Cybersecurity and Innovatove Technology Journal, Vol.3, No.2, 2025, pp. 68-78
e-ISSN 3025-6682. DOI: 10.53889/citj.v3i2.670

Considerations for the Safety Analysis of AI-Enable Systems

Submitted 27 June 2024, Revised 5 November 2024, Accepted 31 December 2025

Christopher W. Green[1]*

[1]Engineering Department, Capitol Technology University, Laurel, United States
Corresponding Email: *cgreen@captechu.edu

Abstract

This study explored the applicability of hazard analysis techniques to Artificial Intelligence/Machine Learning AI-enabled systems, a growing area of concern in safety-critical domains. The study evaluates 127 hazard analysis techniques described in the System Safety Society's System Safety Analysis Handbook (1997) for their relevance to the unique challenges posed by AI-enabled systems. A qualitative criteria-based assessment framework was employed to systematically analyze each technique against key AI-specific considerations, including complexity management, human-AI interaction, dynamic and adaptive behavior, software-centric focus, probabilistic and uncertainty handling, and iterative development compatibility. The evaluation process involved defining criteria to address AI/ML systems' distinctive characteristics, assessing each method's applicability, and ranking techniques based on their alignment with AI-related challenges. Findings indicate that Fault Tree Analysis (FTA) and Human Reliability Analysis (HRA) are highly relevant for performing safety on AI-enabled systems. Other techniques, such as What-If Analysis, require adaptation to address emergent behaviors. This study provides a framework for selecting and tailoring hazard analysis methods for AI-enabled systems, contributing to developing robust safety assurance practices in an increasingly intelligent and autonomous era.

Keywords: Systems Engineering, System Safety, Artificial Intelligence, Machine Learning, Hazard Analysis

INTRODUCTION

Rapid artificial intelligence (AI) and machine learning (ML) integration into critical systems has transformed the healthcare and transportation industries. These AI-enabled systems exhibit autonomous decision-making, pattern recognition, and adaptive learning capabilities, offering unprecedented opportunities for efficiency, innovation, and problem-solving. However, their growing complexity and reliance on dynamic, data-driven algorithms also present unique safety challenges. Unlike traditional systems, AI-enabled systems are characterized by emergent behaviors, probabilistic decision-making, and human-AI interaction dynamics, requiring reevaluating existing safety analysis techniques.

Safety analysis plays a pivotal role in ensuring that systems operate within acceptable levels of risk. Historically, system safety has relied on well-established standards and methodologies, such as those outlined in MIL-STD-882E. These techniques provide structured approaches to identify, assess, and mitigate hazards. However, applying these methods to AI systems, with their iterative development cycles and inherent uncertainties, necessitates adaptation and innovation. Traditional safety frameworks must be reevaluated to address the unique risks associated with AI, including algorithmic biases, emergent system behaviors, and complex interactions between software and hardware components.

This paper explores the suitability of 127 hazard analysis techniques for application to AI-enabled systems. Drawing from the System Safety Society's System Safety Analysis Handbook

(1997), this study evaluates these techniques using a qualitative, criteria-based framework that considers factors such as complexity management, dynamic behavior, and probabilistic risk assessment. Techniques are categorized based on relevance, ease of adaptation, and alignment with AI's unique characteristics (System Safety Society, 1997).

This research examines the applicability of these methods to provide system safety engineers with a comprehensive guide for integrating hazard analysis into the development and deployment of AI-enabled systems. The results highlight which techniques are most relevant and offer insights into how traditional methods can be modified or combined to address the demands of modern AI technologies. This study is a critical step toward ensuring that the safety of AI-enabled systems keeps pace with their increasing complexity and adoption across diverse domains.

