Cybersecurity and Innovatove Technology Journal, Vol.3, No.2, 2025, pp. 85-103
e-ISSN 3025-6682. DOI: 10.53889/citj.v3i2.671

System Safety Preliminary Hazard Analysis (PHA) Using Generative Artificial Intelligence

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 investigated the capability of ChatGPT, an AI-powered generative language model, to perform hazard analysis for complex systems using the ACME Missile System as a case study. Hazard analyses generated by ChatGPT were compared to those detailed in Ericson, Clifton's 2005 publication, Hazard Analysis Techniques for System Safety, focusing on adherence to MIL-STD-882E methodologies. The research addresses general questions regarding the strengths and limitations of ChatGPT in identifying hazards, assessing risks, and proposing mitigation strategies. Through a structured evaluation, the study examines the completeness, accuracy, and alignment of ChatGPT-generated analyses with traditional techniques, identifying areas of strength, such as efficiency and innovative mitigation suggestions, alongside gaps in contextual understanding and methodological consistency. Findings highlight the potential of ChatGPT as a supplementary tool for initial hazard identification, emphasizing the importance of expert validation to ensure reliability in safety-critical applications. This research contributes to understanding AI's role in system safety engineering and integration into existing hazard analysis frameworks.

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

INTRODUCTION

ChatGPT can aid the System Safety Engineer in performing Hazard Analysis by generating a list of hazards based on system descriptions, operational contexts, and known failure modes that align with standards like MIL-STD-882E, ISO 26262, or others. Some of the research questions that will be answered during this research are:

1. How effectively can ChatGPT generate hazard analyses for safety critical systems?
2. What are the strengths and limitations of using ChatGPT in system safety engineering?

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 the following steps (MIL-STD-882E, 2012):

1. Plan: Plan to get system safety involved in a program as soon as possible
2. Identify: Testing, Data, safety situations, scenarios, failures, and conditions that may uncover, define, characterize, or validate hazards
3. Assess: Assess risk; Various standards available
4. Recommend/ Implement Mitigations: Get buy-in from stakeholders
5. Verify Design and Mitigations: Use standards such as MIL-STD-1472 and test results

Some relevant definitions are listed below:

1. Accident: Any unplanned act or event that damages property, material, equipment, cargo, or personnel injury or death when not resulting from enemy action (Navy OP4 & OP5).
2. Mishap: An unplanned event or series of events resulting in death, injury, occupational illness, damage to or loss of equipment or property, or damage to the environment (MIL-STD-882D).
3. 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-882D).
4. Risk: An expression of the impact and possibility of a mishap in terms of potential mishap severity and probability of occurrence (MIL-STD-882D).

Figure 1. shows Ericson's 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.

Hazard analysis is described as the systematic process of identifying hazards, their effects, and causal factors to assess system risk and determine the significance of hazards, enabling the establishment of safety design measures to eliminate or mitigate risks across systems, subsystems, components, software, personnel, and their interrelationships (Erickson, 2005).

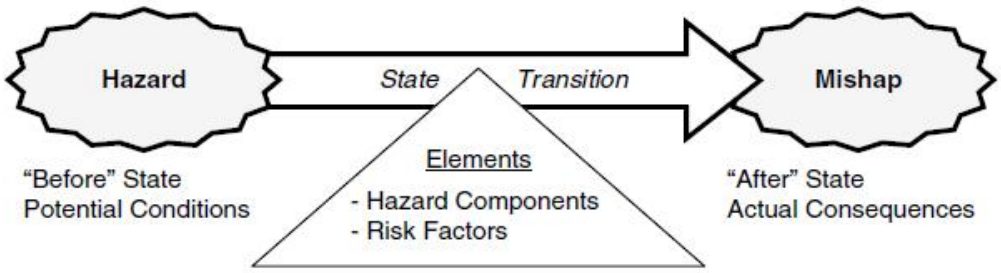

Figure 1. Relationship Between Hazard and Mishap

Types of Hazard Analysis

The DoD's Standard Practice for System Safety (MIL-STD-882E) outlines a comprehensive framework for hazard analysis to identify, assess, and mitigate risks throughout a system's lifecycle. This structured approach supports compliance with safety requirements and ensures effective risk management in safety-critical systems. The types of hazard analysis described in MIL-STD-882E provide a systematic methodology for addressing hazards, with each type tailored to specific phases and components of a system's lifecycle. Below is a listing of some of the primary hazard analysis types.

