WHEN PEACE BECOMES AN EXTERNALITY: STRUCTURAL MISALIGNMENT BETWEEN AI SAFETY, AI ETHICS, AND AI FOR PEACE

Anonymous authors
Paper under double-blind review

ABSTRACT

Artificial intelligence research is increasingly embedded in contexts where its outputs are repurposed for military, security, and large-scale surveillance applications. In response, multiple research paradigms have emerged to address potential harms, most prominently AI safety, AI ethics, and a growing body of work framed as AI for peace. While each paradigm engages with questions of responsibility and risk, their underlying assumptions about where harm arises and how it should be addressed differ substantially. In this paper, we argue that peace is systematically treated as an externality across dominant AI research frameworks. Rather than being integrated as an upstream design and governance constraint, peace-oriented concerns are frequently deferred to downstream applications or policy interventions. Through a comparative analysis of AI safety, AI ethics, and AI for peace, we show how this structural misalignment limits the ability of peace-oriented initiatives to influence research trajectories that feed into militarized and security-driven deployments. We conclude by outlining how repositioning peace as an infrastructural concern within AI research ecosystems can strengthen harm prevention and support more durable peace-oriented outcomes.

1 INTRODUCTION

Artificial intelligence research has become deeply intertwined with domains characterized by heightened security concerns, strategic competition, and institutional power asymmetries. While some AI systems are explicitly developed for defense or security purposes, many originate in civilian research environments before being repurposed within military, intelligence, or large-scale surveillance infrastructures. This pattern reflects long-standing relationships between computation, state power, and security-oriented research agendas Crawford (2021); Winner (1980).

The contemporary scale and generality of AI systems intensify these dynamics. General-purpose models, datasets, and benchmarks circulate across institutional boundaries, enabling downstream integration into security and surveillance systems often without the knowledge or consent of upstream researchers Bowker & Star (2009); Whittaker et al. (2021). These developments have prompted sustained debate within the AI research community about responsibility, harm, and governance.

In response, several influential research paradigms have taken shape. AI safety has focused on preventing catastrophic or system-level failures arising from advanced systems Rahwan et al. (2019). AI ethics has foregrounded normative principles such as fairness, accountability, transparency, and respect for human rights Jobin et al. (2019); Floridi et al. (2018). More recently, AI for peace initiatives have sought to orient AI research toward violence prevention, humanitarian assistance, and conflict mitigation.

Despite this growing attention, peace-oriented concerns often remain marginal within mainstream AI research. Discussions of peace tend to appear downstream, framed as application choices or policy considerations rather than as constraints on research design, dissemination, and governance. This raises a central puzzle for the AI for peace community: why does peace repeatedly appear as an add-on, even in research ecosystems otherwise saturated with ethical reflection?

We argue that this marginalization is not primarily the result of indifference or neglect. Instead, peace becomes externalized due to a structural misalignment between dominant AI research paradigms. AI safety, AI ethics, and AI for peace operate at different levels of abstraction, emphasize different forms of harm, and assume different sites of intervention. As a result, peace is positioned outside the core research pipeline, limiting its capacity to shape upstream decisions whose consequences unfold in militarized and security-driven contexts Birhane et al. (2023).

2 Three Paradigms and Their Assumptions

To understand how peace becomes externalized, it is useful to compare how different AI research paradigms conceptualize harm, responsibility, and intervention.

2.1 AI Safety

AI safety research is primarily concerned with preventing catastrophic or system-level failures arising from advanced AI systems. Typical risks include misalignment between system objectives and human intent, loss of control, and emergent behaviors that exceed design expectations Rahwan et al. (2019). Within this paradigm, harm is framed as a property of system behavior.

Responsibility is therefore located at the level of technical design and control. Interventions emphasize robustness, interpretability, alignment techniques, and formal evaluation regimes. While these efforts address important failure modes, they often abstract away from the institutional, geopolitical, and security contexts in which AI systems are deployed. As a result, questions concerning militarization or surveillance are frequently treated as exogenous to the safety problem itself Floridi et al. (2018).

2.2 AI Ethics

AI ethics scholarship has played a central role in articulating normative concerns related to fairness, accountability, transparency, and human rights Jobin et al. (2019). This paradigm has expanded the scope of AI governance by highlighting social and structural harms that extend beyond technical performance.

However, AI ethics often locates responsibility at the level of individual researchers, organizations, or voluntary principles. Ethical review processes and guidelines raise awareness but frequently lack enforcement power once systems enter complex institutional environments Metcalf et al. (2019). In security-sensitive contexts, ethical commitments may be subordinated to organizational or strategic priorities, limiting their practical influence on peace-related outcomes.

2.3 AI for Peace

AI for peace initiatives explicitly aim to prevent violence, support humanitarian efforts, or contribute to conflict mitigation. This work often emphasizes positive applications, such as early warning systems, humanitarian logistics, or post-crisis recovery.

Yet AI for peace is frequently positioned as an application domain rather than a structural constraint on research production. Peace-oriented projects may coexist alongside research pipelines that also enable military or surveillance deployments. As a result, AI for peace initiatives often lack leverage over upstream research decisions, constraining their ability to reshape the conditions under which harm arises Crawford (2021).

