International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 12 | Dec - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

AI FAILURES, LIMITATIONS, INCAPABILITIES AND THREATS

Abul Hasan Farrukh[1], Atul Chauhan[2], Jayant Kumar Mishra[3]

School of Computer Applications, Babu Banarasi Das University, Lucknow, India

[1]abulhasanfarrukh@gmail.com , [2]atulc8948@gmail.com , [3]jayantmishra9695@gmail.com

Abstract— Artificial Intelligence (AI) has rapidly transformed military, economic, educational, and social sectors, enhancing automation and decision-making while also revealing critical failures, limitations, and threats. This study critically examines these challenges through an extensive review of literature, documented AI failure cases, cognitive science insights, and philosophical analyses. AI shortcomings are categorized into design flaws, algorithmic weaknesses, data biases, deployment challenges, and ethical and interpretability gaps. The research also highlights AI's inherent incapabilities, such as lack of true understanding, moral reasoning, and human-like intelligence. Additionally, it discusses risks like bias, surveillance, cyber threats, misinformation, job displacement, and autonomous system dangers, stressing the need for responsible governance.

Keywords — AI Failures, AI Mistakes and Disasters, AI Limitations, Generative AI Failures, Incapabilities of AI, What can AI do, Threats of AI, Risks of AI, Disadvantages of AI.

I. INTRODUCTION

Artificial Intelligence (AI) refers to the capability of machines and software systems to perform tasks that typically require human intelligence, such as reasoning, learning, perception, decision-making, and problem-solving. The concept of AI dates back to the mid-twentieth century when early computer scientists envisioned machines that could simulate human thought processes. Over the decades, AI has evolved from symbolic reasoning systems and rule-based expert systems to data-driven machine learning and deep learning models capable of handling complex, high-dimensional information.

In recent years, advances in computational power, availability of massive datasets, and breakthroughs in algorithms—particularly neural networks—have accelerated AI development at an unprecedented pace. AI systems are now embedded in everyday applications such as recommendation engines, speech recognition systems, autonomous vehicles, medical diagnosis tools, financial trading platforms, and intelligent tutoring systems. As a result, AI is no longer confined to research laboratories but has become a critical component of modern digital infrastructure.

AI has steadily become a foundational technology across multiple domains, including military operations, healthcare, education, business management, governance, and social media. Organizations increasingly rely on AI-driven decision-making systems to improve efficiency, reduce costs, and gain competitive advantage. Governments deploy AI for surveillance, predictive policing, and public service delivery, while industries integrate AI into supply chain optimization, customer analytics, and automation.

The growing dependence on AI systems has also led to rising expectations regarding their accuracy, objectivity, and reliability. AI is often perceived as neutral, rational, and superior to human judgment, particularly in data-intensive tasks. However, this perception can be misleading. AI systems inherit the assumptions, biases, and limitations embedded in their training data, algorithms,

and deployment contexts. As AI systems take on roles that directly affect human lives—such as diagnosing diseases, approving loans, grading students, or controlling autonomous vehicles—the consequences of failure become more severe. Even minor errors can lead to significant economic losses, ethical violations, or threats to human safety. Therefore, understanding the nature and causes of AI failures is essential for responsible deployment and governance.

One of the defining characteristics of contemporary AI development is the paradox between rapid technological progress and frequent project failure. While AI capabilities continue to advance, a significant number of AI projects fail to meet expectations, are abandoned after deployment, or produce harmful and unintended outcomes. Studies indicate that many organizations struggle to translate AI prototypes into reliable, scalable, and trustworthy systems.

AI failures manifest in various forms, including incorrect predictions, biased outputs, system crashes, hallucinated responses, omission errors (failure to act when required), and commission errors (taking inappropriate or harmful actions). These failures are often exacerbated by factors such as poor data quality, lack of interpretability, overreliance on automation, and insufficient understanding of AI limitations among users and decision-makers.

Another contributing factor is the growing gap between user expectations and actual AI capabilities. Media narratives and commercial incentives often portray AI as intelligent, autonomous, and near-human in its reasoning abilities. This portrayal encourages overconfidence among users and decision-makers, leading to inappropriate delegation of authority to AI systems. When such systems fail, the consequences are magnified due to reduced human oversight and misplaced trust.

The motivation for this research arises from the growing recognition that AI systems, despite their apparent sophistication, lack many fundamental aspects of human intelligence. AI does not possess self-awareness, consciousness, moral reasoning, or genuine understanding of meaning. It operates primarily through pattern recognition and optimization rather than comprehension or intentionality.

High-profile AI failures—such as biased facial recognition systems, autonomous vehicle accidents, flawed predictive policing tools, and unreliable medical AI systems—highlight the risks of overestimating AI capabilities.

This Furthermore, the misuse of AI introduces additional threats, including large-scale surveillance, manipulation of public opinion, cyber-attacks, and the development of lethal autonomous weapons.