1. Preliminary Hazard List (PHL)
2. Preliminary Hazard Analysis (PHA)
3. System Hazard Analysis (SHA)
4. Subsystem Hazard Analysis (SSHA)
5. Operating and Support Hazard Analysis (O&SHA)
6. Functional Hazard Analysis (FHA)
7. Health Hazard Analysis (HHA)
8. Environmental Hazard Analysis (EHA)
9. Software Hazard Analysis
10. Common Cause Analysis (CCA)

Generative AI: ChatGPT and Its Foundations

ChatGPT is a generative AI system developed by OpenAI that emphasizes natural language processing (NLP), and large language models (LLMs). It highlights AI as the computational simulation of human cognitive abilities, enabling tasks like reasoning, learning, and language understanding. Generative AI, which focuses on creating human-like content, underpins ChatGPT's ability to generate coherent, contextually relevant text using probabilistic modeling.

At its core, ChatGPT leverages OpenAI's Generative Pre-trained Transformer (GPT) architecture, which utilizes pre-training and fine-tuning on vast datasets to understand language patterns and context. The transformer architecture, introduced in 2017, facilitates efficient processing of text relationships and long-range dependencies, enabling ChatGPT to maintain conversational coherence over multiple interactions.

The historical development traces AI progress from foundational NLP advancements in the 1950s–1980s, modern breakthroughs in the 1990s–2010s, and OpenAI's milestones with GPT-1 in 2018, GPT-2 in 2019, and GPT-3 in 2020. ChatGPT, initially based on GPT-3.5 and later enhanced with GPT-4, integrates Reinforcement Learning from Human Feedback (RLHF) to improve conversational accuracy and human alignment. The system exemplifies the potential of generative AI in advancing conversational technologies.

Prompt Engineering

Prompt engineering systematically designs and refines input prompts to optimize the outputs generated by large language models (LLMs) such as ChatGPT. Given that LLMs derive their behavior and responses from the instructions they receive, the quality of these inputs directly influences the relevance, accuracy, and clarity of their outputs. A well engineered prompt consists of the task definition, context, constraints, and examples. One prompt formula

emphasizes the following in order of importance.

1. Task (This should start with an action verb such as Give, Generate, List, Analyze. i.e. ”Give me a 3-month training program.”)
2. Context: limit the possibilities (I am a 170lb male)
3. Exemplar: examples (Rewrite this bullet point using this structure: "I accomplished X by the measure Y that resulted in Z."
4. Format: What is the desired result (Output this data in a table with headings "City," "Population," "Age"
5. Tone: casual, formal, witty, enthusiastic, pessimistic

The prompt's structure is depicted below. The first two are needed at minimum to give credible results.

[task]+[context]+[exempler]+[format]+[tone] = productive results

The ACE Missile System: A Conceptual Case Study in Hazard Analysis

The ACE Missile System, described in Clifton Ericson’s Hazard Analysis Techniques for System Safety (2005), is an illustrative example of hazard analysis in safety-critical systems. While hypothetical, the ACE Missile System provides a comprehensive framework for understanding the application of hazard analysis techniques in mitigating risks associated with high-performance military systems. This case study highlights the interdependencies between subsystems, the challenges of ensuring safety in dynamic environments, and the methodologies used to address potential hazards.

Purpose and Context

The ACE Missile System is conceptualized as a high-performance missile designed to demonstrate hazard analysis methodologies throughout its lifecycle. Although not operational, this system represents the complexities of real-world missile systems that operate under diverse and often hostile conditions, including adverse weather, electromagnetic interference, and high mechanical stresses. The ACE Missile System enables engineers to explore systematic approaches to identifying, assessing, and mitigating risks inherent in such safety-critical environments (Ericson, 2005).

System Architecture

The ACE Missile System comprises two primary segments: the missile and the Weapon Control System (WCS). The Missile Segment includes core subsystems such as the warhead, propulsion system, guidance system, destruct system, and structural components. These elements work together to ensure the missile's functionality and performance:

1. Warhead: Houses the payload and initiation mechanisms, presenting risks such as premature

or failed detonation.

2. Guidance System: Ensures accurate navigation and trajectory control, relying on onboard sensors and computational components.

3. Propulsion System: Provides thrust through solid or liquid fuel, with associated hazards like fuel leakage or ignition failure.

4. Destruct System: Allows for controlled missile destruction in case of malfunction, with risks of inadvertent activation or failure to execute.

5. Structural Components: Include the missile body and fins, ensuring aerodynamic stability and mechanical integrity.

The WCS Segment encompasses the external command and control systems, including the operator's console, radar, and system computer. These elements provide real-time operational oversight and facilitate communication with the missile during deployment and flight.

