AI Safety Under Uncertainty: Hallucinations and Unpredictable Failures

Joon S. Park
School of Information Studies, Syracuse University, NY, USA

jspark@syr.edu

Abstract: With the rise of AI-supported tools across mission-critical workflows in medicine, finance, commerce, education, and cybersecurity, their errors and incorrect decisions can pose safety risks. In our broader study of safety challenges in AI applications, we identify and analyze various safety concerns related to AI-supported tools, including hidden dangers in AI-generated content, the misuse of AI-supported tools for cyberattacks, and their societal impacts. In this particular work-in-progress paper, we focus on our analyses of unpredictable failures and hallucinations in AI-supported systems. We analyze how generative AI models can produce fluent, convincing, yet misleading results, and how they can contaminate mission-critical applications, such as healthcare/medical decision-making, software development, cybersecurity, privacy/data governance, and workflows with agentic AI-supported systems. We survey use cases across domains to articulate the vulnerabilities. We discuss how agentic AI-supported workflows can amplify small errors into significant damage by automatically feeding erroneous outputs and actions. We also propose practical countermeasures to minimize the hallucinations and unpredictable failures with continuous monitoring and evaluation. In our future work, we will integrate this analysis with other areas of AI safety to develop a comprehensive framework and strategies.

Keywords: AI safety, AI risk analysis, AI governance

1. Introduction

Today, Artificial intelligence (AI) technologies are influencing decision-making in various applications, including the financial services sector, educational platforms, predictive algorithms for cybercrime detection, and other productivity enhancements (Park, 2025a; Park, 2025b). However, as machine-learning-derived systems settle into positions within workflows rather than coexisting as advisory systems in the background, hallucination errors may pose a critical issue rather than a simple technical failure, with a broader potential for influence, including safety, equity, privacy, organizational stability, and public confidence. AI hallucination is defined as responses returned by an AI system that are convincingly incorrect, inconsistent, inaccurate, or other facts misaligned with the context of the situation. AI safety may shift from a theoretical concern to a practical impact in responsible use-case scenarios across sectors, where erroneous output could ripple effect into material threat vectors. AI-driven safety risks can arise from numerous sources, such as unpredictable system behavior, misleading or incorrect outputs, misuse of AI for breaches and intrusions, privacy and ethical violations, and others.

This work-in-progress paper addresses a primary type of safety risks caused by AI's unpredictable hallucinations in AI-enabled applications. We discuss the unpredictability of AI failure modes and hallucinations, the reasons for their emergence, and their impacts in safety-critical settings. By analyzing technical vulnerabilities to real-world implications, we aim to support system developers, organizational decision-makers, and end users. We believe this study can help minimize potential risks through effective safeguards, policy expectations, and proper operations.

2. Related Work

Researchers define hallucinations as instances in which an advanced language model returns factually inaccurate results and demonstrates similarities to human cognitive errors, and address predictive mechanisms that both the model and its results share, making it prone to mistakes (Tripathi et al., 2024). They classify the AI's hallucination types, considering input-conflicting, context-conflicting, and fact-conflicting, and examine their implications through industry examples, particularly focusing on healthcare and life sciences. Both studies provide evidence of unpredictable failures and risk breakdowns in AI-enabled systems generating content. Some researchers concerned with the physical and digital safety in AI-enabled systems take a more skeptical perspective, arguing that AI will remain a disruptive technology. They further emphasize that advances in deep learning, particularly multi-layer neural networks, computer vision, and anomaly detection, have driven significant performance gains in health monitoring, public safety, and cybersecurity (Simmons & Park, 2024). More researchers emphasize the urgent need to develop AI-enabled systems that effectively control safety throughout data collection, model training, and deployment (Salhab et al., 2024). They also

identify critical safety concerns, including explainability, interpretability, robustness, reliability, bias, and adversarial attacks, and present the need to establish a comprehensive AI safety framework.

