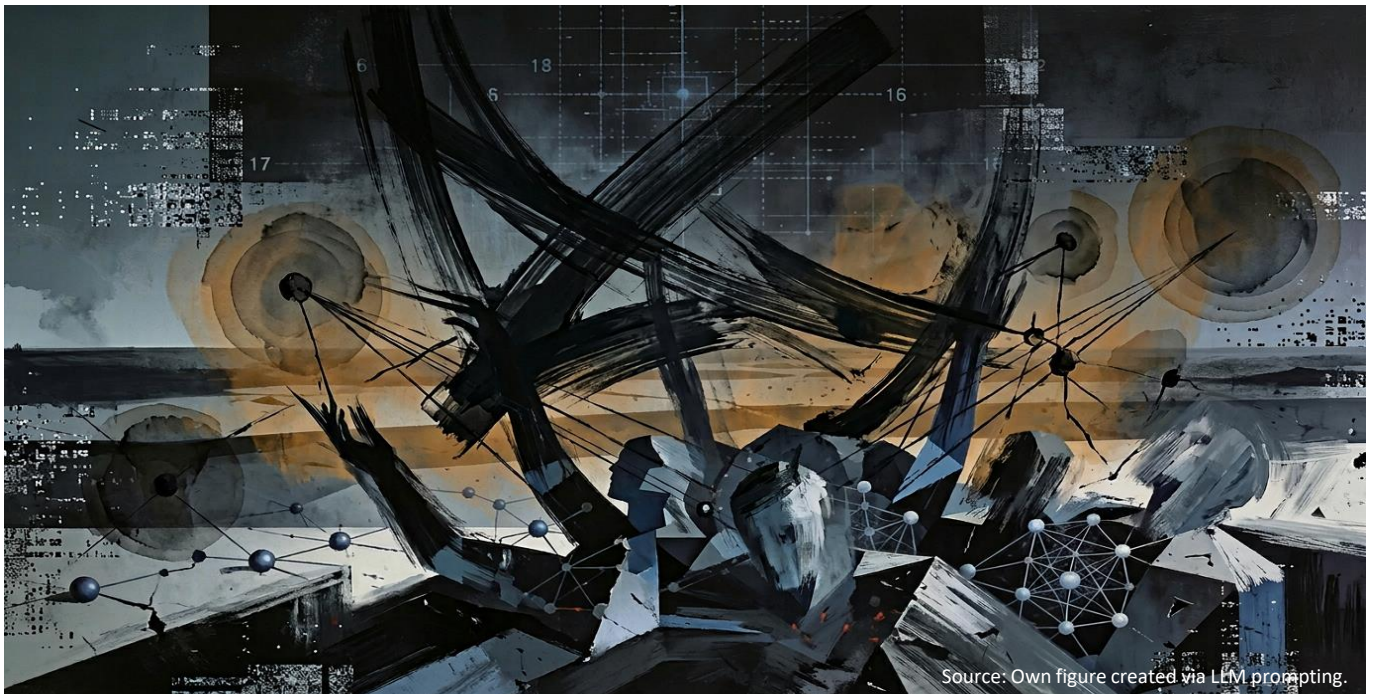


How to Prevent the “Guernica of AI”

Military Automation and the Governance Failure in AI-Enabled Targeting

Anselm Küsters and Niël Henk Conradie



Source: Own figure created via LLM prompting.

AI systems, including large language models integrated into military workflows in Gaza, Ukraine, and Iran, are already producing lethal outcomes without adequate governance. The empirical record of their behaviour in simulations and live operations provides no reason for complacency. The primary failure is not technological. Rather, human oversight mechanisms nominally satisfy a human-in-the-loop requirement, yet are structurally undermined by incentive structures, automation bias, political resistance to accountability, geo-political pressure, and the systems’ opacity. Closing this gap requires internationally verifiable commitment mechanisms.