System Safety

System Safety is "The application of engineering and management principles, criteria, and techniques to achieve acceptable risk within the constraints of operational effectiveness and suitability, time, and cost throughout all phases of the system lifecycle." Many standards and guidelines govern system safety. The System Safety process consists of Planning, Identifying, Assessing, Recommending/ Implementing Mitigations, and Verifying Design and Mitigations. Some relevant definitions are listed below MIL-STD-882E, 2012):

- Accident: Any unplanned act or event that damages property, material, equipment, or cargo, or personnel injury or death when not due to enemy action (Navy OP4 & OP5).
- Mishap: An event or series of events resulting in unintentional death, injury, occupational illness, damage to or loss of equipment or property, or damage to the environment. (MIL-STD-882E).
- Hazard: Any real or potential condition that can cause injury, illness, or death to personnel; damage to or loss of a system, equipment, or property; or damage to the environment (MIL-STD-882E).
- Risk: A combination of the severity of the mishap and the probability that the mishap will occur. (MIL-STD-882E).

Figure 1. shows the relationship between a hazard and a mishap. A hazard and a mishap are two separate states of the same phenomenon linked by a state transition that must occur. You can think of these states as the before and after states. A hazard is a "potential event" at one end of the spectrum that may be transformed into an "actual event" (the mishap) at the other end of the spectrum based upon the state transition (Ericson,2005).

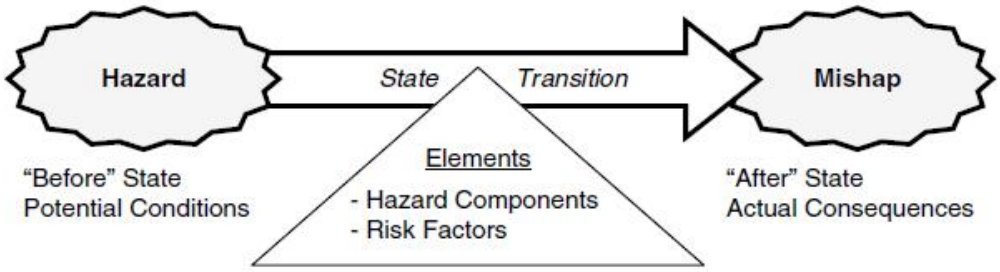

Figure 1. Relationship Between A Hazard And A Mishap

Hazard Analysis

Hazard analysis plays a critical role in system safety in identifying and mitigating risks within a system. However, understanding the distinction between hazard analysis types and techniques is crucial for effective safety management.

A hazard analysis type defines the foundation of the analysis process. It sets the purpose, timing, scope, level of detail, and system coverage, answering questions such as why the analysis is needed, when it should be performed, what it aims to cover, and how deeply it will examine the system. Notably, it does not prescribe the specific methodology for conducting the analysis. Instead, it provides a framework to guide the approach.

On the other hand, a hazard analysis technique focuses on the how—a specific and unique methodology for conducting the analysis. These techniques are tailored to produce targeted results and address particular aspects of system safety. While more than 100 different analysis techniques are available, each is designed to meet the requirements of one or more hazard analysis types.

The discipline of system safety is built on seven primary hazard analysis types, each addressing different risk identification and mitigation dimensions. Each type maps to program development phases within the lifecycle. Together, these types ensure comprehensive coverage of potential system hazards. However, relying on just one analysis type cannot uncover all possible hazards. Effective safety management often requires the integration of multiple, if not all, of these types. The seven hazard analysis types in system safety are:

- Conceptual Design Hazard Analysis Type (CD HAT);
- Preliminary Design Hazard Analysis Type (PD HAT);
- Detailed Design Hazard Analysis Type (DD-HAT);
- System Design Hazard Analysis Type (SD-HAT);
- Operations Design Hazard Analysis Type (OD-HAT);
- Health Design Hazard Analysis Type (HD-HAT)
- Requirements Design Hazard Analysis Type (RD-HAT).

By combining these hazard analysis types with the appropriate techniques, system safety professionals can systematically identify and resolve risks, ensuring the systems' safe and reliable operation (MIL-STD-882E, 2012; Popović & Vasic, 2008; Ericson, 2005).

