AI Safety

Rifky Firmansyah
http://orcid.org/0009-0000-9177-966X
Universitas Esa Unggul, Indonesia

ABSTRACT

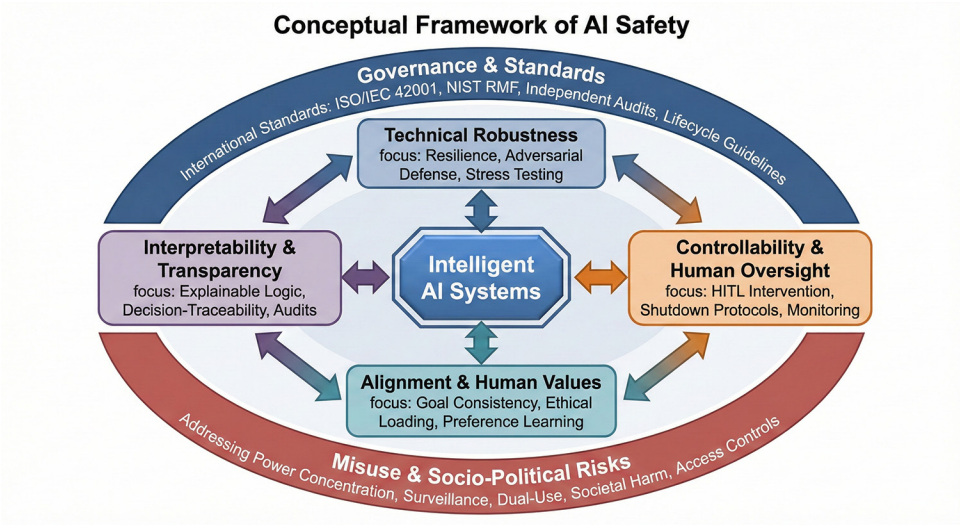
This study proposes a comprehensive framework for AI safety as the foundation for developing secure, controllable, and human-aligned artificial intelligence systems. Through a systematic review of literature from 2021 to 2025, it identifies three core pillars: technical robustness against disruptions, ethical value alignment, and socio-regulatory governance. The analysis covers real-world cases of medical diagnostic vulnerabilities, autonomous vehicle failures, and manipulation risks in large language models (LLMs). Mitigation approaches such as safe reinforcement learning, formal verification, and human-in-the-loop mechanisms are explored. The study also highlights the importance of global standards, including ISO/IEC 23894 and the NIST AI Risk Management Framework, in ensuring transparency and accountability. Findings suggest that cross-disciplinary collaboration among scientists, regulators, and civil society is crucial to building a trustworthy and ethically responsible AI ecosystem.

1. INTRODUCTION: AI SAFETY IN THE ERA OF COMPLEX AUTONOMOUS SYSTEMS

Artificial Intelligence (AI) systems are increasingly embedded in critical infrastructures such as healthcare, autonomous transportation, defense, and digital communication. As these systems scale in complexity and autonomy, they introduce systemic risks that extend beyond conventional software failures, including adversarial vulnerabilities, value misalignment, reduced transparency, and governance gaps. AI Safety therefore emerges as a socio-technical discipline concerned with ensuring that intelligent systems operate reliably, ethically, and accountably in real-world contexts. Addressing these challenges requires the integration of technical robustness, alignment with human values, interpretability, human oversight, and adaptive governance mechanisms across the AI system lifecycle.

DOI: 10.4018/406036

Figure 1. Conceptual Framework of AI Safety



This figure presents AI Safety as a socio-technical framework centered on intelligent AI systems and organized around six interdependent pillars: technical robustness, alignment with human values, interpretability and transparency, controllability and human oversight, misuse and socio-political risks, and governance and standards. The framework highlights the bidirectional relationships between technical safeguards and institutional mechanisms across the AI system lifecycle.

1.1 Technical Robustness and Vulnerability of Modern AI Systems

Modern AI systems, particularly those based on deep learning architectures, have demonstrated remarkable performance across domains such as image recognition, language understanding, and autonomous control. However, this performance is often accompanied by an alarming vulnerability to small, targeted perturbations that can severely disrupt system behavior. Adversarial attacks exemplify this fragility; small, often imperceptible modifications to input data can cause high-confidence misclassifications, even in clinically validated systems. In medical imaging, adversarial noise has been shown to trigger dangerous diagnostic errors in radiology models, highlighting how seemingly minor input perturbations can undermine the reliability of safety-critical AI systems. These findings are not confined to healthcare; similar vulnerabilities exist across a range of AI applications, from autonomous vehicle perception to real-time surveillance and multimodal models.

These observations point to a broader concern regarding robustness in modern AI systems: increases in architectural complexity do not automatically translate into improved resilience against adversarial manipulation. Empirical studies in safety-critical domains indicate that highly parameterized models may exhibit brittle behavior when exposed to distributional shifts or adversarial inputs, particularly when robustness is not an explicit optimization objective. In response, emerging strategies such as model pruning and the integration of attention mechanisms have demonstrated potential for improving robustness while maintaining interpretability and performance (Chen et al., 2021). These technical solutions,