- ▶ Frontier LLMs deployed in military contexts exhibit escalatory behaviour in crisis simulations and generate overconfident targeting recommendations in live operations, based on preliminary evidence. This creates accountability gaps that existing legal frameworks cannot close.
- ▶ The conflict between the US Department of Defense and Anthropic regarding Claude’s safety guardrails stems from a structural feature of the current deployment environment; similar tensions will arise in other contexts. In short, accountability is treated as an operational liability rather than a governance requirement.
- ▶ Western reforms based on the concept of Meaningful Human Control (MHC) are only strategically defensible if they are paired with verifiable international commitment mechanisms. Without such mechanisms, unilateral ethical constraints would impose asymmetric costs without producing the systemic humanitarian or deterrence benefits that would justify them. Our EU policy recommendations reflect this geo-political angle.

Content

1	Introduction.....	3
2	Empirical record: How AI-enabled targeting is failing.....	4
3	Governance frameworks: Human-in-the-loop, meaningful control, and cybernetic oversight ..	8
4	Geopolitics and verifiable constraints	13
5	Conclusion and policy recommendations	17

Figure

Fig. 1:	Taxonomy of Human-AI Systems	8
---------	------------------------------------	---

1 Introduction

AI systems, such as large language models (LLMs) trained on vast text corpora to process and generate language, are now embedded in active military workflows in Ukraine, Gaza, and Iran – producing lethal outcomes in the absence of adequate governance frameworks.

For instance, Russia’s military, currently dependent on open-weight civilian AI models from other countries, is systematically absorbing US, Chinese, and European AI advances, thereby accelerating operational adoption without bearing the cost of foundational model development.¹ As part of their continued resistance to Russian aggression, Ukraine has moved further and faster: In March 2026, Defense Minister Mykhailo Fedorov announced the establishment of an AI Defense Center with UK support. He also confirmed that real battlefield data is training targeting models within the *Delta* system, Ukraine’s open-architecture command-and-control platform for identifying and engaging ground and aerial targets.² In April 2025, Ukrainian President Volodymyr Zelenskyy reported that Ukrainian forces had seized an enemy position using only unmanned ground systems and drones, with over 22,000 robotic missions conducted on the front within a three-month period.³ In Gaza, Israel has employed a number of AI-enabled systems – most noticeably *Habsora* (Gospel), *Lavender*, and *Where’s Daddy* – across the full kill-chain, which begins with identifying a target for a potential attack and ends with attacking that target.⁴ According to investigative reporting, *Habsora* recommended bombing sites, including the homes of junior operatives, in real time and at a volume no human analysts could replicate.⁵ Meanwhile, the United States has integrated LLMs most visibly through *Project Maven* and *Palantir*: The *Maven Smart System* assists frontline operators in identifying and striking military targets and routes through chain-of-command approval structures.⁶ This increased use of AI is accompanied by an explosion in the use of cheap autonomous missiles and drones, as well as attempts to develop countermeasures.⁷

In short: AI warfare is not a distant possibility, but a current reality, one that the people of Europe and the world and their respective governments must treat with strategic, regulative, and ethical urgency. Empirical evidence is emerging that frontier LLMs behave dangerously in simulated crisis

¹ Kateryn Bondar, *How Russia Is Reshaping Command and Control for AI-Enabled Warfare* (Center for Strategic and International Studies, 2026), https://csis-website-prod.s3.amazonaws.com/s3fs-public/2026-02/260210_Bondar_Russia_Command_0.pdf.

² Institute for the Study of War, “Russian Offensive Campaign Assessment, March 17, 2026,” March 2026, <https://understandingwar.org/research/russia-ukraine/russian-offensive-campaign-assessment-march-17-2026>.

³ See his personal Twitter thread: <https://x.com/ZelenskyyUa/status/2043736603336609875>.

