Chapter 2
Designing AI to Prevent Harm (Safety)

Silvio Andrae
https://orcid.org/0000-0002-8586-7812
Independent Researcher, Germany

ABSTRACT

The growing integration of artificial intelligence into safety-critical socio-technical systems increases the risk of physical, psychological, and societal harm that cannot be addressed solely by technical reliability. This chapter positions AI safety as a central and independent objective within AI design and governance, emphasizing the prevention, limitation, and containment of harm throughout the entire lifecycle of AI systems. The analysis adopts a harm-based framework and systematically reviews ethical, regulatory, and technical approaches, including the EU AI Act, product safety and liability regimes, safety testing protocols, risks associated with human–AI interaction, and fail-safe mechanisms. The results indicate that adequate AI safety cannot be achieved solely through ex ante testing or formal compliance. Instead, it requires integrated safety architectures that incorporate continuous monitoring, organizational accountability, interaction-aware design, and robust intervention and containment mechanisms in the presence of irreducible uncertainty.

INTRODUCTION: DEFINITION OF AI SAFETY

Artificial intelligence is increasingly integrated into socio-technical systems, whose operations may directly or indirectly impact physical integrity, psychological well-being, and broader social structures. Within this context, AI safety has become a central design objective, emphasizing the prevention, limitation, and mitigation of harm resulting from the operation, malfunction, or misuse of AI systems. Although

DOI: 10.4018/979-8-3373-6935-8.ch002

AI safety has gained prominence in policy discourse, technical standards, and ethical frameworks, its conceptual boundaries remain analytically contested. They are often conflated with related notions such as IT security, robustness, trustworthiness, and explainability. Therefore, precise conceptual clarification is essential for the systematic analysis of safety-oriented AI design.

Fundamentally, AI safety addresses an AI system's capacity to operate within clearly defined boundaries, ensuring that both anticipated and unforeseen adverse developments do not result in unacceptable physical, psychological, or societal harm (Amodei et al., 2016; Leike et al., 2017). This approach emphasizes harm dynamics and risk consequences over internal model characteristics. Accordingly, safety is not primarily determined by optimal performance or complete epistemic transparency, but by the system's ability to detect deviations, manage uncertainty, and mitigate adverse effects in real-world contexts. The central concern is not whether an AI system functions correctly under ideal conditions, but whether failures, misuse, or contextual mismatches can be identified promptly and their consequences effectively contained.

A key aspect of this approach is recognizing AI safety as a systemic property. Harm seldom arises from isolated technical errors; instead, it typically results from interactions among algorithmic components, human actors, organizational routines, and institutional frameworks (Perrow, 1999; Baxter & Sommerville, 2011). Therefore, safety should not be treated as an add-on after deployment, but must be integrated throughout the entire lifecycle of AI systems, including problem formulation, data and model selection, deployment, monitoring, and decommissioning (ISO/IEC, 2020; NIST, 2023).

Current debates on AI safety reveal divergent conceptual priorities and, at times, competing problem framings. Engineering-oriented perspectives typically define safety in terms of robustness, reliability, and performance guarantees under specified conditions, focusing on technical failure modes and model behavior. In contrast, governance- and ethics-oriented approaches emphasize socio-technical risks arising from deployment contexts, institutional incentives, and human–AI interactions, framing safety as an organizational and regulatory challenge. A third perspective, prevalent in frontier AI and existential risk discussions, addresses long-term, low-probability but high-impact scenarios associated with highly autonomous systems and potential loss of human control. This chapter adopts a harm-based, lifecycle-oriented conception of AI safety that bridges these viewpoints while analytically distinguishing them. While recognizing the importance of technical robustness and long-term foresight, it rejects reducing safety to either model-centric optimization or speculative future risks. Instead, AI safety is understood as a socio-technical practice focused on preventing, limiting, and containing tangible forms of harm throughout development, deployment, and use, especially under persistent uncertainty.