

Large-Scale Literature Review of AI Security and Governance Research from a Managerial Perspective

Mariusz Rafał¹

Abstract

The rapid deployment of AI systems across organizations has increased exposure to security risks, including adversarial attacks on machine learning models. Despite growing research on adversarial AI and AI governance, the literature remains fragmented across technical, policy and managerial domains. This study aims to systematically examine how adversarial AI security is conceptualized and addressed within the broader AI governance and management perspective. The objective is to identify dominant research streams, knowledge structures, and gaps in the integration of technical security insights into AI governance frameworks. This study conducts a large-scale structured literature review combined with network-based bibliometric analysis. A corpus of academic publications related to adversarial AI, machine learning security, and AI governance is collected using targeted keyword searches across major scholarly databases. The analysis maps relationships between papers, authors and key insights to dependency networks. Additionally, abstracts and referenced literature are examined using large language models (LLM) to identify conceptual orientations, research goals, and disciplinary influences within the field. The results reveal significant fragmentation between technical adversarial machine learning research and managerial AI governance practice. Technical studies primarily focus on model robustness and attack–defense mechanisms, while management-oriented research emphasizes regulatory compliance, risk management and organizational oversight. The citation network analysis shows limited cross-referencing between these communities, suggesting weak knowledge transfer across disciplines. The findings also identify a small number of bridging publications that connect technical security insights with AI governance and organizational perspectives. This study contributes to management and information systems research by providing a large-scale, network-based mapping of adversarial AI literature. By integrating bibliometric analysis with the use of LLM, the research highlights structural knowledge gaps that hinder effective organizational governance of AI security risks. The study offers a foundation for future interdisciplinary research and supports the development of more integrated governance frameworks addressing adversarial threats.

Keywords: adversarial AI, machine learning security, AI governance, bibliometric analysis, literature review, LLM, AI security management

Doi: <https://doi.org/10.66414/mlabs.9680>

¹ Mr., Warsaw School of Economics, Poland. ORCID: 0000-0002-4868-3571