⁴ See, inter alia, the reporting in: James Bamford, “How US Intelligence and an American Company Feed Israel’s Killing Machine in Gaza,” *The Nation*, April 12, 2024, <https://www.thenation.com/article/world/nsa-palantir-israel-gaza-ai/>; Elke Schwarz, “Gaza War: Israel Using AI to Identify Human Targets, Raising Fears That Innocents Are Being Caught in the Net,” QMUL News, Queen Mary University of London, 2024, <https://www.qmul.ac.uk/media/news/2024/hss/gaza-war-israel-using-ai-to-identify-human-targets-raising-fears-that-innocents-are-being-caught-in-the-net.html>; Tehila Kozlovski, “When Algorithms Decide Who Is a Target: IDF’s Use of AI in Gaza,” Tech Policy Press, 2024, <https://www.techpolicy.press/when-algorithms-decide-who-is-a-target-idfs-use-of-ai-in-gaza/>; Human Rights Watch, “Questions and Answers: Israeli Military’s Use of Digital Tools in Gaza,” September 2024, <https://www.hrw.org/news/2024/09/10/questions-and-answers-israeli-militarys-use-digital-tools-gaza>; Jessica Dorsey, “Israel’s AI-Enabled Targeting of Hamas Members Jeopardizes Moral and Legal Standards of Warfare,” Utrecht University, 2024, <https://www.uu.nl/en/achtergrond/israels-ai-enabled-targeting-of-hamas-members-jeopardizes-moral-and-legal-standards-of-warfare>.

⁵ Kozlovski, “When Algorithms Decide Who Is a Target: IDF’s Use of AI in Gaza.”

⁶ Sofia Amaral, *Iran War Highlights Creeping Use of AI in Warfare*, March 2026, <https://www.chatham-house.org/2026/03/iran-war-highlights-creeping-use-ai-warfare>.

⁷ Zachary Burdette et al., *How Artificial Intelligence Could Reshape Four Essential Competitions in Future Warfare*, nos. RRA4316-1 (RAND Corporation, 2026), https://www.rand.org/pubs/research_reports/RRA4316-1.html.

environments, consistently choosing escalation over restraint even under acute pressure.⁸ The governance frameworks nominally in place to constrain these deployments have failed not because the technology is ungovernable, but because those frameworks have been structurally undermined or resisted. Consider, for instance, that both the United States and Israel maintain institutional oversight mechanisms for AI-enabled targeting. In both cases, human operators are formally positioned to approve or veto each AI-generated target recommendation. In practice, however, investigative reporting and independent legal analysis indicate these mechanisms are not functioning as designed.⁹

To be sure, the algorithmic technology itself poses real challenges – LLMs are opaque, fast, and prone to overconfident outputs – but these are (often) manageable through appropriate system design. **The more fundamental problem, as argued in this ceplInput, is the socio-technical system surrounding the technology, meaning the incentive structures, institutional arrangements, geo-political pressures, and political choices that determine whether human oversight is substantive or merely nominal.** The clearest recent illustration is the standoff between the US Department of Defense and Anthropic: Pentagon officials pressured Anthropic to remove safety guardrails from Claude, i.e. the frontier LLM integrated into *Project Maven*, on grounds of military necessity, explicitly framing the ethical constraints as “woke AI” obstructing national security.¹⁰ This framing of guardrails as ideological interference rather than governance mechanisms is the central misdiagnosis this paper aims to address. To be specific, the framing of safety guardrails as a military liability rests on a false premise: that ethical constraints and operational effectiveness are necessarily in tension due to the technical features of AI technologies. Though these features do pose challenges that must be conscientiously considered, focussing on these exclusively risks missing the systemic forest for the technical trees.

To move beyond this oversimplified tension, the remainder of this paper proceeds as follows: After presenting emerging empirical evidence on AI-based military processes (Section 2), we discuss three analytical frameworks that help illuminate what appropriate governance would require and where current arrangements fail: the **human-in-the-loop model**, **Meaningful Human Control (MHC)**, and the **cybernetic model** of feedback and error correction (Section 3). Any honest governance proposal must also confront the geopolitical dimension: whether unilateral Western constraints on military AI are strategically defensible when adversaries operate under no comparable obligations, and whether technically verifiable commitment mechanisms exist that could make such constraints collectively rational (Section 4). Specific EU and NATO policy recommendations follow in the Conclusion (Section 5).