The AI-Enabled System

An AI-enabled system integrates artificial intelligence (AI) technologies designed to enhance functionality, decision-making, and operational efficiency. These systems leverage advanced algorithms, data models, and computational techniques to perform tasks that traditionally require human intelligence. Key capabilities of AI-enabled systems include reasoning, learning from experience, pattern recognition, predictive analytics, and autonomous decision-making. Such systems combine AI components with hardware, software, and user interfaces to achieve seamless functionality at both the component and system levels. They are engineered to adapt dynamically to changing environments, process vast data, and execute tasks precisely and consistently. Examples include autonomous vehicles, healthcare diagnostic platforms, industrial robotics, financial analytics tools, and intelligent virtual assistants.

Some of the core components that can make up an AI-enabled system are:

- Data Processing Unit: Handles the acquisition, preprocessing, and transformation of raw data into actionable insights.
- Machine Learning Models: Enable systems to learn from historical data and continuously improve performance. Techniques such as neural networks, decision trees, and reinforcement learning are often employed.
- Decision-Making Algorithms: Provide real-time responses and adaptive behavior based on input data and learned patterns.
- Human-AI Interfaces: Facilitate user interaction and the AI system, ensuring usability and accessibility.
- Integration Frameworks: Allow seamless communication and cooperation between AI components and other subsystems, such as sensors, actuators, and external devices.

AI-enabled systems are distinguished by their ability to handle uncertainty, evolve with iterative updates, and operate autonomously or semi-autonomously in dynamic and unpredictable environments. They present unique opportunities for innovation while introducing challenges such as ensuring safety, managing bias, and addressing ethical considerations.

AI Safety

AI safety is crucial because artificial intelligence systems are increasingly integrated into critical and high-stakes domains, making their safety essential to protect individuals, organizations, and society. Artificial Intelligence Safety Engineering (AI Safety) was first coined

in 2010 by Roman Yampolskiy and described in his book “Artificial Intelligence Safety and Security.” It emerged from computer science with research on autonomous vehicles and is relatively young and underfunded outside of industry (Yampolskiy, 2015). Some of the challenges of performing safety engineering on AI-enabled systems are:

- Complexity and Unpredictability of AI Systems
- Lack of Standardized Safety Frameworks
- Ethical and Societal Challenges
- Regulatory and Legal Challenges
- Integration into Existing Safety Workflows
- Economic and Resource Constraints

The literature on AI safety shares the commonality that AI Safety should (Amodei et al, 2016; Dobbe, R., 2022; Johnson, B., 2022; Yampolskiy, R., 2019):

- Focus on Risk Mitigation:
- Align with Human Values:
- Develop Adaptable System for Real-Time Safety:
- Have an Interdisciplinary Approach:
- Focus on Preventative Measures:
- Address Complexity and Uncertainty:

Because of these challenges, AI-enabled systems' hazard analysis should evolve beyond traditional frameworks. By expanding its scope, adopting dynamic methodologies, and incorporating ethical, societal, and cybersecurity considerations, hazard analysis can effectively address the unique risks posed by AI. However, this transformation requires innovation, interdisciplinary collaboration, and alignment with evolving regulatory standards. These adaptations are essential to ensure hazard analysis remains the foundation of safe and reliable AI deployment (Carter et al., 2022; Cummings, M.L., 2024; Garvin & Kimble, 2022; Martelaro et al., 2022).

METHOD

The System Safety Analysis Handbook has 127 hazard analysis techniques. Each technique was reviewed, and techniques specific to a particular domain, such as radiation, human factors, and nuclear, were ruled out. The remaining techniques were evaluated for their applicability to AI-enabled systems. A qualitative criteria-based assessment was used. This method systematically analyzes each technique against specific challenges and characteristics. Below is an outline of the method used (System Safety Society, 1997; Amodei et al., 2016).