Operational Phases

The operational lifecycle of the ACE Missile System is divided into several critical phases (Ericson, 2005):

1. Storage and Transportation: Secure storage in land or shipboard facilities and safe transportation between locations.

2. Installation and Standby: Integration of the missile into launch tubes and maintenance of an alert state.

3. Launch and Flight: Execution of the launch sequence and real-time navigation to the designated target.

This phased approach reflects the complexity of operations, emphasizing the need for meticulous planning and hazard mitigation at each stage.

Key Hazards and Risks

Ericson's case study outlines a range of hazards associated with the ACE Missile System, categorized by subsystem:

1. Warhead Hazards: Premature detonation caused by external triggers, such as electromagnetic interference and failure to detonate, rendering the missile ineffective.

2. Propulsion Hazards: Fuel leaks that lead to fires, explosions, and ignition failures compromising thrust and trajectory.

3. Guidance System Hazards: Signal interference or software errors that result in navigation failures and off-course trajectories.

4. Structural Hazards: Failures caused by mechanical stress, corrosion, or damage during handling and storage.

5. Destruct System Hazards: Accidental activation of the destruct mechanism and failure to destruct, causing unintended consequences.

6. Launch System Hazards: Misalignments or malfunctions in the launch mechanism, leading to inaccurate targeting or instability during deployment.

These hazards underscore the interconnected nature of missile systems, where the failure of one subsystem can propagate across the entire system.

Preliminary Hazard Analysis

The preliminary hazard analysis (PHA) technique is a safety analysis tool for identifying hazards, their associated causal factors, effects, level of risk, and mitigating design measures when detailed design information is unavailable. The PHA provides a methodology for identifying and collating hazards in the system and establishing the initial system safety requirements (SSRs) for design from preliminary and limited design information. The PHA intends to affect the safety design as early as possible in the development program. The methodology for the PHA is listed in Table 1. The inputs to the PHA are design knowledge, hazard knowledge, Preliminary Hazard List (PHL) which is the initial step in the hazard analysis process, aiming to identify potential hazards early in the system lifecycle, and top-level Mishaps (TLMs) which are the most undesirable events. Ericson provides these inputs in his text. The outputs of the PHA are captured in a PHA Worksheet. The structure of this worksheet is described in Table 2. Risk measures are the product of mishap severity and probability and are depicted in Table 3.

Table 1. PHA Methodology

Step	Task	Description
1	Define System	Define, scope, and bound the system. Define the mission, mission phases, and mission environments. Understand the system design, operation, and major system components.
2	Plan PHA	Establish PHA definitions, worksheets, schedules, and processes. Identify system elements and functions to be analyzed.
3	Establish safety criteria.	Identify applicable design safety criteria, safety precepts/principles, safety guidelines, and safety-critical factors.
4	Acquire data.	Acquire all of the necessary design, operational, and process data needed for the analysis.
5	Conduct PHA.	a. List and evaluate each PHL and TLM for hazards. b. Identify new hazards. c. Evaluate hazards as thoroughly as design detail allows. d. Document process.
6	Evaluate risk.	Identify the level of mishap risk presented for each identified hazard, both with and without hazard mitigations in the system design.
7	Recommend corrective	Recommend corrective action necessary to eliminate or mitigate identified hazards. Work with the design organization to translate the recommendations into SSRs. Also, identify safety features already in the

Step	Task	Description
	action.	design or procedures that are present for hazard mitigation.
8	Monitor corrective action.	Review test results to ensure safety recommendations and SSRs effectively mitigate hazards as anticipated.
9	Track hazards.	Transfer newly identified hazards into the HTS. Update the HTS as hazards, hazard causal factors, and risks are identified in the PHA.
10	Document PHA.	Document the entire PHA process and PHA worksheets in a PHA report. Include conclusions and recommendations.

Table 2. PHA Worksheet

Column Name	Description
System	This entry identifies the system under analysis.
Subsystem/Function	This entry identifies the subsystem or function under analysis.
Hazard Number	identifies the number assigned to the identified hazard
Hazard	the specific hazard being postulated and evaluated.
Causes	Identifies conditions, events, or faults that could cause the hazard to exist and the events that can trigger the hazardous elements to become a mishap or accident.
Effects	identifies the effects and consequences of the hazard, should it occur.
Mode	identifies the system mode(s) of operation, or operational phases, where the identified hazard is of concern.
Initial Mishap Risk Index (IMRI)	This provides a qualitative measure of mishap risk significance for the potential effect of the identified hazard, given that no mitigation techniques are applied to it.
Recommended Action	Establishes recommended preventive measures to eliminate or mitigate the identified hazards
Final Mishap Risk Index (FMRI)	Provides a qualitative measure of mishap risk for the potential effect of the identified hazard, given that mitigation techniques and safety requirements are applied to the hazard.
Comments	Provides a place to record useful information regarding the hazard or the analysis process
Status	The current status of the hazard is that it is either open or closed.