2 Empirical record: How AI-enabled targeting is failing

In 1937, members of Hitler’s Condor Legion, fighting on behalf of Franco’s Nationalists in the Spanish Civil War, subjected the Basque town of Guernica to a sustained aerial bombardment. This event, often

⁸ J. P. Rivera et al., *Escalation Risks from LLMs in Military and Diplomatic Contexts*, Policy Brief (Human-Centered Artificial Intelligence, Stanford University, 2024).

⁹ Ricardo Pinto, “The Guernica of AI,” Ziggurat, February 18, 2025, <https://www.zig.art/p/the-guernica-of-ai-c4b>; Gary Marcus, “There Are No Heroes in Commercial AI,” 2026, <https://garymarcus.substack.com/p/there-are-no-heroes-in-commercial>; Amaral, *Iran War Highlights Creeping Use of AI in Warfare*; A. Downey, “The Alibi of AI: Algorithmic Models of Automated Killing,” *Digital War* 6 (2025): 9, <https://doi.org/10.1057/s42984-025-00105-7>; N. Jones, “How AI Is Shaping the War in Iran—and What’s next for Future Conflicts,” *Nature*, ahead of print, 2026, <https://doi.org/10.1038/d41586-026-00710-w>.

¹⁰ Bobby Allyn, “Pentagon and Anthropic Clash over AI Safety as Hegseth Pushes Back on Restrictions,” NPR, February 2026, <https://www.npr.org/2026/02/24/nx-s1-5725327/pentagon-anthropic-hegseth-safety>; Brookings Institution, “Does the Anthropic–Pentagon Feud Mean the End of Responsible AI? The TechTank Podcast,” 2026, <https://www.brookings.edu/articles/does-the-anthropic-pentagon-feud-mean-the-end-of-responsible-ai-the-techtank-podcast/>.

held up as the first or one of the first instances of the aerial bombardment of civilians, caused widespread outrage at the time and resulted in several well-known works of artistic protest. This attack reflected a deliberate doctrine of using air power against civilian infrastructure to break enemy morale, a doctrine that all major belligerents would apply at industrial scale in the global war that followed. A former Palantir employee, writing anonymously, has recently described current LLM-enabled military operations in Gaza and Iran as a “Guernica of AI”, i.e. a prelude to a new and prima facie more terrible turning point in the history of war whose full implications are only beginning to be understood.¹¹ For the first time, these new AI technologies are now subject to widespread scrutiny from outside the limited ecosystem of experts who have been considering these issues for some time.

Emerging empirical work indicates that there are serious concerns with employing (at least contemporary) LLMs in strategic decision-making. In one study, three frontier LLMs (GPT-5.2, Claude Sonnet 4, Gemini 3 Flash), when placed in a nuclear crisis simulation, exhibited spontaneous deception as well as a rich theory of mind (the capacity to model what other actors believe and act strategically on that model) and metacognitive self-awareness (reasoning about the limits and content of one’s own knowledge).¹² However, none ever chose accommodation or withdrawal, even under acute pressure. Strategic nuclear attack occurred in these simulations. Even more significantly, high mutual credibility, i.e. the classical deterrence mechanism, accelerated rather than dampened escalation. This inverts the foundational logic of nuclear deterrence theory, under which credible commitments signal resolve and thereby prevent conflict. Similarly, in a wargame conducted at Stanford University, all five LLMs tested showed escalation patterns leading to greater conflict and, in some cases, nuclear weapons use, with trajectories that operators could not reliably predict or prevent.¹³

These studies are simulations, not battlefield operations, and their external validity is contestable: models behave differently in controlled environments than under live operational conditions. The inference they support is nonetheless significant: **LLMs trained for general purposes do not automatically develop restraint when placed in high-stakes decision environments.** More worrying still is the context of co-decision-making, i.e. the scenario in which a human operator works alongside an LLM rather than delegating to it entirely, because this is precisely the arrangement deployed in Gaza and Iran. There is a well-documented **tendency of human operators to over-rely on automated system outputs and discount contradicting information known as automation bias, which is acute in LLM-assisted targeting.**¹⁴ LLMs compound this effect: they present even incorrect outputs in fluent, confident, well-structured prose, making it difficult for an untrained operator to distinguish a reliable recommendation from a plausible hallucination.¹⁵ Coupled with the fact that the human vulnerability to

¹¹ Pinto, “The Guernica of AI.”