For students, researchers, policymakers, and practitioners, there is an urgent need to systematically analyze why AI systems fail, what their inherent limitations are, and how their misuse can pose serious threats to individuals and society. This research aims to contribute to that understanding by integrating technical, cognitive, organizational, and ethical perspectives.

II. PAST WORK AND PROBLEM FORMULATION

A literature review plays a critical role in establishing the academic foundation of any research study. It enables the researcher to examine existing knowledge, identify research trends, understand theoretical perspectives, and locate gaps that justify the present work. In the context of Artificial Intelligence (AI), reviewing past work is particularly important due to the interdisciplinary nature of the field, which spans computer science, information systems, cognitive science, philosophy, ethics, and social sciences.

This chapter reviews prior research related to AI failures, limitations, and threats. It critically examines studies on information system failures, AI deployment challenges, algorithmic limitations, cognitive critiques, and documented AI failure incidents. The chapter then synthesizes these findings to formulate the research problem addressed in this dissertation.

Some of the past research of AI and employment gives a complex interaction between advancement of technologies, employees demand and job transformation. Most of the previous studies highlighted the risk of automation that displacing the routine and manual jobs. An estimation by Frey and Osborne report that, around 47% of U.S. occupations is at high risk by AI, which sparks the global concern about job loss and AI-driven.

Organizations Recent studies indicate that a large proportion of AI projects fail to transition successfully from experimentation to production. While AI prototypes may perform well in controlled environments, real-world deployment introduces challenges related to data variability, system integration, scalability, and user trust.

One of the most extensively studied aspects of AI failure is algorithmic bias. Bias occurs when AI systems produce systematically unfair or discriminatory outcomes due to biased training data, flawed feature selection, or inappropriate optimization objectives. Research has documented bias in facial recognition systems, hiring algorithms, credit scoring models, and predictive policing tools.

III. CLASSIFICATION AND ANALYSIS OF AI FAILURES

As Artificial Intelligence systems increasingly operate in real-world, high-stakes environments, understanding the nature of their failures becomes critical. Unlike traditional software systems, AI systems exhibit failures that are often probabilistic, context-dependent, and difficult to predict or explain. These failures arise not only from coding errors but also from data issues, algorithmic behaviour, environmental variability, and human–machine interaction.

A. Introduction

This chapter presents a structured classification of AI failures based on technical, operational, and cognitive dimensions. By categorizing failures into meaningful groups, the study aims to provide a clearer understanding of why AI systems fail and how such failures differ fundamentally from conventional information system failures.

As Artificial Intelligence systems move beyond experimental settings into large-scale deployment, understanding their failure modes becomes increasingly important. Unlike traditional software systems, AI applications operate under conditions of uncertainty, learn from data rather than explicit instructions, and often function autonomously once deployed. These characteristics make AI failures more complex, less predictable, and harder to diagnose than failures in conventional information systems.

Fig. 1 The Reality of AI in Business

B. Framework

AI failures can be broadly defined as situations where an AI system produces outcomes that are incorrect, harmful, misleading, unethical, or inconsistent with its intended purpose. Unlike deterministic systems, AI failures often occur even when the system functions "as designed," highlighting deeper structural issues.

The classification framework adopted in this study is informed by prior literature, incident databases, and the research methodology outlined in the synopsis.

C. Omission Error

Omission errors occur when an AI system fails to take an action that is required in a given situation. These failures are especially critical in safety-sensitive domains such as healthcare, autonomous driving, and industrial automation. For example, an AI-based medical diagnostic system may fail to flag a rare but life-threatening condition due to insufficient representation in training data.

Omission errors often stem from conservative decision thresholds, incomplete data coverage, or limited generalization ability. While designers may intentionally bias systems toward caution to reduce false positives, this can increase the risk of harmful inaction.

D. Commission Error

Commission errors occur when an AI system takes an incorrect, inappropriate, or harmful action. Examples include autonomous vehicles misinterpreting road conditions or AI systems generating incorrect yet confident recommendations.

Commission errors are particularly dangerous because they can create a false sense of reliability. Users may trust AI outputs without sufficient verification, leading to overreliance and reduced human oversight.

Both omission and commission errors illustrate the inherent trade-offs in AI system design, where reducing one type of error often increases the likelihood of the other.

IV. FUNDAMENTAL CONSTRAINTS AND RISKS OF AI

While Artificial Intelligence systems have achieved impressive performance in specific tasks, their limitations and vulnerabilities become increasingly evident as they are deployed in complex, real-world environments. The failures discussed in the previous chapter are not merely accidental or temporary shortcomings; many arise from fundamental structural and conceptual constraints inherent in current AI paradigms.

1) Introduction

The rapid evolution and deployment of Artificial Intelligence systems have intensified the need to critically examine their underlying constraints and associated risks. While AI technologies are often celebrated for their efficiency, scalability, and ability to process vast amounts of data, such strengths can obscure the limitations that emerge when these systems interact with complex real-world environments. Understanding these limitations is essential not only for improving technical performance but also for safeguarding ethical, social, and institutional values.