Criteria Definition

To evaluate the hazard analysis techniques, key criteria were established based on the unique challenges and requirements of AI-enabled systems:

- Complexity Management: Can the method handle the intricacies of AI decision-making, concurrent processes, and emergent behaviors?
- Human-AI Interaction: Does the method address risks associated with human interaction with AI, including errors in interpreting or misusing AI outputs?
- Dynamic and Adaptive Behavior: Can the technique analyze risks arising from AI systems' adaptive nature, such as retraining or evolving algorithms?
- Software-Centric Focus: Is the technique suited to identifying software-related hazards, such as algorithmic errors, bias, or unintended behavior?
- Probabilistic and Uncertainty Handling: Does the method account for uncertainties in AI predictions and the probabilistic nature of many AI-driven decisions?
- Iterative Development Compatibility: Is the technique flexible enough to be applied at different stages of AI system development, particularly in iterative workflows?

Technique Analysis

A structured approach was adopted to evaluate various hazard analysis techniques, focusing on three key criteria. First, each technique was thoroughly reviewed to understand its purpose, scope, and traditional application areas. This analysis provided insight into each method's foundational intent and established uses. Second, consideration was given to these techniques' potential adaptation or extension to address AI-specific challenges. This involved exploring how existing methodologies could be reimagined to tackle the unique complexities introduced by AI technologies.

RESULTS AND DISCUSSION

Hazard Analysis Techniques for AI-Enabled Systems

Below is an evaluation of hazard analysis techniques from your list, specifically selected for their applicability to AI-enabled systems. These techniques are chosen based on their ability to address challenges unique to AI, such as software faults, human-AI interaction, complex system dynamics, and emergent behavior. The applicable techniques and their relevance are listed below.

1. Accident Analysis*:* Useful for understanding accidents caused by AI decisions or failures. Can analyze how AI-related mishaps occurred and propose preventive measures.

2. Action Error Analysis*:* Examines errors in decision-making, which is particularly relevant for AI systems executing autonomous actions.

3. Barrier Analysis*:* Identifies barriers that prevent AI-related hazards, such as safety interlocks or ethical safeguards in AI systems.

6. Cause-Consequence Analysis*:* Traces cause-and-effect relationships to understand how AI faults propagate and lead to hazardous outcomes.

7. Change Analysis*:* AI systems evolve through updates or retraining. This technique evaluates the impact of changes on system safety.

32. Failure Modes and Effects Analysis (FMEA)*:* Identifies failure modes in AI algorithms, sensors, or integration points with hardware, assessing their effects on system performance.

33. Failure Modes, Effects, and Criticality Analysis (FMECA)*:* Adds criticality ranking to FMEA, prioritizing AI-related failures that could lead to catastrophic outcomes.

36. Fault Tree Analysis (FTA)*:* Models logical pathways to AI system failures, such as software bugs, misclassification errors, or sensor malfunctions.

40. Hazard and Operability Study (HAZOP)*:* Examines AI operations for deviations from expected behavior, such as unintended outputs or unsafe decisions.

42. Hardware/Software Safety Analysis: Evaluates the interaction between AI software and hardware to identify integration risks.

44. Human Error Analysis*:* Focuses on human-AI interaction, examining errors introduced by operators, designers, or users interacting with AI systems.

46. Human Reliability Analysis (HRA)*:* Examines the reliability of human responses to AI outputs, particularly in high-stakes applications like autonomous vehicles or healthcare.

47. Interface Analysis*:* Investigates the interfaces between AI components, human operators, and other subsystems to identify integration hazards.

62. Petri Net Analysis*:* Models concurrent processes in AI systems, such as decision-making or data processing pipelines, to identify unsafe conditions.

64. Preliminary Hazard List*:* This list identifies potential hazards early in AI system development, helping guide detailed analysis.

65. Probabilistic Hybrid Analytical System Evaluation Tool: Incorporates probabilistic approaches to evaluate uncertainties in AI predictions and system reliability.