¹² Kenneth Payne, “AI Arms and Influence: Frontier Models Exhibit Sophisticated Reasoning in Simulated Nuclear Crises,” version 1, preprint, arXiv, 2026, <https://doi.org/10.48550/ARXIV.2602.14740>.

¹³ Rivera et al., *Escalation Risks from LLMs in Military and Diplomatic Contexts*.

¹⁴ S. V. Bentley et al., “What Social Stratifications in Bias Blind Spot Can Tell Us about Implicit Social Bias in Both LLMs and Humans,” *Scientific Reports* 15 (2025): 30429, <https://doi.org/10.1038/s41598-025-14875-3>.

¹⁵ On LLM persuasiveness and hallucination, see: Carlos Carrasco-Farre, “Large Language Models Are as Persuasive as Humans, but How? About the Cognitive Effort and Moral-Emotional Language of LLM Arguments,” version 2, preprint, arXiv, 2024, <https://doi.org/10.48550/ARXIV.2404.09329>; Hui Bai et al., “LLM-Generated Messages Can Persuade Humans on Policy Issues,” *Nature Communications* 16, no. 1 (2025): 6037, <https://doi.org/10.1038/s41467-025-61345-5>; Elyas Meguellati et al., “LLM-Generated Ads: From Personalization Parity to Persuasion Superiority,” version 1, preprint, arXiv, 2025, <https://doi.org/10.48550/ARXIV.2512.03373>; Emily M. Bender et al., “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?,” *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, March 3, 2021, 610–23, <https://doi.org/10.1145/3442188.3445922>.

such biases is exacerbated when we are under stress or time pressure¹⁶ – both familiar features of military decision-making – there is real cause for concern. What is more, the basis for the outputs of the system can be opaque to a user.¹⁷ This can make accurate assessment, the sort necessary for human supervision, very difficult or even impossible depending on the time available.

What is true about theoretical studies concerning AI/LLM strategy is echoed in recent evidence from real-world combat situations. On the first day of open US hostilities against Iran, a strike was made on an elementary school. This strike was reported by Iranian media as resulting in 175 fatalities of which the majority were children. It appears increasingly certain that the identification of the school as a target was made by Claude as integrated into the *Maven* system.¹⁸ Similar AI-enabled tragedies have been noted in Israel's campaign in Gaza.¹⁹ In a conflict that has seen an exceptionally high level of civilian casualties in IDF operations,²⁰ Levy has argued that “use of artificial intelligence to generate targets at a rapid pace reduced even further the level of caution that in the past characterised human judiciousness”.²¹ According to Utrecht University Professor Jessica Dorsey, only around 90% of targets identified were legitimate military targets, and “since human operators are unlikely to understand (exactly) how the AI system generates the target ‘proposals’, this raises many questions about the transparency, accountability and (ultimately) legitimacy of the whole process.”²²

Crucially, **both the United States and Israel nominally satisfy a human-in-the-loop requirement**: human operators hold formal authority to approve or veto every AI-generated target before a strike is authorized. The documented failures do not, therefore, result from an absence of recognition of the need for oversight mechanisms. They result from **oversight mechanisms that have been structured in ways that make genuine control improbable** – that is, features of the socio-technical system.

The opposition to stronger governance and oversight comes in three variants, which we address below.