Moreover, the limitations of AI are closely linked to emerging threats. Technical fragility can amplify security risks, cognitive incapabilities can lead to ethical failures, and organizational overreliance can magnify the consequences of system errors. These interdependencies mean that limitations and threats cannot be analysed in isolation; they form a connected risk landscape that requires holistic assessment.

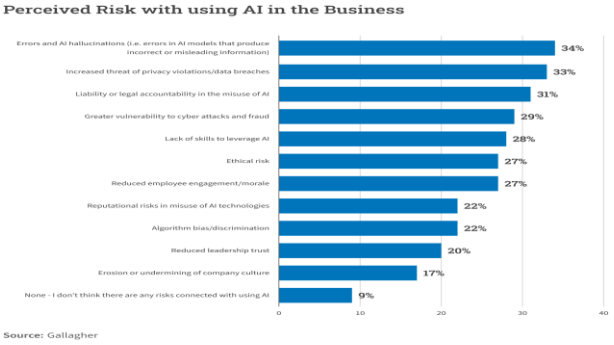

Fig. 2 Perceived Risk of AI in Business

2) Technical Limitations

By framing AI limitations and threats as interconnected phenomena, this chapter provides a critical foundation for evaluating the long-term implications of AI integration into societal, economic, and security-sensitive domains. The insights presented here are intended to inform both academic discourse and practical decision-making, emphasizing the need for responsible and human-centred AI development.

AI systems learn from historical data and infer patterns based on statistical correlations. This dependence makes them vulnerable to incomplete, biased, or outdated datasets. Unlike humans, AI systems cannot question the validity or relevance of their data; they treat patterns as truth regardless of context.

As a result, AI performs poorly in situations where historical data is unavailable, misleading, or unrepresentative. This limitation is particularly evident in rare events, edge cases, and rapidly changing environments.

3) Cognitive Incapabilities

One of the most fundamental cognitive limitations of Artificial Intelligence is the complete absence of consciousness and self-awareness. Consciousness enables humans to experience subjective states, reflect on their own thoughts, recognize intentions, and understand the consequences of actions. AI systems, by contrast, operate entirely without subjective experience or internal awareness.

Another major cognitive incapability of AI systems is the lack of true semantic understanding. Although AI systems can manipulate symbols, generate fluent language, and recognize complex patterns, they do not understand meaning in the human sense. Their operations are based on syntactic relationships and statistical associations rather than comprehension of concepts, intentions, or context.

4) Ethical, Moral, Societal, and Economic Limitations

As Artificial Intelligence systems increasingly influence decisions that affect human lives, ethical and moral concerns have become central to discussions on responsible AI development. While AI technologies are often promoted as objective and impartial, closer examination reveals that they lack the intrinsic ethical capacity required to navigate morally complex situations.

Ethical behaviour in AI systems is not an inherent property but an externally imposed construct, shaped by design choices, data inputs, and governance mechanisms.

V. RESULT, DISCUSSION AND CONCLUSION

The concluding chapter of this dissertation serves to integrate and reflect upon the insights developed throughout the study on Artificial Intelligence failures, limitations, incapabilities, and threats. While earlier chapters focused on detailed analysis, classification, and discussion, this chapter adopts a holistic perspective, drawing together the key arguments and evaluating their broader implications. The aim is not only to summarize findings but also to contextualize their significance in relation to ongoing technological, organizational, and societal developments.

REFERENCES

1. Westenberger, J., Schuler, K., & Schlegel, D. (2022). *Failure of AI projects: Understanding the critical factors*." **Procedia Computer Science**, 196, 69–76.
2. Banerjee, D. N., & Chanda, S. S. (2020). *AI failures: A review of underlying issues*. arXiv preprint arXiv:2008.04073.
3. Williams, R., & Yampolskiy, R. V. (2021). *Understanding and avoiding AI failures: A practical guide*. **Philosophies**, 6(3), 53.
4. Chollet, F. (2019). *On the measure of intelligence*. arXiv preprint arXiv:1911.01547.
5. Searle, J. R. (1980). *Minds, brains, and programs*. **Behavioral and Brain Sciences**, 3(3), 417–457.
6. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete problems in AI safety*. arXiv preprint arXiv:1606.06565.
7. Uesato, J., Kumar, S., 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**, 34(04), 11475–11482.
8. Buolamwini, J., & Gebru, T. (2018). *Gender shades: Intersectional accuracy disparities in commercial gender classification*. **Proceedings of the Conference on Fairness, Accountability, and Transparency**, 77–91.
9. Floridi, L., Cowls, J., Beltrametti, M., et al. (2018). *AI4People—An ethical framework for a good AI society*. **Minds and Machines**, 28, 689–707.
10. Yampolskiy, R. V. (2015). *Artificial intelligence safety engineering: Why machine ethics is a wrong approach*. **Philosophy & Technology**, 28(3), 389–396.
11. Azelya, I., & Filin, S. A. (2025). Introduction: The Role of AI in Transforming Management Research. Involvement International Journal of Business, 2 (1), 39–44.