66. Probabilistic Risk Assessment (PRA): Quantifies the likelihood and impact of AI system failures using statistical methods, particularly useful for high-complexity AI models.

69. Process Hazard Analysis: This applies to AI systems managing dynamic processes, ensuring operational risks are understood and mitigated.

75. Root Cause Analysis: Identifies the root causes of AI-related failures, such as training data biases or software bugs.

102. Risk-Based Decision Analysis: Evaluates trade-offs in AI decision-making systems to prioritize safety-critical actions.

110. Software Failure Modes and Effects Analysis (SFMEA): Focuses on software-specific failure modes in AI systems, such as algorithm errors or memory overflow.

111. Software Fault Tree Analysis: Builds logical fault trees for AI algorithms, identifying critical failure pathways in software.

112. Software Hazard Analysis: Examines risks specific to AI software, such as unexpected behavior from neural networks or reinforcement learning.

121. Technique for Human Error Prediction (THERP): Predicts potential errors in human-AI interactions, especially in safety-critical applications.

127. What-If Analysis: Explores hypothetical scenarios where AI systems fail, identifying potential hazards and mitigation strategies.

Comparative Evaluation

A comprehensive comparative framework was developed to rank the techniques based on their suitability for addressing AI-related safety challenges. The evaluation categorized the techniques into three distinct levels of applicability:

- Direct Applicability: Some techniques, such as Fault Tree Analysis (FTA) and Hazard and Operability Analysis (HAZOP), were identified as readily applicable to AI systems with little to no adaptation required. These methods could seamlessly integrate into AI-specific contexts, leveraging their existing strengths.
- Moderate adaptation Required: Certain techniques, such as Safety Failure Modes and Effects Analysis (SFMEA), showed promise for application to AI systems but required moderate extensions to address issues unique to neural networks and other AI-specific architectures.
- Limited Applicability: A few techniques, like those traditionally focused on confined space safety, were determined to have limited relevance to AI systems unless significantly reconfigured. These methods were deemed less practical for addressing the nuanced hazards associated with AI.).

Integration Considerations

In addition to evaluating the individual techniques, the analysis examined how they could be effectively integrated into existing safety workflows for AI-enabled systems. Techniques were categorized based on their roles in hazard identification, assessment, and mitigation, enabling a systematic approach to safety management. Furthermore, complementary methods were identified to support a layered analysis strategy. For instance, Failure Modes and Effects Analysis (FMEA) was recommended for detailed component-level analysis, while Probabilistic

Risk Assessment (PRA) was suggested for addressing system-wide risks. This multi-tiered approach ensured comprehensive coverage of potential hazards, enhancing the robustness of AI system safety workflows.

The evaluation of hazard analysis techniques for AI-enabled systems highlights the growing need to adapt traditional safety methodologies to the unique challenges posed by artificial intelligence. AI systems introduce complexities such as dynamic behavior, probabilistic decision-making, and human-AI interactions that conventional techniques may not fully address. The categorization of techniques into direct applicability, moderate adaptation, and limited applicability offers a practical framework for selecting appropriate methods.

For instance, Fault Tree Analysis (FTA) and Human Reliability Analysis (HRA) provide robust tools for addressing AI systems' reliability and interaction challenges without extensive modification. However, methods such as What-If Analysis illustrate the potential for expanding traditional frameworks to account for AI-specific concerns, such as algorithmic biases or emergent behaviors. This discussion underscores the importance of adopting a layered approach, combining complementary techniques to ensure comprehensive hazard identification and mitigation.

The integration considerations further emphasize that safety workflows must evolve to accommodate iterative AI development cycles. Safety assessments must keep pace with AI models' rapid retraining and deployment, ensuring hazard analyses are embedded in continuous development processes.