Table 3. Risk Measures

Severity	Probability
I. Catastrophic	A. Frequent
II. Critical	B. Probable
III. Marginal	C. Occasional
IV. Negligible	D. Remote
	E. Improbable

METHOD

A systematic review of literature was done on Journal articles in the last 10 years on the following sites. Engineering Village, Arxiv, IEEE, and Google Scholar. The searches were done with Table 1 & and Table 2 in mind. A structured methodology should be followed to compare the hazard analysis generated by ChatGPT to the analysis in Ericson, Clifton's 2005 publication Hazard

Analysis Techniques for System Safety. This ensures an objective evaluation of the depth, accuracy, and alignment of ChatGPT's outputs with Ericson's techniques.

1. Define Evaluation Criteria

Establish clear criteria to compare the two analyses effectively. Key aspects include:

1) Completeness:
   - Does the hazard analysis identify all significant hazards outlined in Ericson's example?
   - Are potential sources of hazards adequately explored?
2) Accuracy:
   - Does ChatGPT identify the hazards consistent with those in Ericson's analysis?
   - Are the severity and likelihood categories appropriately assessed?
3) Methodological Alignment:
   - Does ChatGPT follow the methodologies described in MIL-STD-882E and Ericson's publication?
   - Are the identified hazards structured similarly to those in Ericson's example?
4) Mitigation Strategies:
   - Are the proposed mitigation strategies as robust and practical as those in Ericson's example?
   - Do both analyses use similar approaches to address identified risks?
5) Terminology and Frameworks:
   - Does ChatGPT use the language and categorization consistent with MIL-STD-882E and Ericson's frameworks?
6) Contextual Application:
   - Does ChatGPT accurately reflect the system context (e.g., ACME Missile System)?
   - Are system-specific details adequately incorporated?

2. Perform a Side-by-Side Comparison

Using a tabular or structured approach, directly compare each component of the hazard analysis. For example:

Table 4. Side by Side Comparison of Hazard Analyses

Evaluation Aspect	ChatGPT-Generated Analysis	Ericson's Analysis	Observations
Hazard Identification	List of hazards from ChatGPT output.	List of hazards in Ericson's book.	Are all critical hazards identified in both?
Severity & Likelihood	Severity/likelihood ratings by ChatGPT.	Ratings in Ericson's example.	Are ratings consistent or diverging?
Mitigation Strategies	Mitigation approaches suggested by	Strategies outlined in Ericson.	Are strategies comparable in detail and practicality?

Evaluation Aspect	ChatGPT-Generated Analysis	Ericson's Analysis	Observations
	ChatGPT.		
Analysis Framework	Methods/tools used by ChatGPT (e.g., PHA).	Techniques in Ericson's example.	Does ChatGPT align with Ericson's methods?
Completeness of Analysis	Depth of hazard exploration by ChatGPT.	Scope and depth in Ericson's book.	Are there missing or extra hazards in ChatGPT's output?

3. Test for Contextual Understanding

Evaluate whether ChatGPT effectively incorporates the context of the **ACME Missile System**:

- Review system-specific hazards (e.g., propulsion failures, payload risks).
- Compare how well each analysis addresses unique system interactions.

4. Analyze Methodology Alignment

Cross-check the methods used by ChatGPT against Ericson's recommended techniques:

1) Preliminary Hazard Analysis (PHA):

- Does ChatGPT capture high-level hazards as described by Ericson?

2) Functional Hazard Analysis (FHA):

- Are functional failures identified and assessed with similar depth?

3) Subsystem Hazard Analysis (SSHA):

- Does ChatGPT dive into subsystem-level risks with similar granularity?

5. Assess Gaps and Strengths

Identify strengths and weaknesses of ChatGPT's analysis compared to Ericson's:

- Strengths:
  - Does ChatGPT capture additional hazards or suggest innovative mitigations?
- Weaknesses:
  - Are certain critical hazards or mitigation steps missing in ChatGPT's output?

6. Visualize Results

Summarize the comparison with:

- Tables: Highlighting matches and discrepancies.
- Charts: Displaying coverage or alignment (e.g., a bar graph showing the percentage of hazards identified).
- Narrative Analysis: Explaining major findings and gaps.

7. Incorporate Expert Validation

- Involve safety engineering experts to review both analyses and provide insights on their relative

accuracy and applicability.

- Use Ericson's analysis as the benchmark for expert judgment.