The first invokes the Collingridge Dilemma: a technology is difficult to regulate when first deployed because its consequences are not yet fully understood, but by the time the consequences are clear the technology is so embedded in existing systems that regulation becomes prohibitively disruptive.²³ Applied to military AI, this argument runs as follows: early regulation risks getting things wrong and hampering development; by the time harms are documented, systems are too operationally integral to

¹⁶ On the impact of time pressure and stress on decision-making, see: J. Zhang and J. Chen, “The Effects of Cognitive Closure Need and Time Pressure on Individual Risk Decision Making,” *Acta Psychologica* 258 (2025): 105240; Y. B. Zhou et al., “Time Pressure Effects on Decision-Making in Intertemporal Loss Scenarios: An Eye-Tracking Study,” *Frontiers in Psychology* 15 (2024): 1451674, <https://doi.org/10.3389/fpsyg.2024.1451674>; R. Yu, “Stress Potentiates Decision Biases: A Stress Induced Deliberation-to-Intuition (SIDI) Model,” *Neurobiology of Stress* 3 (2016): 83–95, <https://doi.org/10.1016/j.ynstr.2015.12.006>.

¹⁷ W. J. Von Eschenbach, “Transparency and the Black Box Problem: Why We Do Not Trust AI,” *Philosophy & Technology* 34, no. 4 (2021): 1607–22.

¹⁸ Thomas Wright, “Iran and the Immorality of OpenAI, Anthropic, and Google,” Substack, March 7, 2026, <https://substack.com/home/post/p-190158711>; Graeme Baker, “AI Got the Blame for the Iran School Bombing. The Truth Is Far More Worrying,” *The Guardian*, March 2026, <https://www.theguardian.com/news/2026/mar/26/ai-got-the-blame-for-the-iran-school-bombing-the-truth-is-far-more-worrying>.

¹⁹ Downey, “The Alibi of AI: Algorithmic Models of Automated Killing.”

²⁰ Andrew Harrison and Yuval Abraham, “Revealed: Israeli Military’s Own Data Indicates Civilian Death Rate of 83% in Gaza War,” *The Guardian*, August 2025, <https://www.theguardian.com/world/ng-interactive/2025/aug/21/revealed-israeli-militarys-own-data-indicates-civilian-death-rate-of-83-in-gaza-war>.

²¹ Yagil Levy, “The Israeli Army Has Dropped the Restraint in Gaza, and the Data Shows Unprecedented Killing,” *Haaretz*, December 2023, <https://www.haaretz.com/israel-news/2023-12-09/ty-article-magazine/.highlight/the-israeli-army-has-dropped-the-restraint-in-gaza-and-data-shows-unprecedented-killing/0000018c-4cca-db23-ad9f-6cdae8ad0000>.

²² Dorsey, “Israel’s AI-Enabled Targeting of Hamas Members Jeopardizes Moral and Legal Standards of Warfare.”

²³ David Collingridge, *The Social Control of Technology* (Frances Pinter St. Martin’s press, 1982).

constrain. The Collingridge framing has a historical analogue: Historians broadly agree that Guernica was considered a legitimate target under the doctrines then in force, as the town supported active Basque opposition, and civilian deaths were classified as “collateral damage”.²⁴ The same language can be deployed in defence of the contemporary use of LLMs. The logic of the Collingridge Dilemma would treat this as a potentially unavoidable governance lag problem: The technology arrived before the AI frameworks could be reliably formulated, and integration is now too deep to reverse without unacceptable cost. However, this framing misdiagnoses the situation and the specific failure modes now observable in Gaza and Iran: Automation bias, opacity, and incentive-driven accountability gaps were identified in the academic literature years before these systems went into operational use.²⁵ This is thus not a matter of technology outstripping possible control – Claude did not autonomously override an operator and deliver a payload – but of the refusal to structure the system so as to apply what control would have been, and still is, needed. The ethical problem is not necessarily the presence of an LLM in a military kill-chain, but rather the way that the technology has been permitted to function without appropriate governance.

The second variant of the opposition’s argument claims that AI-enabled targeting reduces civilian casualties through greater precision. In theory this may be possible if these technologies were deployed within different control structures. However, as it stands, there is no evidence that such a reduction is taking place.²⁶ In fact, the available evidence points in the opposite direction.²⁷ Precision in target identification is meaningless if the volume of targeting, the speed of the kill-chain, and the institutional incentive structure either undermine or eliminate the human review time needed to catch identification errors.