3. Unpredictable AI Failures and Hallucinations

AI-enabled systems may fail in unpredictable, unreproducible, and unexplainable ways in safety-critical and high-stakes environments. Hallucination is one of the most common ways such failure arises. Hallucinations are especially dangerous because we may simply accept a well-written claim that sounds confident and polished, even when it is false or unverified. Modern generative AI models are likely trained and optimized to continue a prompt that sounds contextually appropriate, rather than to verify the relations between claims and corresponding evidence. Therefore, when the input is incomplete, ambiguous, or even internally inconsistent, the AI models may "fill in" the missing content that seems related to the expected results. From a systems assurance perspective, this may create a more serious problem, especially in mission-critical systems (Farahmand, 2023). A major source of uncertainty is the gap between how a generative model is trained and how it is actually used in the system environment. Models typically inherit blind spots from their training data, and the AI-enabled system can make critical errors at runtime due to them. Sometimes, ordinary changes, such as new user types, workflow shifts, or different operating environments, may trigger unexpected problems. Malicious users can provide adversarial inputs or crafted prompts that can lead the model to produce hallucinations or unsafe actions (Majeed & Hwang, 2023). People often treat AI output as a time-saving substitute for careful review, especially when it looks polished on the surface. As a result, small errors or shaky assumptions can transmit faster and farther than they would in a purely human workflow.

For instance, in healthcare AI-enabled systems, analyses of hallucinated outputs have found recurring patterns that consist of confident claims about biomarkers or treatments with no solid support, inaccurate summaries of clinical details, and even polished, "guideline-like" recommendations that look professional but aren't reliably grounded in evidence (Aditya, 2024; Verma et al., 2025). In software engineering, a generative AI might produce flawed code, unsafe configurations, or misleading vulnerability descriptions that seem professional and correct. This might be benign in normal situations, but under adversarial conditions, with crafted inputs or exploit attempts, the same code can break in dangerous ways or provide attackers with advantages. Furthermore, AI-supported cybersecurity management can increase risk by insecure patterns, such as weak authentication procedures, insecure parsing, or improper cryptographic usage. It can also deliver security advice with unverified confidence (Gupta et al., 2023; Bhuiyan & Park, 2025). AI-supported intrusion detection systems (IDS) and monitoring can also fail in unpredictable ways, and when they do, the fallout can be serious. The damage from a persuasive hallucination is not only bad information, but also bad process, such as misdirected decisions, misplaced priorities, and poor allocation of defensive resources (Dhanushkodi & Thejas, 2024). A poor configuration or inaccurate filtering of a vulnerability can rapidly spread across teams, tools, and systems, leading to correlated vulnerabilities and a growing attack surface.

Another concern about hallucinations is how they interact with governance and privacy. When a generative AI confidently, but incorrectly, interprets what counts as private data, how long data can be retained, or what information can be shared with others, it can unintentionally guide an organization into policy violations or legally questionable practices. Researchers address that AI's ability to profile, reason, and process data at scale raises privacy concerns, and that misleading AI guidance about what is safe/acceptable can substantially amplify those concerns. From a more practical perspective, the concern is that an AI tool can provide a highly convincing narrative that falsely reassures people about governance and privacy.

Finally, as AI-supported systems become more agentic, where multiple AI agents work in a chained workflow (i.e., outputs from one step become inputs to other steps) and more deeply embedded in mission-critical environments, hallucinations can cause greater damage by amplifying actions rather than simply spreading misinformation. Systems that flow model outputs into plans, decisions, actions, and feedback loops can escalate the potential damage caused by even small mistakes. A single hallucinated assumption about system state, constraints, decisions, or interdependencies can cascade through subsequent steps in the workflow and become integrated with real operational processes. For instance, an informational error in an agentic AI-supported system can quickly escalate into an operational failure with serious consequences (Xenos & Serpanos, 2025).