Example Summary

Objective: Determine whether ChatGPT can replicate the hazard analysis techniques and results described by Ericson for the ACME Missile System.

RESULTS AND DISCUSSION

- ChatGPT aligns with Ericson in identifying [X]% of the hazards and matching [Y]% of the severity/likelihood ratings.
- Divergences occur in [specific areas], indicating gaps in [functional understanding/methodological application].
- ChatGPT offers additional [innovations/errors] in mitigation strategies.

- **Recommendations**:
  - Use ChatGPT for initial hazard identification but validate findings against established methodologies like those in Ericson's work.

By systematically evaluating ChatGPT's output against Ericson's analysis, this approach ensures a thorough understanding of ChatGPT's strengths and limitations in hazard analysis.

Preliminary Hazard Analysis Worksheet

Subsystem/ Function	Hazard No.	Hazard	Causes	Effects	Mode	IMRI
Missile Structure	PHA-1	Structural failure during flight	Manufacturing defect, design error	Missile crash, death/injury	Flight	1D
Warhead	PHA-2	Premature detonation due to extreme vibration	Structural failure, improper mounting	Explosion, personnel injury	Flight, Transport	1D
Destruct Subsystem	PHA-3	Delayed destruct command	Delayed destruct command	Communicat ion delay, processing lag	Extended risk to unintended areas	1D
Fuel Subsystem	PHA-4	Fuel contamination	Improper storage, aging infrastructure	Reduced thrust, engine failure	Storage, Maintenan ce	2D
Electrical Systems	PHA-5	Electrical short circuit	Damaged wiring, environmental exposure	Subsystem failure, potential fire	All Phases	1D
Radar	PHA-6	Radar blackout due to environmental factors	Fog, rain, or electromagnetic interference	Loss of situational awareness	Flight	2D
Guidance	PHA-7	Loss of	GPS signal	Off-course	Flight	1D

Subsystem/ Function	Hazard No.	Hazard	Causes	Effects	Mode	IMRI
and Navigation		navigation control	interference, software bug	trajectory, mission failure		
Operator Console	PHA-8	Unauthorized access	Weak cybersecurity measures	Incorrect commands, system misuse	Standby, Launch	1C
Thermal Management	PHA-9	Overheating of onboard electronics	Insufficient cooling, environmental heat	System shutdown, fire risk	All Phases	2D
Launch Platform	PHA-10	Platform instability during launch	Uneven terrain, mechanical failure	Off-target launch, system damage	Launch	1D
Battery Systems	PHA-11	Battery overheating	Overcharging, damaged cells	Fire, subsystem failure	Maintenance, Operation	1D
Payload Deployment	PHA-12	Failure to deploy the payload	Mechanical failure, software glitch	Mission failure, wasted payload	Flight	1C
Communication Subsystem	PHA-13	Data loss during the mission	Signal interference, hardware failure	Loss of situational awareness	All Phases	1D
Software	PHA-14	Unintended system activation	Software bugs, external hacking	Unsafe system behavior	All Phases	1D
Environmental Control	PHA-15	Corrosion of critical components	Prolonged exposure to moisture or chemicals	Structural failure, reduced reliability	Storage, Transport	1E
Transport	PHA-16	Damage during transportation	Inadequate securing, mishandling	Structural damage, increased hazards	Transport	1D

Key Additions

1. Thermal Management: Added risks of overheating electronics and batteries.
2. Payload Deployment: Risks of mechanical or software failures preventing payload deployment.
3. Environmental Risks: Corrosion of critical components due to moisture exposure.
4. Unauthorized Access: Cybersecurity vulnerabilities lead to potential system misuse.
5. Platform Stability: Launch instability due to terrain or mechanical issues.

Next Steps

- Validate the mitigations in real-world testing scenarios.
- Refine risk levels (IMRI and FMRI) based on updated operational data.
- Develop more targeted training and procedural guidelines for hazard-prone phases.

This expanded PHA provides a deeper understanding of the potential risks in the ACE Missile System and outlines comprehensive strategies for mitigation. Let me know if you need further refinements or additions!

Comparison of ChatGPT-Generated PHA and Document PHA for the ACE Missile System

The performance of ChatGPT in generating a Preliminary Hazard Analysis (PHA) for complex systems demonstrates notable alignment with traditional methodologies, such as those outlined in standard documents like MIL-STD-882E. Furthermore, ChatGPT expands the scope of conventional analyses by addressing emerging risks associated with modern technological advancements. This section evaluates the alignment, differences, and scope expansion observed in ChatGPT's PHA output compared to document-based analyses.