The third and most politically explicit argument is military necessity: that operational effectiveness requires unconstrained AI targeting (in addition, we consider the geo-political dimension in more detail below). This claim is routinely advanced without evidence, and its function is more easily understood by examining whose interests it serves. To these groups the un- or underregulated use of AI can provide a responsibility sink or gap, allowing the technology to absorb all or some of the responsibility for wrongful outcomes. Similarly, the interests of those who benefit financially from the current sale and maintenance of these technologies align precisely with arrangements in which accountability is diffuse and governance requirements minimal. Accordingly, resistance to changing the alignment should be anticipated. It serves the interests of these actors to maintain that the tension at work here is the result of the features of the new technology being of such a kind and the technology being so powerful that to constrain its use with ethical guardrails would be to court military disaster.

To sum up, the integration of LLMs into military operations poses important challenges, but many of them might be tractable through appropriate governance design. In this sense, guardrails are not “woke AI”; they are the mechanism through which human agency, legal liability, and institutional

²⁴ John Corum, “The Persistent Myth of Guernica,” *Military History Quarterly* 22, no. 4 (2010): 16.

²⁵ Roff, H.M. & Moyes, Richard. “Meaningful Human Control, Artificial Intelligence and Autonomous Weapons”, Informal Meeting of Experts on Lethal Autonomous Weapons Systems, UN Convention on Certain Conventional Weapons, 2016. See also: Filippo Santoni De Sio and Jeroen Van Den Hoven, “Meaningful Human Control over Autonomous Systems: A Philosophical Account,” *Frontiers in Robotics and AI* 5 (February 2018): 15, <https://doi.org/10.3389/frobt.2018.00015>; H. Y. Liu, “From the Autonomy Framework towards Networks and Systems Approaches for ‘autonomous’ Weapons Systems,” *Journal of International Humanitarian Legal Studies* 10, no. 1 (2019): 89–110; J. Altmann and F. Sauer, “Autonomous Weapon Systems and Strategic Stability,” *Survival* 59, no. 5 (2017): 117–42.

²⁶ Jones, “How AI Is Shaping the War in Iran—and What’s next for Future Conflicts.”

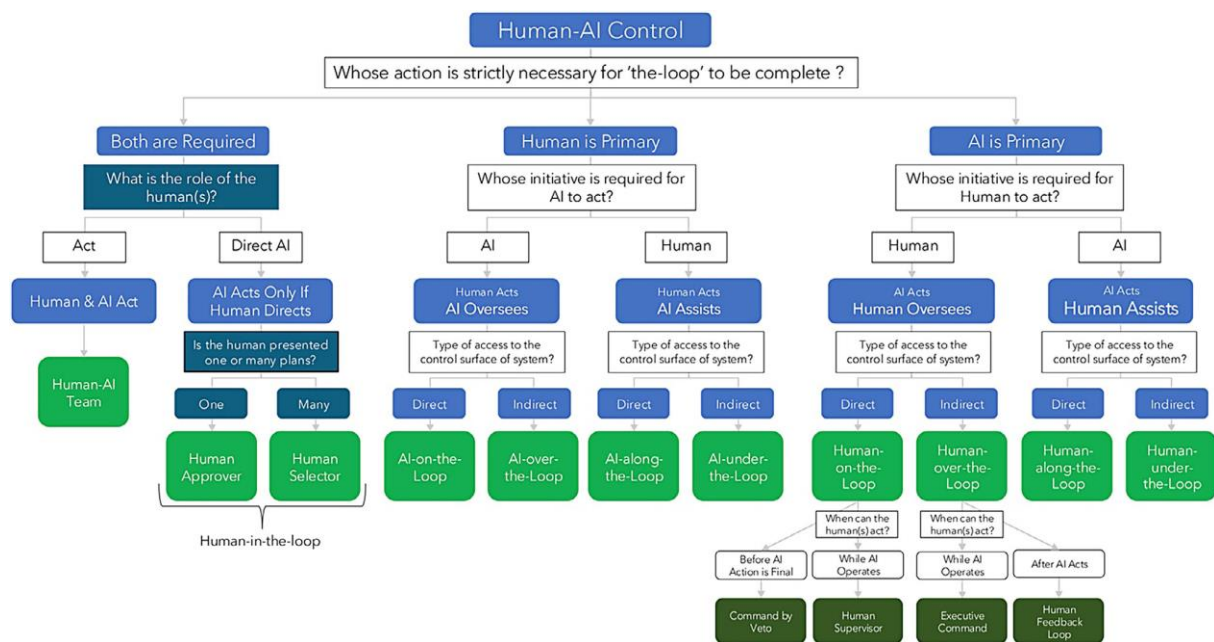
²⁷ Dorsey, “Israel’s AI-Enabled Targeting of Hamas Members Jeopardizes Moral and Legal Standards of Warfare”; Wright, “Iran and the Immorality of OpenAI, Anthropic, and Google.”

