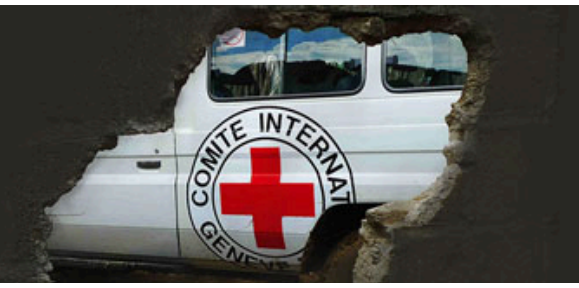
HUMANITARIAN
LAW & POLICY



The (im)possibility of responsible military AI governance

December 12, 2024, Accountability / Analysis / Artificial intelligence in military decision-making / Conduct of Hostilities / IHL / New Technologies / Special Themes

11 mins read

Elke Schwarz
Professor of Political Theory at Queen Mary University London

During 2024, efforts to address the governance of military artificial intelligence (AI) have gained momentum. Yet in the same year, we have also witnessed the growing use of AI decision support systems during armed conflict, and it is becoming clearer that such systems may pose a significant challenge to peace and stability. These developments raise questions about the current approach toward military AI governance.

In this post, Elke Schwarz, Professor of Political Theory at Queen Mary University London, argues that efforts toward governance are complicated by a number of factors intrinsic to contemporary AI systems in targeting decisions. She highlights three in particular: (1) the character of current AI systems, which rests on iteration and impermanence; (2) the dominance of private sector producers in the sector and the financial ethos that grows from this; and (3) the expansive drive implicit in AI systems themselves, especially predictive AI systems in targeting decisions. These realities of AI suggest that the risks are perhaps greater than often acknowledged.

Momentum behind military AI governance is growing. A number of recent initiatives have dedicated space to high-level conversations and negotiations on how to tackle the use of artificial intelligence in the military domain and govern the shift toward greater autonomy in weapon systems. Events like the Responsible AI in the Military Domain (REAIM) summits, for example, are illustrative of this momentum, and several documents have emerged as a result of such meetings, including the 2023 REAIM *Political Declaration* on the Responsible Military Use of Artificial Intelligence and Autonomy, and the outcome document from the 2024 REAIM summit, which endorses a *blueprint for action*.

This year, we have also witnessed the increased roll out and use of AI enabled systems, including AI decision-support systems, in conflict, despite the many "*risks and inefficiencies*" they present. And increasingly, we get a sense that these systems may not herald fewer civilian victims, swift wins, or peace, as it is so often claimed by proponents of military AI. Quite the contrary. The oft-lauded potential benefits of new, AI systems for decision support which include the potential to "*protect civilians and civilian objects*", and the frequently *proclaimed promise* of a swift end to warfare with AI have not materialized. Violence is surging, *children are dying* at a staggering rate, and tensions are growing, not diminishing.

Something is amiss. Unless we examine more carefully why the embrace of AI decision-support systems and other AI applications in military matters perhaps poses much more of a risk than a benefit, politics and policy making will always be on the back foot, resigned to react to industry hype-cycles on one hand, and the stark realities on the ground on the other.

In other words, there are tensions in the speculative aspirations for military AI on one hand, and the present-day realities of conflict with AI on the other, and governance seems to be taking place in the lacuna between the two.

The impermanence of AI as a technique

The foundations of predominant AI, as a statistical data processing technique, and machine learning, as its underlying logic, are iterative. And just like any other software system in the commercial realm, AI systems need frequent updates to stay relevant and functional, even more so in the contested space of battle.

It is suggested that unmanned aerial vehicle (UAV) systems, for example, need to be updated every six to 12 weeks in order to remain effective. AI systems, with significantly more complexity, likely need more frequent updates to adjust to the fluid and adversarial nature of conflict. But with each substantial update, vital systems-aspects may become compromised. Ongoing checks and evaluations are the minimum requirement for military AI systems to be fielded. This takes time, and requires the willingness to prioritise taking this time – a scarce commodity in warfare, where speed of action is paramount. Moreover, extremely robust ethical procedures need to be in place that allow for the possibility that a system will *not* be used if it has not been tested and evaluated appropriately.

Adding to this, the field of AI is developing at pace, at least in terms of scale, but with every new iteration of AI, new problems emerge. Large Language Models (LLMs) are a case in point. Implicit system biases and human automation bias are well-known problems for most 'traditional' AI systems. However, new problems arise with LLMs, such as "*hallucinations*" and the radical *anthropomorphizing* of AI systems. LLMs are merely the latest innovation in AI – there are likely many new minor and major variants on the horizon, each with its own mandates to find ways to use it and consequent problems. Releasing more AI products onto the market will *likely create new problems* faster than it can possibly address existing ones.

New challenges and problems worthy of ethical considerations will emerge from every new iteration – and subsequent implementation – of AI. Reactiveness to the latest AI capabilities cannot be the way forward in regulating AI systems, or finding appropriate norms. The inherent impermanence of AI means we can only ever see what the red lines for a system should be once the systems are already in play. Then the focus becomes risk management, rather than responsibility per se.

The vested financial interests – and attitudes – of stakeholders

The military AI market is *lucrative*. In recent years, it has attracted many non-traditional defense actors who have begun to shape the sector in ways that favors the logics of Silicon Valley industries and its products. This development is bolstered by venture capital (VC) investors who have discovered the defense sector as a market with high potential – some *USD 130 billion* in VC investments have been injected into military technology startups since 2021.

VC investors expect high returns for the risks they take on, and the risks are not insubstantial. *Nearly 90%* of all startups eventually fail, but those that make it big usually yield outsized returns. However, VC investments operate with a different logic than traditional investments in the defense sector. The timelines are shorter, and the promises made for the future need, utility and valuations of the startups are more exaggerated. In order to produce the expected returns on the high-risk, high-reward bets of VC investment, the defense sector must function more like Silicon Valley at large.

