International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Governing the Ungoverned: AI Safety Research as a Bridge Between Humanitarian Principles and Algorithmic Power

Christine Wawira Nyaga
Afrinet Global

Abstract - The rapid spread of artificial intelligence across military and humanitarian settings has brought into view a governance gap that is structural rather than incidental. AI systems are being built and deployed far faster than the legal, ethical and institutional frameworks needed to oversee them. This paper argues that AI safety research, understood as a policy and governance discipline rather than a technical specialism, offers a principled and practical way through this impasse. Drawing on the three thematic tensions identified by Humanitarian Alternatives, namely operational contexts, professional reconfigurations and interdisciplinary dilemmas, the paper shows that the frameworks and institutional proposals developed within the AI safety field address directly the challenges facing humanitarian actors. AI safety governance is not an adjacent concern. It is the connective tissue the humanitarian sector is missing, sitting between the principles it holds and the algorithmic systems it is increasingly being asked to rely on.

Keywords: AI safety, AI governance, humanitarian principles, algorithmic accountability, regulatory frameworks, international humanitarian law, dual-use AI

1. INTRODUCTION: A GOVERNANCE GAP IN REAL TIME

When the Israeli military deployed a targeting system called Lavender in Gaza, assigning algorithmic risk scores to thousands of individuals, and when Ukrainian drones began navigating autonomously through electronic jamming to reach their targets, the world crossed a threshold. These were not controlled experiments reviewed by ethics committees. They were live deployments of AI systems in lethal, complex environments, the same environments where humanitarian workers operate, where civilians seek protection, and where international humanitarian law is supposed to govern the conduct of all parties.

The humanitarian sector has responded with alarm, and the alarm is warranted. But alarm does not produce governance. What the sector needs is a coherent framework: one that can define, in operationally meaningful terms, what responsible AI deployment looks like; one that can hold states, militaries and technology developers to account; and one that gives humanitarian organisations the tools to engage critically with AI systems they may be pressured, or simply tempted, to adopt.

That framework exists, at least in substantial part. AI safety research, as a policy and governance discipline, has been building it for over a decade. This paper argues that the humanitarian sector's engagement with AI needs to be grounded in the vocabulary, methods and institutional proposals that safety governance has produced. Equally, safety researchers need the humanitarian field's empirical grounding, normative authority and political legitimacy to make their proposals matter beyond academic circles.

The argument proceeds in four steps. Section 2 maps the structural governance deficit connecting military AI proliferation to the pressures facing humanitarian actors. Section 3 shows how AI safety governance frameworks respond to each of the three thematic tensions in this special issue. Section 4 proposes a concrete institutional architecture for embedding safety principles into humanitarian AI practice. Section 5 considers the political economy of reform and the conditions under which progress becomes possible.

2. THE STRUCTURAL GOVERNANCE DEFICIT

2.1 Speed, Opacity and the Erosion of Accountability

The core challenge AI poses to the humanitarian sector is institutional rather than technical. AI systems are being designed, deployed and scaled at a pace that consistently outruns the regulatory, legal and ethical frameworks intended to govern them. This is not a

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

failure of any particular government or organisation. It is a structural feature of the current AI development landscape, one in which first-mover advantages are substantial, transparency obligations remain thin, and liability regimes are largely underdeveloped.

In the military domain, the consequences are most severe. The principle of *distinction*, the obligation under international humanitarian law to differentiate between combatants and civilians, rests on the assumption that a human being capable of moral and legal accountability is making targeting decisions. When that decision is delegated, even partially, to an algorithmic system, the accountability chain fractures. Who bears responsibility when a system operating within an accepted error margin produces civilian casualties? The developer? The commanding officer who authorised its deployment? The procuring state? Existing legal frameworks offer no clean answer, and that ambiguity is being exploited in practice.

In the humanitarian domain, the governance deficit is less dramatic but no less consequential. Humanitarian organisations face growing pressure to adopt AI tools that promise efficiency gains in logistics, needs assessment and fundraising. Most lack the institutional capacity to audit those tools, understand their failure modes or evaluate their longer-term effects on the populations they serve. Smaller organisations, in particular, often cannot distinguish between AI systems that are genuinely useful and safe and those that carry embedded biases or create dependencies on external providers they have no power to hold to account.