Alignment with Hazard Identification

ChatGPT's PHA aligns closely with traditional analyses, accurately identifying critical hazards commonly associated with complex missile systems. Key hazards identified by both ChatGPT and the reference document include inadvertent warhead initiation, fuel system malfunctions (e.g., leaks and ignition failures), and structural failures under operational stress. These hazards represent core safety concerns within missile systems and are consistently addressed across both approaches.

Additionally, both analyses emphasize risks such as electromagnetic interference and battery-related fire hazards. These alignments demonstrate ChatGPT's ability to effectively capture foundational risks in missile systems, reflecting a solid understanding of established hazard analysis frameworks.

Enhanced Detail in Mitigation Strategies

While the traditional document's PHA provides detailed causes, modes, and effects of hazards, ChatGPT's analysis augments this by introducing comprehensive mitigation strategies. For example, ChatGPT elaborates on modern approaches such as implementing anti-jamming technologies to mitigate electromagnetic interference and incorporating advanced thermal protection systems to safeguard high-voltage electronics. These contributions indicate that ChatGPT not only identifies risks but also enhances risk management through forward-looking solutions aligned with current technological capabilities.

Introduction of New Hazards

One of the most significant contributions of ChatGPT's PHA is its identification of additional hazards not explicitly covered in the document analysis. These include:

1. Thermal Management Risks:
   - Overheating of onboard electronics, especially in prolonged operations or high-stress environments.
   - Proposed mitigations include advanced heat dissipation materials, real-time thermal monitoring systems, and redundant cooling mechanisms.
2. Cybersecurity Vulnerabilities:
   - Unauthorized access to the operator console or weapon control system, potentially compromising command integrity.
   - Mitigation strategies include enhanced encryption protocols, multi-factor authentication, and regular system penetration testing.
3. Payload Deployment Failures:
   - Mechanical or electronic malfunctions leading to the inability to release the payload.
   - Mitigation measures such as redundant deployment mechanisms and pre-flight testing were suggested.

These additional hazards reflect ChatGPT's capacity to account for risks arising from contemporary technological advancements, broadening the scope of traditional hazard analyses.

Broadening the Analytical Scope

ChatGPT's ability to identify emerging risks such as cybersecurity vulnerabilities and advanced thermal management issues signifies an evolution in hazard analysis capabilities. Traditional hazard analysis frameworks often focus on physical and mechanical risks, whereas ChatGPT's inclusion of digital and systemic risks acknowledges the growing complexity of modern systems. This expanded scope aligns the analysis with the demands of evolving military and aerospace technologies, where digital transformation introduces new vulnerabilities.

Implications for System Safety

The observations suggest that ChatGPT has the potential to complement traditional hazard analysis by:

- Streamlining the identification of foundational risks.
- Enhancing the depth and breadth of mitigation strategies.
- Addressing emerging risks associated with modern technologies, which may be overlooked in conventional analyses.

However, these findings also highlight the importance of domain experts' validation to

ensure the accuracy, relevance, and applicability of AI-generated analyses. Integrating AI tools like ChatGPT into system safety workflows could provide substantial benefits, but careful oversight is necessary to mitigate potential limitations.

Comparison Table

**Aspect	ChatGPT PHA	Document PHA
Structural Failure	Identified during flight; recommends quality assurance and redundancy testing.	Same; highlights manufacturing defects and design errors.
Fuel System	It adds fuel contamination risks and suggests filtration and regular testing.	Focuses on fuel tank leakage and ignition, focusing on material improvements.
Warhead Risks	Includes premature initiation and failure to initiate with detailed mitigation strategies like dual-arm signals.	Similar focus but emphasizes external triggers like bullets, heat, and shrapnel.
Missile Destruct	Addresses both inadvertent and failed to destruct; suggests dual-command controls and RF shielding.	Matches with a focus on command errors and transmission faults.
Thermal Management	Identifies overheating of electronics as a hazard; proposes enhanced cooling and thermal sensors.	Not explicitly covered.
Cybersecurity Risks	Highlights unauthorized access and system misuse; suggests multi-factor authentication and penetration testing.	Not explicitly covered.
EMR Hazards	Matches risks like personnel injury and ignition of explosives; recommends shielding and operational safe zones.	Same risks but lacks detailed mitigation strategies.
Battery System	Identifies fire risks from electrolyte leakage and overcharging; suggests tamper-resistant designs.	Covers leakage risks but lacks specifics on tamper-proofing.

Strengths and Gaps

The application of ChatGPT to Preliminary Hazard Analysis (PHA) offers valuable insights into its capabilities as a complementary tool for traditional, document-based hazard analysis. Both approaches exhibit distinct strengths, while each has gaps that highlight the limitations of AI and traditional methodologies in addressing the complexities of safety-critical

systems. This section examines the comparative strengths and gaps of ChatGPT-generated and document-based PHAs, providing a framework for understanding their potential integration.