Faster timelines for contracts, an embrace of the fail-and-iterate ethos prevalent in Silicon Valley, taking changes and making risky bets – these are the cornerstones that have worked to produce enormous returns for commercial AI products, largely unencumbered by regulatory limits or ethical boundaries. This *ethos is increasingly advocated* as guidelines for the defense sector and military culture more broadly. I have written *elsewhere* in more depth on this dynamic and the outsized influence *VC interest wield on the culture of defense* in the US. VC funded military startups still only make up a relatively small percentage of the overall defense market, but they vie for a more substantial market share by crafting an environment in which their products become indispensable. And to do so, actors with vested interests at times put forward *narratives* that are evocative and are in tension with the aims and ethos of established norms and laws designed to foster restraint in the use of force, not an expansion.

VC companies invest significant sums of money *in lobbying* and in *incentivizing former military and policy staff* to come on board to help create a favorable policy attitude toward military AI systems. Unless we acknowledge the tension in interests by the various invested stakeholders in this military AI domain, effective governance is, to put it bluntly, unlikely. This is a political matter par excellence because it is a matter of power. The ethos and the ethical foundations of these startups and their financial backers matter in the wider context of AI governance and responsibility frameworks. And they deserve scrutiny if responsibility is to be foregrounded.

The logic of AI and our (human) relationship with it

AI is, in its technological foundations, expansionist. In order to function well and as envisioned, it needs large volumes of relevant data and effective interlacing of AI systems. Self-driving cars, for example, would work perfectly if the pedestrians and objects could all be fitted with sensors that correspond to the AI systems in autonomous vehicles (in theory).

With this, AI is also expansionist in a philosophical sense. In 1988, the philosopher Günther Anders made the *following observation*: "Every machine is expansionistic, that is to say, imperialistic; each creates its own service – and colonial empire. And they demand from this colonial empire that it is on hand to work to the same standards as the machine does. .... The machine's hunger for accumulation is insatiable." To put it differently, AI needs more AI in order to function well. And with this everything will be drawn into its wake until, finally, all human affairs will bend to its logic.

And the more intricately humans are embedded in AI systems, and systems of systems, the more likely it is that systems logics determines the practical course of action. As the saying goes: if you have a hammer, you tend to see every problem as a nail. Or, to put it differently, once a system is sufficiently ubiquitous in any context, the question shifts from: "should I use this system", to "how can I use the system more widely" – a mission creep of sorts.

Consider, for example, the problematic allure an AI-enabled decision-support system might hold, not just for the increased speed and scale of finding potential targets and actioning these, but also as modes of "discovering" targets, rogue elements that become marked as "suspicious" and therefore as potentially actionable, in geographies across the globe. This is, of course, a somewhat familiar practice; recall the production of targets through a so-called '*disposition matrix*' in the context of the global war on terror during the early 2010s. Such predictive, machine-learning-based AI systems have attracted *growing interest* from military organizations. With multi-domain connected AI decision support systems that work across geographies, greater emphasis on target *discovery* – rather than target recognition – is a highly plausible development in the wider trajectory of AI in the targeting process.

The wider logic of conflict suggests that such a move will increase suspicion, enmity and will strain relations, not mitigate them. And this is precisely where we are right now with expanding and escalating tensions across the globe.

Conclusion

When the UN Secretary General Antonio Guterres *notes* that the time to act on autonomous weapon systems is now, and that "human agency must be preserved at all cost", he is absolutely right – but the challenges to ensure this are enormous. And the challenge is not just relevant to full autonomy in weapon systems, but it also concerns the growing reliance on AI decision-support systems for military targeting.

I applaud all initiatives that seek to bring together multiple stakeholders in an effort to collectively address this new technology that will indeed transform us, and our relationship to each other and the world, if we let it. These fora are urgent and crucial. But unless there is a willingness *not* to rely on AI if the risks exceed the actual benefits, if the interests of financial stake-holders stand in contrast to the wider aims of the international community, and if the dynamics that the use of AI systems sets loose is contrary to the aims and ethos of limiting violence in war, "responsible AI in the military domain" will remain a chimera.

In other words, if the realities of AI are incommensurable with the ideals of responsible AI, then a greater focus must be placed on when and how *not* to use AI systems in military actions.

See also:

- Erica Harper, *Will AI fundamentally alter how wars are initiated, fought and concluded?*, September 26, 2024
- Matthias Klaus, *Transcending weapon systems: the ethical challenges of AI in military decision support systems*, September 24, 2024
- Wen Zhou & Anna Rosalie Greipl, *Artificial intelligence in military decision-making: supporting humans, not replacing them*, August 29, 2024

Tags: AI, AI decision support systems, artificial intelligence, autonomous weapon systems, AWS, conduct of hostilities, human control, human dignity, IHL, modern warfare

You may also be interested in:

Beyond prevalence: new approaches to measuring sexual- and gender-based violence prevention in conflict settings

12 mins read

Accountability / Analysis / Artificial intelligence in military decision-making / Conduct of Hostilities / IHL / New Technologies / Special Themes

Zuleyka Piniella & Jessica Lenz

When discussing the measurement of sexual- and gender-based violence (SGBV) prevention in humanitarian settings, reactions ...

Trying to square the circle: the ICRC AI Policy

16 mins read

Accountability / Analysis / Artificial intelligence in military decision-making / Conduct of Hostilities / IHL / New Technologies / Special Themes Pierrick Devidal

The development of artificial intelligence (AI) technologies brings significant opportunities, and risks, for principled humanitarian ...