2.2 The Military-Humanitarian Imbalance

This edition of Humanitarian Alternatives draws attention to a political asymmetry that deserves to be taken further than the funding question. Military investment in AI substantially exceeds humanitarian investment, and the technologies developed for warfare are increasingly shaping the environments in which aid is delivered. But the deeper issue is about control over infrastructure. Who controls the systems through which populations are classified, needs are assessed and decisions are made?

Satellite imagery interpreted by AI systems funded by defence ministries may be the same imagery used to map damage after a natural disaster. Biometric systems developed for border enforcement may be repurposed for refugee registration. The data infrastructures built for military intelligence and those available to humanitarian actors are not as separate as convention suggests, and the terms on which humanitarian organisations access shared infrastructure are set by actors whose primary accountability lies with their own governments, not with affected populations.

AI safety governance research has developed tools to engage seriously with this asymmetry. The concepts of value alignment, corrigibility, and principal hierarchies help articulate what genuine accountability to affected populations would actually require of an AI system, as distinct from accountability to the organisations deploying it. These are not abstract philosophical constructs. They have direct implications for procurement standards, system design requirements, and the conditions humanitarian organisations should place on their technology partnerships.

3. AI SAFETY GOVERNANCE AS A RESPONSE TO HUMANITARIAN TENSIONS

3.1 Operational Contexts: Governing Algorithmic Warfare

The challenge of algorithmic warfare is, in governance terms, a problem of what AI safety researchers call value alignment under distributional shift. A targeting system trained on historical patterns of combatant behaviour will perform unpredictably when those patterns change, as they invariably do in the fluid and complex settings of urban conflict. An error margin that appears acceptable as a statistical parameter translates, in practice, into the deaths of people the system was never designed to target.

AI safety governance contributes three concrete things here. The first is the concept of meaningful human control, which was developed primarily in the autonomous weapons debate but applies much more broadly across the sector. It provides a legally and ethically tractable standard for evaluating whether AI systems preserve or erode human accountability in consequential decision-making. Humanitarian actors and legal scholars can use this concept to challenge the legitimacy of AI-enabled targeting policies and to push for treaty obligations that give it operational rather than merely declaratory force.

The second contribution is the practice of adversarial testing and structured red-teaming. Just as weapons systems are subject to legal review under Additional Protocol I of the Geneva Conventions, AI-enabled systems should be subject to rigorous independent evaluation examining their performance under the conditions most likely to arise in humanitarian settings: dense civilian

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

populations, degraded communications and information environments saturated with disinformation. This is not a radical proposition. It is a straightforward extension of legal review obligations that already exist.

The third contribution is the concept of corrigibility, which means designing systems that remain genuinely open to human correction and oversight rather than optimising autonomously toward fixed objectives. A corrigible system flags its own uncertainty, defers to human judgement in ambiguous situations and supports rather than displaces the deliberative processes that international humanitarian law requires. Advocating for corrigibility as a design standard for military AI gives humanitarian actors a specific, technically grounded demand to bring to intergovernmental forums alongside the normative arguments they already make effectively.

3.2 Professional Reconfigurations: Safety by Design Within Humanitarian Organisations

The reconfiguration of humanitarian roles by AI is, at its core, a question of institutional design: who sets the conditions under which AI tools are used, and on what basis? AI safety governance has developed the concept of deployment constraints, meaning the oversight mechanisms, accountability structures and contextual requirements that should be in place before a given AI system is put to use in a given setting. Humanitarian organisations that internalise this framework can move from being passive consumers of AI products to being active architects of the conditions under which those products operate.

Consider needs assessment. An AI system that processes thousands of field reports to identify emerging crises is, in principle, a valuable tool. In practice it is valuable only if the organisation deploying it understands what data the system was trained on, how it performs across different linguistic and cultural settings, and where it is most likely to fail. AI safety governance research has produced mature frameworks for exactly this kind of analysis, commonly called system cards: structured documentation of what a system was trained on, what it was evaluated against, and what risks it is known to carry. Requiring system cards as a condition of procurement would materially shift the burden of transparency toward vendors and away from buyers who currently bear all the informational risk.

