigen, A. Hafver, and F. B. Pedersen (2021). Utilization of risk priority number to systems-theoretic process analysis: A practical solution to manage a large number of unsafe control actions and loss scenarios. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability 235*(1), 92–107.

Koessler, L. and J. Schuett (2023). Risk assessment at AGI companies: A review of popular risk assessment techniques from other safety-critical industries.

Leveson, N. (1995). *SafeWare: System Safety and Computers*. Addison-Wesley.

Leveson, N. G. (2012). *Engineering a Safer World: Systems Thinking Applied to Safety*. Engineering Systems. The MIT Press.

Leveson, N. G. and J. P. Thomas (2018). STPA handbook. pp. 1–188.

Maritime Technologies Forum (2024). Guidelines for the development of liquefied hydrogen bunkering systems and procedures.

Miller, T., P. Howe, and L. Sonenberg (2017). Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences.

Rausand, M. and S. Haugen (2020). *Risk Assessment: Theory, Methods, and Applications* (1 ed.). Wiley.

Rismani, S., R. Dobbe, and A. Moon (2024). From silos to systems: Process-oriented hazard analysis for ai systems. *arXiv preprint arXiv:2410.22526*.

Tamburini, F., M. Iaiani, and V. Cozzani (2025). Analysis of system resilience in escalation scenarios involving LH2 bunkering operations. *Reliability Engineering & System Safety 257*, 110816.