Ethical AI in the Management of Defence Projects:
Navigating the Challenges for Responsible Innovation

Luis Javier SerranoTamayo
http://orcid.org/0000-0001-6449-4781
Cardiff Metropolitan University, UK

Babatunde Abiodun Salami
http://orcid.org/0000-0002-9175-2923
Cardiff Metropolitan University, UK

ABSTRACT

The availability of powerful AI tools is impacting multiple sectors, including defence. The integration of AI in defence has ethical challenges and concerns, and the vulnerabilities introduced by reliance on AI systems are among the pressing issues that policymakers, military leaders, and ethicists must grapple with. AI in military contexts involves balancing the potential benefits against the risks of escalation, civilian harm, and unintended consequences. This chapter explores two fundamental concepts of AI ethics applied to the defence sector: accountability and trust. Accountability relates to humans' final responsibility in using AI for military applications. Trust is an ongoing journey to a desired state in which humans and machines optimise their relations and balance the desired outcomes with the unintended consequences. A qualitative content analysis of policy documents, academic papers, books, reports, and a series of official and unofficial documents in the burgeoning field of AI ethics in defence contributes to the necessary discussion of AI ethical principles.

INTRODUCTION

The integration of artificial intelligence (AI) into military operations is a paradigm shift in modern warfare, transitioning from theoretical discussions to tangible, battlefield-deployed systems that are reshaping strategic and tactical calculations (Center for Strategic and International Studies, 2025; Hadlington et al., 2023). The primary challenge for nations is no longer whether AI should be used, but rather how it can be responsibly managed within the high-stakes, ethically complex environment of national defence, as they increasingly employ AI to enhance capabilities in intelligence, surveillance,

DOI: 10.4018/406029

command and control, and autonomous systems (Borchert, 2024; Payne, 2024). The development of corresponding legal and ethical frameworks is consistently outpaced by the rapid pace of technological advancement, resulting in a critical gap between capability and governance that necessitates urgent attention (Blanchard & Taddeo, 2024; Canca, 2023).

This chapter addresses the complex landscape of ethical AI in defence project management by emphasising two fundamental pillars that are indispensable for bridging this gap: accountability and trust. Although there is a growing international consensus on the importance of high-level ethical principles (Defense Innovation Board, 2019; UK Ministry of Defence, 2022), their operationalisation is the true test (Taddeo et al., 2024). Abstract guidelines must be transformed into concrete, verifiable mechanisms that are integrated into the entire lifecycle of a defence project, from design and procurement to deployment and post-mission review (Devitt et al., 2020). To establish clear lines of responsibility and cultivate a justifiable basis for trust in systems that may operate at velocities and scales beyond direct human control, it is necessary to transcend broad principles (Osoba, 2024; Glikson & Woolley, 2020).

Confronting this challenge necessitates confronting the distinctive challenges posed by the defence context. The regulatory endeavour is complicated and the distinction between permissible and impermissible use is blurred by the dual-use nature of AI, in which civilian innovations are rapidly repurposed for military applications (Uzer, 2024; McFaul et al., 2025). Additionally, the conflict between the necessity to uphold international humanitarian law and the operational effectiveness of the mission requires a comprehensive approach that recognises ethical considerations as essential components of mission success and legitimacy, rather than as constraints (Fanni & Giancotti, 2023; Taddeo, 2024). This study examines the practical pathways for establishing accountability and trust in defence AI, focusing on the frameworks, processes, and cultural shifts necessary to ensure that these potent technologies are developed and deployed in an ethically sound and effective manner.

BACKGROUND

The risks and concerns associated with the growing development of AI have indeed brought ethics to the forefront as a rapidly evolving field to address the implications of the use of AI technologies. AI, defined generally as the theory and development of computer systems able to perform tasks normally requiring human intelligence (Jobin et al., 2019), is recognised as a pervasive, general-purpose technology comparable in influence to electricity or the internal combustion engine (Payne, 2024). Since around 2000, there has been a significant and persistent growth in AI research, development, and application; this acceleration of AI into commercial products and services, particularly since the early 2010s (Szabadföldi, 2021), has led to a significant increase in ethical guidelines released globally, with 88% of identified policy documents being issued after 2016 (Canca, 2023).

Although AI systems should respect human rights and values, be accountable and transparent, and benefit all humanity, multiple concerns have arisen. A global consensus has emerged around five core ethical principles: transparency and explainability, justice and fairness, non-maleficence, responsibility and accountability, and privacy (Defense Innovation Board, 2019). These principles address fundamental societal fears, such as the possibility of AI increasing inequality if justice and fairness considerations are neglected. For instance, transparency is frequently cited as a "proethical condition" necessary to enable other ethical practices, such as fostering trust (Jobin et al., 2019). Privacy and data protection concerns are paramount, often presented in relation to data security