The concern that AI-generated proposal writing will displace substantive social innovation is legitimate and worth taking seriously. But the appropriate response is not refusal. It is governance. Safety researchers have developed use policy frameworks that distinguish clearly between appropriate and inappropriate applications of a given AI system. A humanitarian consortium that agrees on shared use policies for AI in grant writing, specifying for instance that AI may support clarity and structure but may not generate claims about programme outcomes, can protect the integrity of its work without forgoing real efficiency gains that matter most to under-resourced organisations.

3.3 Interdisciplinary Dilemmas: Truth, Accountability and Equity

The three cross-cutting challenges identified in this special issue, relating to the integrity of information, the ethics of automated decisions, and the risk of AI-generated inequality, each find substantive counterparts in AI safety governance research.

On the integrity of information. The proliferation of AI-generated synthetic content poses a direct threat to humanitarian action, which depends on the credibility of witness testimony, field documentation and evidential records. AI safety research on detection and content provenance provides the foundation for a humanitarian digital integrity framework. Organisations such as the Coalition for Content Provenance and Authenticity have developed open standards for authenticating digital content. Humanitarian organisations should treat these as a baseline operational requirement, and donors should consider making adherence a condition of funding.

On automated decision-making. The AI safety principle of human-in-the-loop design ensures that consequential decisions involve a human decision-maker who understands the basis for the system's recommendation. This provides a clear and defensible standard for humanitarian AI governance, and it is not merely a technical preference. In contexts where AI recommendations concern the allocation of life-saving resources, the identification of beneficiaries, or the safety of aid workers, accountability requires a human being who can explain and defend the decision. Humanitarian organisations should develop decision rights frameworks specifying which categories of decision may involve AI, to what degree, and under what oversight conditions.

On equity. The risk that AI amplifies existing inequalities between large and small organisations, between the Global North and South, and between well-connected and marginalised communities is well-documented in the fairness and safety research literature. The policy responses that literature has developed, including open-source models, shared data infrastructure and differential access

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

pricing, translate directly into a humanitarian technology equity agenda. The sector has strong normative grounds and significant convening power to advocate for AI infrastructure that is treated as a global public good rather than a proprietary competitive advantage.

4. AN INSTITUTIONAL ARCHITECTURE FOR HUMANITARIAN AI SAFETY

Translating this analysis into practice requires institutional commitment, not only policy statements. The following represents a minimum viable governance structure for responsible AI adoption across the humanitarian sector.

The first element is a Humanitarian AI Safety Board: an independent, multistakeholder body with a mandate to develop and maintain standards for AI use in humanitarian action, to conduct or commission audits of AI systems used by major actors in the sector, and to provide technical assistance to smaller organisations that lack in-house expertise. Its authority should derive from endorsement by major humanitarian donors, UN agencies and leading NGO consortia. Independence from any single state or commercial technology provider would need to be structurally protected from the outset, not merely affirmed in founding documents.

The second element is a mandatory incident reporting system, modelled on the safety reporting culture that transformed commercial aviation. The analogy is worth dwelling on. Aviation's remarkable safety record over the past half-century came not primarily from improvements to individual aircraft but from systematic, industry-wide learning from incidents and near-misses, sustained by a reporting culture that made disclosure routine rather than exceptional. A humanitarian AI incident database, built with appropriate confidentiality protections and accessible to organisations of all sizes, would allow the sector to learn collectively from its errors rather than each organisation repeating the same mistakes independently.

The third element is a set of model procurement clauses that humanitarian organisations can incorporate into contracts with AI vendors. These clauses would require system cards as a condition of supply, mandate notification of significant model updates or performance changes, establish audit rights, and set out liability arrangements where a system causes demonstrable harm. Developed collaboratively by legal, technical and humanitarian experts, such clauses would substantially reduce the transaction costs of responsible procurement for organisations that currently lack the capacity to draft such provisions themselves.

The fourth element is a sector-wide capacity-building programme. This is not a call for aid workers to become machine learning engineers. It is a call for the kind of AI literacy that enables a field coordinator to ask meaningful questions about a predictive model allocating resources, or a programme manager to assess whether an AI-generated needs analysis reflects what they observe on the ground. That level of critical literacy is achievable at scale, and it is essential.

