AI Failures, Limitations, Incapabilities, and Threats

Submitted By
Abul Hasan Farrukh, Jayant Kumar Mishra, Atul Chauhan

Under The Supervision Of
Dr. Vijay Prakash

Master Of Computer Applications School Of
Computer Applications Babu Banarasi Das
University
Bbd City, Faizabad Road, Lucknow (U.P.) - 226028, India

Introduction:

Artificial intelligence (AI) is steadily weaving itself into military, economic, and social life, reshaping their very foundations in ways that many of us can already see unfolding around us. Given the scale and reach of these changes, it has become crucial to make sure AI systems are built in ways that are not only technically reliable but also ethically responsible and socially beneficial. One of the main goals of this discussion is to bring together the scattered pieces of knowledge about risks connected to AI, sharpen the key ideas so they are clearer, and reduce the difficulty of engaging with these issues by presenting them in a way that feels systematic yet understandable.

To make sense of AI safety, it helps to place the issue in the wider context in which these systems are created and used. The choices and interactions of people and institutions—whether they are developers, policymakers, military leaders, or other influential actors—will shape this landscape in decisive ways. Since AI touches so many areas, drawing on established frameworks offers us useful tools for looking at the actors involved, their relationships, and the wide-ranging consequences of AI. These frameworks are intentionally flexible, giving us ways to compare different kinds of intelligence, whether we are talking about individuals, corporations, governments, or autonomous machines.

Over the past ten years, the growth of AI has been striking. It has steadily moved into everyday use while becoming more accessible in commercial settings, raising the expectations of organizations eager to benefit from it. Unsurprisingly, the pace of adoption has increased sharply. Yet, research shows a sobering reality: many AI projects fail to deliver what was promised. For practitioners and researchers alike, this makes it even more important to figure out what separates successful projects from unsuccessful ones so that the real potential of AI can be unlocked.

Although there is already a large amount of research on why information systems (IS) projects succeed or fail – particularly in areas like Enterprise Resource Planning (ERP)—AI brings its own challenges that make those older models incomplete. The complexity of AI algorithms, combined with the sweeping organizational changes that usually come with introducing AI, means we need to rethink and expand the factors that predict success. In other words, the traditional IS success models need to be adjusted so they fit the realities of AI.

Motivation:

Artificial Intelligence (AI) has rapidly penetrated military, economic, educational, and social domains, becoming a fundamental driver of automation and decision-making. Its widespread integration has created immense expectations—yet many AI systems fail in real-world deployments due to conceptual flaws, design oversights, and unpredictable contextual challenges.

Despite significant advancements, AI still lacks core elements of human cognition, moral reasoning, adaptability, and comprehension. Failures such as incorrect predictions, unintended actions, biases, hallucinations, or harmful decisions highlight the gap between perceived AI capabilities and actual performance.

As organizations invest heavily in AI, identifying why AI systems fail, understanding their structural limitations, and recognizing emerging threats from misuse (e.g., cyber-attacks, surveillance, lethal autonomous systems) has become urgent.

Brief Literature Survey:

Early research in Information Systems (IS) identifies success and failure factors for technology adoption; however, AI introduces unique complexities—algorithmic opacity, data-driven learning, hardware constraints, and context dependency—which existing IS models cannot fully explain.

Scholars such as Bauer et al. examined organizational readiness for AI, while Baier et al. studied deployment challenges in machine learning. Further philosophical and cognitive science discussions—such as Searle's Chinese Room argument and critiques by François Chollet—highlight AI's inability to achieve true understanding or generalization.

The AI Failure Incident Database documents numerous cases of malfunction, ranging from omission and commission errors to inappropriate automated actions and cyber breaches. Meanwhile, literature on AI threats explores powerful uses and misuses in domains such as healthcare, education (AIED), and autonomous weapons.

Thus, contemporary literature points towards two urgent gaps:

1. Understanding why AI fails across technical and social dimensions.
2. Developing trustworthy and interpretable AI solutions.