Strengths of Document-Based PHA

- Comprehensive Hazard Identification

Traditional document-based PHAs excel in systematically identifying hazards, their causes, and effects. These analyses often adhere rigorously to established standards, such as MIL-STD-882E, ensuring alignment with best practices in system safety.

- Detailed Mishap Risk Assessments

A hallmark of document-based PHAs is the inclusion of quantitative metrics, such as the Initial Mishap Risk Index (IMRI) and Final Mishap Risk Index (FMRI). These indices provide detailed risk assessments based on severity and likelihood, enabling precise prioritization of hazards and mitigation strategies.

- Adherence to Established Frameworks

Document-based analyses are deeply rooted in methodologies outlined by safety standards and frameworks. This rigor ensures consistency, reliability, and regulatory compliance, making them a trusted foundation for safety practices.

Strengths of ChatGPT-Generated PHA

- Inclusion of Emerging Risks

ChatGPT introduces hazards associated with modern technologies not explicitly addressed in traditional PHAs. Examples include cybersecurity vulnerabilities, thermal management issues, and risks related to digital control systems. This expanded scope demonstrates ChatGPT's adaptability to the evolving landscape of safety-critical systems.

- Innovative Mitigation Strategies

ChatGPT provides forward-looking mitigation strategies tailored to address both traditional and modern hazards. For instance, advanced encryption protocols and multi-factor authentication are suggested for cybersecurity risks, while real-time thermal monitoring systems are proposed to mitigate overheating risks. These recommendations highlight ChatGPT's capacity to integrate contemporary solutions into hazard analysis.

- Emphasis on Redundancy and Testing

A notable feature of ChatGPT's PHA is its focus on redundancy and iterative testing. This emphasis aligns with best practices in managing uncertainties associated with emerging technologies, ensuring robust safeguards against potential failures.

Gaps in Document-Based PHA

- Limited Exploration of Modern Risks

Traditional PHAs often focus on physical and mechanical hazards, with limited emphasis on risks associated with emerging technologies. For instance, cybersecurity vulnerabilities and software-specific hazards are underrepresented, potentially leaving critical gaps in contemporary systems analysis.

- Depth of Mitigation Strategies

While document-based PHAs provide detailed risk assessments, their mitigation strategies can lack the depth required to address complex, modern challenges. For example, risks related to electromagnetic interference (EMI) or high-voltage electronics may not include comprehensive countermeasures leveraging current technological advancements.

Gaps in ChatGPT-Generated PHA

- Absence of Quantitative Risk Metrics

ChatGPT cannot generate detailed quantitative assessments such as IMRI and FMRI scores. These metrics are essential for prioritizing hazards and determining the relative effectiveness of proposed mitigation strategies. The absence of such metrics limits the precision of ChatGPT's analyses.

- Potential Overlap or Deviation

Some hazards identified by ChatGPT may overlap with existing hazards or deviate from the intended system focus. For instance, broad categorizations of risks can dilute efforts to address critical system-specific issues, leading to inefficiencies in resource allocation.

- Reliance on Generalized Knowledge

While ChatGPT can identify a wide array of hazards, its reliance on pre-trained data limits its capacity to address system-specific intricacies without supplementary inputs or domain-specific fine-tuning.

Conclusion

The comparative analysis of ChatGPT and document-based PHAs reveals their complementary strengths and unique limitations. Traditional PHAs excel in their systematic rigor, adherence to established frameworks, and quantitative assessments. Conversely, ChatGPT offers expanded coverage of emerging risks, innovative mitigation strategies, and a focus on modern challenges, particularly cybersecurity and digital systems.

Integrating ChatGPT into hazard analysis processes can enhance efficiency, broaden the scope of risk identification, and introduce contemporary solutions. However, these benefits must be balanced with the need for expert oversight to validate AI-generated analyses and address their limitations. Together, these approaches can offer a synergistic pathway for improving the comprehensiveness and adaptability of system safety practices in increasingly complex

technological landscapes.

CONCLUSION

The ChatGPT-generated PHA complements the document's analysis by addressing additional hazards and modern risks while maintaining alignment with the core hazards outlined in the document. Together, these analyses provide a more comprehensive safety perspective for the ACE Missile System. Combining insights from both sources would strengthen hazard mitigation planning and enhance system safety.

REFERENCES

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On The Dangers Of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language Models Are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.

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

Department of Defense (DOD). (2010). *Joint Software Systems Safety Handbook (SSSH)*. Naval Ordnance Safety and Security Activity

Dhamani, N. (2024). *Introduction To Generative AI* (1st ed.). Manning Publications Co. LLC.