5. THE POLITICAL ECONOMY OF REFORM

The institutional architecture proposed here is technically feasible and ethically well-grounded. Whether it proves politically achievable will depend on three dynamics.

The first concerns the relationship between the humanitarian sector and major technology providers. The largest AI developers have substantial commercial interests in the sector: as a source of data, as a site of reputational legitimacy, and increasingly as a direct revenue stream. The humanitarian sector's leverage in this relationship is real but largely underused. Collectively, humanitarian organisations represent a meaningful market, a source of morally authoritative advocacy, and a reputational asset that technology companies actively seek. The sector should use that leverage deliberately, conditioning engagement with technology providers on substantive commitments to transparency, safety standards and equitable access rather than accepting vendor terms as given.

The second concerns the intergovernmental negotiation landscape. Recent years have brought significant acceleration in international AI governance activity, from the EU AI Act to the Bletchley Declaration to the work of the UN High-Level Advisory Body on AI. These processes have, on the whole, underweighted humanitarian perspectives. The humanitarian sector has both the standing and the substantive expertise to contribute more effectively, particularly on the question of AI in conflict, where no other community of practice carries equivalent operational knowledge. Developing a coherent humanitarian position on key governance questions and building the diplomatic capacity to advance it should be a leadership priority for the sector.

The third dynamic is internal. The credibility of the sector's advocacy on AI governance depends in part on the integrity of its own AI practices. Organisations that adopt AI tools without adequate governance, that use AI to optimise fundraising at the expense of

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

programme quality, or that allow efficiency pressures to erode meaningful engagement with affected communities will find it difficult to demand accountability from others. The governance reforms proposed in this paper are not only a matter of external advocacy. They are a matter of institutional integrity, and the connection between the two is one the sector should be clear-eyed about.

6. CONCLUSION

The humanitarian sector is at a consequential moment it did not choose. The AI systems now being deployed in conflict zones and disaster response settings are not waiting for governance frameworks to mature. They are actively shaping the operational contexts, professional practices and moral terrain of humanitarian action as it unfolds.

AI safety governance research has developed the conceptual tools, institutional models and, in many cases, the specific policy proposals needed to respond to this challenge. What it has lacked is the empirical authority, normative grounding and political legitimacy that the humanitarian sector carries. Bringing these two traditions into genuine and sustained dialogue offers the most credible path toward AI systems that serve protection rather than displacement, accountability rather than opacity, and human dignity rather than operational convenience.

That dialogue will not happen without deliberate effort. It requires investment in institutions, in capacity, in diplomacy, and in the willingness to hold our own practices to the same standards we seek from others. For people living through conflict and crisis, the cost of not making that investment falls on those least able to bear it. That is reason enough to begin.

REFERENCES

[1] Amodei, D. et al. (2016). Concrete Problems in AI Safety. arXiv:1606.06565.
[2] Bode, I. and Watts, T. (2023). Meaning-less Human Control: Lessons from Air Defence Systems for Lethal Autonomous Weapon Systems. One Earth Future Foundation.
[3] Bryson, J. and Winfield, A. (2017). Standardizing ethical design for artificial intelligence and autonomous systems. Computer, 50(5), 116-119.
[4] Coalition for Content Provenance and Authenticity (2023). C2PA Technical Specification v1.3.
[5] Gabriel, I. (2020). Artificial intelligence, values, and alignment. Minds and Machines, 30(3), 411-437.
[6] Geneva Academy (2023). Autonomous Weapon Systems and International Humanitarian Law: New Challenges and Developments.
[7] Human Rights Watch and IHRC (2023). Stopping Killer Robots: Country Positions on Banning Fully Autonomous Weapons.
[8] International Committee of the Red Cross (2021). ICRC Position on Autonomous Weapon Systems.
[9] Krakovna, V. et al. (2020). Avoiding Side Effects in Complex Environments. NeurIPS 2020.
[10] Roff, H. and Moyes, R. (2016). Meaningful Human Control, Artificial Intelligence and Autonomous Weapons. Article 36 Briefing Paper.
[11] UN Secretary-General (2023). Interim Report: Governing AI for Humanity. UN High-Level Advisory Body on AI.
[12] UNOCHA (2023). Humanitarian Impact of Digital Technologies: Navigating AI in the Aid Sector.