Problem Formulation:

Although AI systems are designed to emulate human cognitive functions, several issues persist:

a. Failures in Real-World Performance

AI often fails due to:

- Omission errors (failure to act when required)
- Commission errors (taking unintended or incorrect actions)
- Misinterpretation of commands or environmental cues
- Hardware limitations and environmental variability
- Lack of moral or contextual judgment

b. Structural Limitations

- AI systems rely on statistical patterns, not comprehension, making them vulnerable when confronted with unfamiliar inputs, biased training data, or incomplete information.

c. Incapabilities in Human-Like Intelligence

AI lacks:

- Self-awareness
- Conscious reasoning
- Semantic understanding
- Cross-context generalization
- Moral or ethical judgment

d. Societal and Security Threats

Unchecked AI development can result in:

- AI-driven cyber-attacks
- Large-scale surveillance
- Manipulation of public behavior
- Labour displacement
- Threats from autonomous weapons
- Risks from self-improving AGI

Objectives:

1. To analyze the major causes of AI failures, including omission/commission errors, design flaws, data issues, and hardware constraints.
2. To evaluate the inherent limitations and incapabilities of AI, especially its lack of semantic understanding, general intelligence, and adaptive reasoning.
3. To identify existing and emerging threats arising from AI misuse, including cyber-attacks, surveillance risks, and impacts on human autonomy.

Methodology / Planning of Work:

The study will follow a structured, research-oriented methodology:

a. Literature Review

- Review foundational AI theory, cognitive science perspectives, and philosophical critiques.
- Analyze empirical studies on AI deployment success and failure.
- Examine documented incident reports (e.g., AI Failure Incident Database).

b. Classification of AI Failures

Failures will be categorized based on:

- Design errors (omission, commission)
- Algorithmic limitations (bias, hallucinations)
- Data-related issues (quality, noise, incompleteness)
- Operational constraints (hardware, environment)
- Ethical and interpretability gaps

c. Evaluation of AI Limitations and Incapabilities

- Analyze AI's lack of cognitive abilities using interdisciplinary frameworks.
- Compare AI reasoning processes with human cognition.
- Assess AI's dependency on statistical models and inability to generate original knowledge.

d. Threat Assessment

- Conduct a conceptual analysis of threats arising from AI misuse (cybersecurity, education, health, surveillance, autonomous systems).
- Use SWOT or risk analysis frameworks to identify weaknesses and vulnerabilities.

e. Synthesis and Reporting

- Integrate findings to present a coherent, research-backed argument.
- Draft recommendations for safer and more trustworthy AI systems.

Expected Outcomes:

1. Comprehensive understanding of AI failures through well-classified categories and real-world examples.
2. Clear identification of AI's inherent limitations, showing why AI cannot yet match human intelligence.
3. Insightful analysis of AI threats, including misuse scenarios and societal risks.

References:

1. Westenberger, J., Schuler, K., & Schlegel, D. (2022). Failure of AI Projects: Understanding the Critical Factors. Procedia Computer Science, 196, 69–76. Elsevier B.V.
https://doi.org/10.1016/j.procs.2021.11.074

2. Banerjee, D. N., & Chanda, S. S. (2020). AI Failures: A Review of Underlying Issues.
arXiv preprint arXiv:2008.04073.
https://doi.org/10.48550/arXiv.2008.04073

3. Azelya, I., & Filin, S. A. (2025). Introduction: The Role of AI in Transforming Management Research. Involvement International Journal of Business, 2 (1), 39–44.
https://doi.org/10.62569/iijb.v2i1.108

4. Williams, R., & Yampolskiy, R. V. (2021). Understanding and Avoiding AI Failures: A Practical Guide. Philosophies, 6(3), 53. MDPI.
https://doi.org/10.3390/philosophies6030053

5. Rice, J. (2022). Tesla Autopilot and Collisions with Stationary Emergency Vehicles: Car for the Future or Doomed for Failure? Case Study Report, Rochester Institute of Technology.

6. Uesato, J., Kumar, S. M., Szepesvári, C., Erez, T., Rusu, A. A., & Legg, S. (2020). Rigorous Agent Evaluation: An Adversarial Approach to Uncover Catastrophic Failures. Proceedings of the AAAI Conference on Artificial Intelligence.
https://doi.org/10.48550/arXiv.2008.04073