CONCLUSION

This study systematically evaluates 127 hazard analysis techniques, identifying their relevance and adaptability for AI-enabled systems. The findings demonstrate that while some techniques are directly applicable, others require tailored modifications to address AI-specific risks effectively. Techniques such as FTA and HRA emerged as critical tools for assessing system reliability and human interaction, while methods like SFMEA and What-If Analysis offer opportunities for targeted adaptations.

This study contributes to the growing body of knowledge in AI safety, offering a framework for selecting and customizing hazard analysis methods. By aligning traditional techniques with the unique characteristics of AI systems, this study supports the development of robust safety practices that can be applied across various industries, from healthcare to autonomous vehicles. The results highlight the necessity of combining techniques to address component- and system-wide risks, ensuring a holistic approach to safety in AI-enabled environments.

This criteria-based qualitative evaluation ensures a systematic and context-aware assessment of hazard analysis techniques for AI systems. By aligning each technique's purpose and strengths with the challenges posed by AI, this method provides a robust framework for determining their applicability.

SUGGESTIONS

Future research should focus on expanding the adaptation of traditional hazard analysis techniques to address AI-specific challenges. This includes developing new methodologies tailored to address dynamic and adaptive behaviors, algorithmic biases, and the uncertainties inherent in AI systems. Additionally, integrating machine learning algorithms into hazard analysis workflows could automate the identification of emerging risks, providing real-time insights into system safety.

Another promising avenue for research is exploring AI's role in enhancing hazard analysis techniques. For example, using AI-driven models to predict potential hazards or simulate complex system interactions could offer new capabilities for safety assurance.

Finally, industry collaboration is essential for validating the applicability of these techniques in real-world AI-enabled systems. Conducting case studies across diverse sectors, such as healthcare, manufacturing, and defense, will provide practical insights into the effectiveness of adapted hazard analysis methods. These efforts will ensure that the proposed frameworks remain relevant and practical in addressing the evolving challenges of AI system safety.

REFERENCES

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete Problems In AI Safety. *arXiv.org*. https://arxiv.org/abs/1606.06565

Carter, H. G., Chan, A., Vinegar, C., & Rupert, J. (2022). Proposing The Use Of Hazard Analysis For Machine Learning Data Sets. *Journal of System Safety, 58*(2). https://doi.org/10.56094/jss.v58i2.253

Cummings, M. L. (2024). A Taxonomy For AI Hazard Analysis. *Journal of Cognitive Engineering and Decision Making, 18*(4), 327–332. https://doi.org/10.1177/15553434231224096

Dobbe, R. (2022). System Safety And Artificial Intelligence. In *FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (p. 1584). https://doi.org/10.1145/3531146.3533215

Department of Defense (DOD). (2012). *MIL-STD-882E: Department Of Defense Standard Practice System Safety*.

Ericson, C. A. (2005). *Hazard Analysis Techniques For System Safety* (2nd ed.). John Wiley & Sons.

Garvin, T., & Kimbleton, S. (2021). Artificial Intelligence As Ally In Hazard Analysis. In *American Institute of Chemical Engineers 2020 Spring Meeting and 16th Global Congress on Process Safety* (August 16–20, 2020).

Johnson, B. (2022). Metacognition For Artificial Intelligence System Safety – An Approach To Safe And Desired Behavior. *Safety Science, 151*, 105743. https://doi.org/10.1016/j.ssci.2022.105743

Martelaro, N., Smith, C. J., & Zilovic, T. (2022). Exploring Opportunities In Usable Hazard Analysis Processes For AI Engineering. *arXiv preprint*. https://arxiv.org/abs/2203.15628

Popović, V. M., & Vasić, B. (2008). Review Of Hazard Analysis Methods And Their Basic Characteristics. *FME Transactions, 36*(4).

System Safety Society. (1997). *System Safety Analysis Handbook* (2nd ed.).

Yampolskiy, R. V.. (2019). *Artificial Intelligence Safety And Security*. CRC Press/Taylor & Francis Group. https://doi.org/10.1201/9781351251389