3 Peace as a Structural Externality

Taken together, these paradigms reveal a recurring pattern: peace is treated as an externality. AI safety prioritizes technical failure modes, AI ethics emphasizes principles and awareness, and AI for peace focuses on downstream applications. None of these paradigms, in isolation, fully address how AI research infrastructures channel capabilities into security and military contexts.

Table 1: Comparison of dominant AI research paradigms and their treatment of harm, responsibility, and intervention

PARADIGM	PRIMARY HARM FRAMING	SITE OF RESPONSIBILITY	TYPICAL INTERVENTION MODE
AI Safety	System-level failure, misalignment, loss of control	Technical design and model behavior	Alignment methods, robustness, evaluation benchmarks
AI Ethics	Normative violations, social and rights-based harms	Individual researchers and organizations	Ethical principles, review processes, impact assessments
AI for Peace	Violence, instability, humanitarian harm	Downstream applications and policy contexts	Peace-oriented applications, mitigation and response

When peace is framed as an outcome rather than a design constraint, responsibility for preventing harm is deferred. Researchers may recognize ethical risks while lacking institutional mechanisms to intervene. Peace-oriented initiatives may demonstrate localized benefits without altering the broader research ecosystems that enable conflict-related uses Bowker & Star (2009); Suchman (2007).

Table 1 illustrates how peace is systematically displaced across paradigms. While each framework addresses genuine forms of harm, they locate responsibility at different points in the research lifecycle, leaving upstream research infrastructures largely unexamined.

4 Governance, Neutrality, and Structural Limits

The persistence of peace as an externality is further reinforced by dominant norms of neutrality within scientific research cultures. Neutrality is often framed as a professional virtue, allowing researchers to develop general-purpose tools while remaining agnostic about downstream use. However, scholarship in science and technology studies has long shown that claims of neutrality frequently align with existing power structures Winner (1980); Latour (2005).

In AI research, neutrality often enables research outputs to remain compatible with security-driven deployment environments while diffusing responsibility for harm. Ethical and peace-oriented commitments are articulated without corresponding authority to contest use Metcalf et al. (2019); Whittaker et al. (2021). This dynamic helps explain why peace is rarely framed as a failure condition within research evaluation, whereas safety and ethics violations are.

Reframing peace as an infrastructural concern requires challenging neutrality not as an epistemic ideal, but as an institutional positioning. Decisions about openness, scalability, and dissemination are normative choices that shape whose interests are ultimately served Benjamin (2019).

5 Peace as an Upstream Design and Governance Constraint

Treating peace as an infrastructural concern requires a shift in how responsibility is conceptualized across the AI research lifecycle. Rather than viewing peace as an aspirational outcome or an application-specific objective, this perspective positions peace as a constraint that operates upstream, shaping decisions about what kinds of research are pursued, how results are disseminated, and under what conditions reuse is considered legitimate.

Across contemporary AI research ecosystems, governance mechanisms tend to focus on compliance, risk management, or post-hoc accountability. Safety evaluations assess whether systems behave as intended, ethical reviews examine whether principles are violated, and peace-oriented initiatives often assess whether deployments mitigate or exacerbate harm. While these mechanisms address important dimensions of responsibility, they largely presuppose that core research trajectories are already fixed. As a result, they intervene after epistemic and material commitments have been made, when the capacity to redirect or refuse harmful pathways is significantly diminished Bowker & Star (2009); Suchman (2007).

Reframing peace as an upstream constraint challenges this sequencing. It asks whether research agendas, funding structures, and dissemination norms should be evaluated not only in terms of technical merit or ethical compliance, but also in terms of their compatibility with non-violent and non-militarized futures. From this perspective, peace-oriented evaluation does not replace safety

or ethics; rather, it interrogates the assumptions that allow research outputs to flow seamlessly into security-driven infrastructures without triggering governance friction Whittaker et al. (2021).

Importantly, this framing avoids reducing peace to individual moral intent. Structural analyses of AI governance consistently show that responsibility is distributed across infrastructures, institutions, and incentive regimes Rahwan et al. (2019). Individual researchers may act in good faith while participating in systems that systematically externalize harm. Peace as an infrastructural constraint therefore emphasizes collective responsibility and institutional design over individual virtue.

This approach also clarifies why neutrality functions as a stabilizing force within existing research ecosystems. Claims of neutrality enable research outputs to remain maximally reusable and scalable, qualities that are highly valued in academic and industrial settings. However, as science and technology studies have long demonstrated, infrastructure is never neutral with respect to power or consequence Winner (1980); Latour (2005). When neutrality is left uninterrogated, it effectively privileges downstream actors with the greatest capacity to appropriate research outputs, including military and security institutions.

Positioning peace upstream makes this asymmetry visible. It highlights how decisions about openness, generality, and abstraction are not merely technical choices, but governance decisions that shape whose interests are ultimately served Crawford (2021); Benjamin (2019). Peace-oriented governance, on this view, does not require prohibiting research or prescribing specific applications. Instead, it introduces friction into research pipelines that otherwise treat militarized reuse as an unremarkable extension of scientific progress.