Domkundwar, I., Mukunda, N. S., & Bhola, I. (2024). Safeguarding AI Agents: Developing And Analyzing Safety Architectures. arXiv. https://arxiv.org/abs/2409.03793

Ericson, C. A. (2016). *Hazard Analysis Techniques For System Safety* (2nd ed.). Wiley.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. The MIT Press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative Adversarial Nets. *Advances in Neural Information Processing Systems*, 27, 2672–2680.

Gupta, N. K., Chaudhary, A., Singh, R., & Singh, R.. "ChatGPT: Exploring the Capabilities and Limitations of a Large Language Model for Conversational AI," *2023 International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT)*, Faridabad, India, 2023, pp. 139-142, doi: 10.1109/ICAICCIT60255.2023.10465811.

Martelaro, N., Smith, C. J., & Zilovic, T. (2022). Exploring Opportunities in Usable Hazard Analysis Processes for AI Engineering. arXiv preprint arXiv:2203.15628.

Nouri, A., Cabrero-Daniel, B., Törner, F., Sivencrona, H., & Berger, C. (2024). Engineering Safety Requirements For Autonomous Driving With Large Language Models. *2024 IEEE 32nd International Requirements Engineering Conference (RE)*, Reykjavik, Iceland, 218–228. https://doi.org/10.1109/RE59067.2024.00029

Nouri, A., Cabrero-Daniel, B., Törner, F., Sivencrona, H., & Berger, C. (2024). Welcome Your New AI Teammate: On Safety Analysis By Leashing Large Language Models. *2024 IEEE/ACM 3rd International Conference on AI Engineering – Software Engineering for AI (CAIN)*, Lisbon, Portugal, 172–177.

Phoenix, J., & Taylor, M. (2024). *Prompt Engineering For Generative AI* (1st ed.). O'Reilly Media.

Qi, Y., Zhao, X., Khastgir, S., & Huang, X. (2023). Safety Analysis In The Era Of Large Language Models: A Case Study Of STPA Using ChatGPT. arXiv. https://arxiv.org/abs/2304.01246

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving Language Understanding By Generative Pre-Training. *OpenAI Technical Report*.

Roumeliotis KI, Tselikas ND. ChatGPT and Open-AI Models: A Preliminary Review. Future Internet. 2023;15(6):192-. doi:10.3390/fi15060192

Russell, S. J., Norvig, P., & Davis, E. (2010). *Artificial Intelligence: A Modern Approach* (3rd ed.). Prentice Hall.

Santhosh, R., Abinaya, M., Anusuya, V., & Gowthami, D.. "ChatGPT: Opportunities, Features and Future Prospects," *2023 7th International Conference on Trends in Electronics and Informatics (ICOEI)*, Tirunelveli, India, 2023, pp. 1614-1622, doi: 10.1109/ICOEI56765.2023.10125747.

Sivakumar, M., Boaye Belle, A., Shan, J., & Khakzad Shahandashti, K. (2023). GPT-4 and Safety Case Generation: An Exploratory Analysis. arXiv. https://arxiv.org/abs/2312.05696

Solanki, S. R., & Khublani, D. K. (2024). *Generative Artificial Intelligence: Exploring The Power And Potential Of Generative AI* (1st ed.). Apress. https://doi.org/10.1007/979-8-8688-0403-8

Stephans, R. A. (2004). System Safety For The 21st Century: The updated and revised edition of *System Safety 2000* (2nd ed.). Wiley.

Summary Of The 2018 Department of Defense *Artificial Intelligence Strategy: Harnessing AI to Advance Our Security and Prosperity*. Department of Defense.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.

Weidinger, L., Rauh, M., Marchal, N., Manzini, A., Hendricks, L. A., Mateos-Garcia, J., Bergman, S., Kay, J., Griffin, C., Bariach, B., Gabriel, I., Rieser, V., & Isaac, W. (2023). Sociotechnical Safety Evaluation Of Generative AI Systems. arXiv. https://arxiv.org/abs/2310.11986

Wu, T., et al. (2023). A Brief Overview Of Chatgpt: The History, Status Quo And Potential Future Development. *IEEE/CAA Journal of Automatica Sinica*, 10(5), 1122–1136. https://doi.org/10.1109/JAS.2023.123618

Yazdi, M., Zarei, E., Adumene, S., & Beheshti, A. (2024). Navigating The Power Of Artificial Intelligence In Risk Management: A Comparative Analysis. *Safety*, 10(2), 42. https://doi.org/10.3390/safety10020042