4. Conclusions and Future Strategies

In our broader study of safety challenges in AI applications, we identify and analyze additional safety concerns with AI-supported tools, including hidden dangers in AI-generated content, the misuse of AI-supported tools

for cyberattacks, and their broader societal impacts. In this particular work-in-progress paper, we focused on our analyses of unpredictable failures and hallucinations in AI-supported systems, including their types, causes, and impacts on workflows, especially in mission-critical applications. We characterized how generative AI models can produce fluent, convincing, yet misleading results, and contaminate mission-critical applications, such as healthcare/medical decision-making, software development, cybersecurity, privacy/data governance, and workflows involving agentic AI-supported systems. In our future work, we will integrate this analysis with other areas related to AI safety to develop and introduce a comprehensive framework and strategies for AI safety.

In more reliable AI-supported systems, especially for mission-critical applications, layered assurance at the technical level should provide comprehensive validation and context-aware evaluation embedded at all stages, including data collection, curation, training, pre-deployment validation, and post-deployment monitoring. Evaluation and verification should consider the measures of real operational failures and runtime uncertainty, in addition to reference accuracy. In mission-critical applications, human oversight and workflow controls are necessary. Human-in-the-loop review should be mandatory when incorporated as a workflow decision point for high-consequence decisions, such as clinical recommendations. In cybersecurity fallbacks, baseline assumptions should be that AI-generated artifacts (e.g., code, playbooks, threat summaries) are unverified and require systematic review prior to acceptance for operational use, thereby minimizing exploitable opportunities for undetected vulnerabilities, false positives/negatives, or recommendations. Ultimately, AI safety rests on more than technical safeguards. It also depends on strong governance, clear accountability aligned with regulations, proper operation, and ongoing training. To prevent privacy violations and compliance gaps, organizations should implement principles into daily practice through training and continuous monitoring of connected components.

References

Aditya, G. (2024) Understanding and Addressing AI Hallucinations in Healthcare and Life Sciences, *International Journal of Health Sciences*, 7(3), pp. 1–11.

Bhuiyan, S., Park, J. S. (2025) Cybersecurity Threats and Mitigation Strategies in AI Applications. Journal of the Colloquium for Information Systems Security Education (CISSE), 12(1), 7. April.

Dhanushkodi, K. and Thejas, S. (2024) AI Enabled Threat Detection: Leveraging Artificial Intelligence for Advanced Security and Cyber Threat Mitigation, *IEEE Access*, 12, pp. 173127–173136.

Farahmand, F. (2023) A System Engineering Approach to AI Security and Safety, *Computer*, 56(11), pp. 118–122.

Gupta, M. et al. (2023) From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy, *IEEE Access*, 11, pp. 80218–80245.

Majeed, A. and Hwang, S.O. (2023) When AI Meets Information Privacy: The Adversarial Role of AI in Data Sharing Scenario, *IEEE Access*, 11, pp. 76177–76195.

Park, J. S. (2025a) Towards AI-augmented Teaching for Higher Education. In Proceedings of the 6th International Conference on Artificial Intelligence in Education Technology (AIET), July 29–31.

Park, J. S. (2025b) Course-Specific AI TA: Implementation and Impact on Student Learning. In Proceedings of the 7th International Workshop on Artificial Intelligence and Education (WAIE), September 27–29.

Simmons, T., Park, J. S. (2024) Cybersecurity Challenges and Opportunities with Generative AI. In Proceedings of the 23rd European Conference on Cyber Warfare and Security (ECCWS), June 27–28.

Salhab, W. et al. (2024) A Systematic Literature Review on AI Safety: Identifying Trends, Challenges, and Future Directions, *IEEE Access*, 12, pp. 131762–131784.

Tripathi, S., Griffith, H., & Rathore, H. (2024). Assessing hallucination in large language models under adversarial attacks. In Proceedings of the 9th International Conference on Mobile and Secure Services (MobiSecServ). IEEE.

Verma, A. et al. (2025) *Generative Artificial Intelligence and Cybersecurity Risks: Issues and Challenges ICT Analysis and Applications*. Springer, pp. 321–327.

Xenos, G. and Serpanos, D. (2025) Agentic AI: A New Paradigm for Cyber–Physical Systems Cybersecurity, *Computer*, 59(1), pp. 180–183.