Finally, understanding peace as a design and governance constraint aligns AI for peace with broader efforts to move from reactive to anticipatory governance. Rather than responding to harm once systems are deployed, anticipatory approaches seek to shape the conditions under which harm becomes more or less likely Floridi et al. (2018). In this sense, AI for peace contributes not only a set of applications, but a critical perspective on how research infrastructures distribute agency, risk, and accountability over time.

By embedding peace upstream, AI research communities gain the capacity to ask different questions: not only whether systems are safe or ethical, but whether the research ecosystems that produce them systematically foreclose non-violent alternatives. This shift strengthens the analytical and practical relevance of AI for peace, positioning it as a necessary complement to existing safety and ethics paradigms rather than a downstream add-on.

6 Conclusion

AI for peace initiatives operate within a research landscape already shaped by safety and ethics frameworks. While these paradigms address important dimensions of harm, their structural misalignment contributes to the persistent marginalization of peace-oriented concerns. Peace becomes an externality not because it is unvalued, but because it is systematically positioned downstream of research decisions that determine how AI capabilities are produced, circulated, and ultimately mobilized. This paper has argued that repositioning peace as an infrastructural concern alters the locus of responsibility within AI research ecosystems. Rather than treating peace as a contingent outcome of deployment choices, an infrastructural perspective foregrounds how upstream design, governance, and dissemination decisions preconfigure the range of downstream possibilities. In doing so, it exposes the limits of relying solely on technical safety or normative ethics to address harms that are structurally enabled.

Crucially, this reframing does not demand the abandonment of existing paradigms. Instead, it reveals how AI for peace can function as a necessary complement—one that interrogates neutrality, challenges the deferral of responsibility, and reorients attention toward the institutional conditions under which harm prevention becomes possible. By embedding peace upstream, AI research communities gain the capacity not only to mitigate harm after the fact, but to contest the trajectories that make violent and security-driven deployments appear inevitable. If peace continues to be treated as an externality, AI research will remain structurally aligned with the very systems it seeks to regulate. Repositioning peace as infrastructure is therefore not an aspirational add-on, but a prerequisite for any serious claim that AI research can contribute to durable, non-violent futures.

GENERATIVE AI DISCLOSURE

The authors used large language models in a limited assistive capacity for language refinement, clarity, and structural editing of the manuscript. All conceptual contributions, arguments, theoretical framing, and citations were developed, selected, and verified by the authors. No generative AI system was used to generate original research content, arguments, or conclusions. The authors retain full responsibility for the content of this work.

ETHICS STATEMENT

This paper presents a conceptual and critical analysis of AI research paradigms and their relationship to peace, militarization, and governance. It does not involve human subjects, personal data, or experimental deployment of AI systems. As such, no institutional ethics approval was required.

The work is motivated by concerns about the downstream use of general-purpose AI research in military, security, and surveillance contexts. The analysis is intended to support harm prevention, responsible research governance, and peace-oriented outcomes. The paper does not promote, optimize, or enable military or weapons-related AI systems, and it avoids operational or tactical detail that could be repurposed for such uses.

All arguments are presented at a structural and analytical level, consistent with the aims of the AI for Peace workshop and with broader norms of responsible AI research.

FUNDING DISCLOSURE

The authors declare that there was no targeted funding from military, defense, or security agencies for the research presented in this paper.

REFERENCES

Ruha Benjamin. *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity Press, Cambridge, UK, 2019.

Abeba Birhane, Pratyusha Ria Kalluri, William Agnew, Myra Cheng, Katherine Owens, and Luca Soldaini. The surveillance ai pipeline. *arXiv preprint arXiv:2309.15084*, 2023.

Geoffrey C. Bowker and Susan Leigh Star. Infrastructure and social complexity. *Science, Technology, & Human Values*, 34(1):97–118, 2009.

Kate Crawford. *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press, New Haven, CT, 2021.

Luciano Floridi, Josh Cowls, Monica Beltrametti, et al. Ai4people—an ethical framework for a good ai society. *Minds and Machines*, 28(4):689–707, 2018.

Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399, 2019.

Bruno Latour. *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford University Press, Oxford, UK, 2005.

Jacob Metcalf, Emanuel Moss, and Danah Boyd. Owning ethics: Corporate logics, silicon valley, and the institutionalization of ethics. *Social Research*, 86(2):449–476, 2019.

Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, et al. Machine behaviour. *Nature*, 568(7753):477–486, 2019.

Lucy A. Suchman. *Human–Machine Reconfigurations: Plans and Situated Actions*. Cambridge University Press, Cambridge, UK, 2007.

Meredith Whittaker, Kate Crawford, Roel Dobbe, Genevieve Fried, Elizabeth Kaziunas, Varoon Mathur, and Sarah Myers West. Foundations of ai governance: Mapping power and accountability. *AI Now Institute Report*, 2021.

Langdon Winner. Do artifacts have politics? *Daedalus*, 109(1):121–136, 1980.