Chapter 1
Advancing Women in Artificial Intelligence (AI) Safety and Artificial Intelligence (AI) Security, Especially in Healthcare

Amalisha Sabie Aridi
https://orcid.org/0000-0002-7869-5530
Washington University of Science and Technology, USA

ABSTRACT

Artificial intelligence (AI) safety and security constitute interdependent disciplines essential for mitigating both unintended and deliberate harms arising from autonomous systems. AI safety aims to ensure that intelligent agents operate predictably and in alignment with human values, while AI security safeguards systems against adversarial manipulation, data poisoning, and model exploitation. Together, these domains underpin the ethical and operational integrity of AI technologies across sectors, particularly in high-stakes environments such as healthcare and governance. Nevertheless, the persistent underrepresentation of women, who comprise less than one-third of AI professionals globally, poses a critical challenge to the field's inclusivity, legitimacy, and epistemic diversity. This commentary argues that women's participation in AI safety and security is not merely a matter of representation, but a prerequisite for developing systems that are equitable, accountable, and aligned with societal values. Women's perspectives are indispensable in identifying algorithmic bias, confronting gendered harms such as deepfakes, and designing governance frameworks that advance fairness and transparency. Simultaneously, AI technologies themselves can serve as tools for empowerment, mentorship, and

DOI: 10.4018/979-8-3373-9918-8.ch001

confidence-building among women in professional roles. Ensuring women's engagement is therefore both an ethical and strategic policy imperative for achieving robust, socially responsive AI safety and security practices.

INTRODUCTION

In the rapidly evolving digital age, the convergence of artificial intelligence (AI), law, and ethics has produced a complex terrain where technological innovation frequently outpaces governance and regulatory oversight. Within healthcare organizations, AI safety has emerged as an essential interdisciplinary field devoted to ensuring that intelligent systems operate reliably, predictably, and in alignment with both professional ethics and public health values (Gyevnar & Kasirzadeh, 2025). As AI systems increasingly influence clinical diagnostics, patient triage, and administrative workflows, the legal and ethical implications of their deployment have become profound. AI safety within healthcare is not limited to technical precision; it entails preventing unintended behaviors such as diagnostic errors, triage misclassifications, or inequitable patient outcomes, while maintaining compliance with established legal frameworks such as the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR).

Nevertheless, AI safety extends beyond algorithmic accuracy or software reliability, it is fundamentally a moral and legal imperative. The goal is to ensure that innovation advances patient welfare while preserving autonomy, consent, and data integrity. For instance, an AI-assisted diagnostic platform must not only perform with clinical accuracy but also adhere to data protection statutes and align with ethical norms governing informed consent and patient privacy (Mahanta & Maringanti, 2023). As such, AI safety serves as a critical bridge between technological innovation and the principles of justice, transparency, and accountability that underpin modern healthcare ethics and law.

Closely related yet distinct, AI security occupies a complementary role within the healthcare ecosystem. It focuses on protecting AI-driven medical systems from deliberate threats, including adversarial manipulation, data poisoning, model theft, and the exploitation of hospital IT infrastructure (Rahaman et al., 2024). The stakes in healthcare are uniquely high: adversaries could manipulate diagnostic algorithms to produce false results, compromise sensitive medical data through model inversion, or disrupt the functioning of AI-enabled medical devices. These threats raise pressing legal and ethical concerns around liability, patient harm, and the adequacy of institutional cybersecurity measures.

Moreover, the dual-use nature of AI technologies, capable of simultaneously enhancing clinical care and enabling cyberattacks or misinformation, demands

Cyber Bullying in the Digital Age: Challenges, Impact, and Strategies for Prevention

Ria Ghosh, Meetu Malhotraand Naresh Kumar (2025). *Combating Cyberbullying With Generative AI (pp. 151-180).*

www.irma-international.org/chapter/cyber-bullying-in-the-digital-age/369058