International Conference on Science, Engineering & Management Trends

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

"AI Cancers": Systemic Limitations Threatening the Integrity of Artificial Intelligence

Sagar Gupta
Department of Computer Science

Abstract- As Artificial Intelligence (AI) systems permeate critical sectors like healthcare, finance, and governance, deep-rooted limitations have begun to surface—often referred to metaphorically as "AI cancers." These include systemic issues such as algorithmic bias, hallucination, goal misalignment, data poisoning, overfitting, and lack of explainability. Like cancer in a biological organism, these flaws can spread undetected, undermining trust, accuracy, and societal safety. This paper explores the nature, origin, and consequences of these "AI cancers," while outlining the emerging strategies to detect, contain, and remediate them.

Keywords- Artificial Intelligence (AI), AI Cancers, Algorithmic Bias, AI Hallucination, Goal Misalignment.

I. INTRODUCTION

Artificial Intelligence has advanced from rule-based automation to large-scale neural networks capable of language understanding, image generation, and autonomous decision-making. However, alongside this growth, critical flaws have emerged that, if left unchecked, can cause significant harm—both technically and ethically. Researchers and critics have begun referring to these flaws as "AI cancers" (Hendrycks et al., 2023), denoting persistent, often hidden issues that can metastasize through training data, model architecture, or deployment pipelines. Recognizing and mitigating these "cancers" is essential for building trustworthy, robust, and safe AI systems.

II. TAXONOMY OF AI CANCERS

Algorithmic Bias
Description: AI systems often inherit or amplify social, racial, or gender biases present in the training data. Example: COMPAS, an algorithm used in criminal sentencing, was found to disproportionately rate Black defendants as high-risk (Angwin et al., 2016). Impact: Reinforces systemic inequality; undermines fairness.

Hallucination
Description: Large language models (LLMs) like GPT and PaLM generate fluent yet factually incorrect or fabricated content—a phenomenon known as "hallucination."
Example: ChatGPT may confidently provide incorrect citations or non-existent laws.
Impact: Degrades trust in AI-generated content, especially in high-stakes domains like healthcare or legal work (Ji et al., 2023).

Goal Misalignment
Description: When an AI system's objective function diverges from human intent, it may pursue goals that are technically optimal but socially harmful.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Example: A content recommendation engine may maximize engagement at the cost of user well-being or polarization (Zhou et al., 2020).
Impact: Unintended consequences; ethical failures; safety risks.

Data Poisoning and Adversarial Attacks
Description: AI systems can be manipulated via poisoned inputs during training or adversarial perturbations during inference.
Example: Slight pixel changes to an image can cause a classifier to mistake a stop sign for a yield sign (Szegedy et al., 2014).
Impact: Security vulnerabilities; untrustworthy AI in critical systems.

Overfitting and Brittleness
Description: Models trained too closely on specific data distributions fail to generalize to unseen scenarios.
Example: Medical imaging models that perform well in a research hospital but fail in rural clinics due to different equipment or demographics (Oakden-Rayner, 2020).
Impact: Lack of robustness; poor real-world performance.

Lack of Explainability
Description: Deep learning models are often "black boxes," making it difficult to understand or justify their predictions.
Example: A neural network rejects a loan application, but the reason is opaque to users and regulators.
Impact: Regulatory non-compliance; reduced human trust; ethical concerns.

III. WHY THESE ARE CALLED 'CANCERS'

The term “AI cancer” is metaphorical but appropriate for several reasons:

- Silent Growth: Many of these issues remain undetected during model development and surface only post-deployment.
- Metastatic Nature: Biases or design flaws in foundation models get transferred across thousands of downstream applications.
- Systemic Harm: Like biological cancer, they threaten the integrity of the entire system—technically, ethically, and socially.
- Difficulty of Cure: Once deployed at scale, flawed AI systems are difficult to audit or correct.

IV. CASE STUDIES

Microsoft Tay (2016)
An AI chatbot released on Twitter, Tay began posting offensive and racist content within 24 hours due to adversarial user interactions—demonstrating both goal misalignment and lack of content safeguards.

Amazon’s AI Hiring Tool
An AI-based hiring tool trained on past data began downgrading resumes that included the word "women’s," revealing deep-seated gender bias inherited from historical data (Dastin, 2018).

V. TOWARD DETECTION AND CONTAINMENT

To address AI cancers, researchers are developing defensive strategies:

- Bias Auditing (e.g., Fairlearn, Aequitas)
- Explainability Tools (e.g., SHAP, LIME)

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

- Robust Training Protocols (e.g., adversarial training)
- Alignment Research (e.g., reinforcement learning with human feedback)
- Model Interpretability Layers (e.g., attention maps)

AI governance frameworks like the EU AI Act and the U.S. NIST AI Risk Management Framework are also pushing for accountability, fairness, and auditability.

VI. CONCLUSION

AI holds transformative potential—but like a biological system, it is not immune to internal pathologies. These "AI cancers"—bias, hallucination, misalignment, and more—require urgent attention from researchers, developers, ethicists, and policymakers. Only with proactive detection and systemic mitigation can we ensure AI evolves in a direction that benefits humanity rather than undermines it.

REFERENCES

1. Angwin, J., et al. (2016). Machine Bias. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
2. Dastin, J. (2018). "Amazon scraps secret AI recruiting tool that showed bias against women." Reuters. https://www.reuters.com/article/idUSKCN1MK08G
3. Hendrycks, D., et al. (2023). "Natural Safety Problems in AI." arXiv preprint arXiv:2301.07368. https://arxiv.org/abs/2301.07368
4. Ji, Z., et al. (2023). "Survey of Hallucination in Natural Language Generation." ACM Computing Surveys, 55(12), 1–38. https://doi.org/10.1145/3571730
5. Oakden-Rayner, L. (2020). "Exploring large-scale public medical image datasets." Academic Radiology, 27(1), 106–112. https://doi.org/10.1016/j.acra.2019.09.024
6. Szegedy, C., et al. (2014). "Intriguing properties of neural networks." arXiv preprint arXiv:1312.6199. https://arxiv.org/abs/1312.6199
7. Zhou, X., et al. (2020). "Reinforcement learning for recommender systems: A survey." ACM Computing Surveys, 52(1), 1–38.