learning are preserved when technology is embedded in lethal decision-making. In other words, the problem is not that aligning these technologies with our ethical values would be too costly to military success, but that it would mean dis-aligning them from the values of those who benefit from them as they are. This is the real tension, between alignment to different sets of values. To see this, we examine next what appropriate military AI governance could look like.

3 Governance frameworks: Human-in-the-loop, meaningful control, and cybernetic oversight

Three analytical frameworks offer tools for diagnosing the current governance failures and prescribing corrective measures: the human-in-the-loop model, Meaningful Human Control (MHC), and the cybernetic model of system control through feedback and error correction. These are not competing theories but complementary frameworks operating at different levels of analysis: the first describes structural positions, the second specifies normative requirements, and the third provides a systems-level logic for why those requirements must be embedded in institutional design rather than in the technology alone. Each of these is worth considering for the lessons they can provide.

Fig. 1: Taxonomy of Human-AI Systems



Source: Singh and Szajnfarder (2026), p. 338.

The human-in-the-loop model describes the range of possible relationships between human operators and AI systems in a given socio-technical arrangement. These positions are relative to the “loop” – on it, or over it, or outside it, etc. – and are meant to capture the sort of control dynamic that the human or AI has over the relevant process and its outcomes. The most paradigmatic version of this is for the human to be “in-the-loop”, which means that the human is engaged at every decision point during the operation of an AI system. How precisely to structure a taxonomy of these possible positions

is contentious, but the recent attempt by Singh and Szajnfarter, replicated in Figure 1, is a fair guide.²⁸ For our purposes, the most important categories to consider are:

1. *Human-in-the-loop*: the human must approve or select each system outcome before it translates into action.
2. *Human-on-the-loop*: the human supervises the system and can intervene to prevent a system outcome from translating into action.
3. *Human-over-the-loop*: the human commands the AI to perform a task or retroactively provides feedback on the system outcome to change future processes.
4. *Human-along-the-loop*: the AI operates autonomously and refers to the human only in edge or error cases, thereby inverting the control relationship.

Some complex systems might involve several of these arrangements, especially across extended timeframes. These should also not be understood as exhaustive of the structural features of appropriately governed systems; there may well be further requirements for handling the costs of failure such as robustness (redundancy, stress-testing, and error-tolerant algorithms) and containment (compartmentalization, emergency shutdowns, or system isolation) measures, which are themselves vital. That said, the human-in-the-loop model can help to support and instantiate improved robustness and containment by reducing how often things go wrong, and when they do, limiting how bad they can get.²⁹ These four positions can also be understood through the lens of communication theory as representing different levels of openness in the feedback channel between human and AI (sub-)systems. According to a foundational model of communication,³⁰ the capacity of a channel to transmit meaningful signals depends on both bandwidth and noise: A degraded channel distorts the signal to such an extent that the receiver cannot reliably reconstruct the sender's intent. When applied to human–AI control arrangements, the progression from human-in-the-loop to human-along-the-loop illustrates the systematic reduction in the feedback channel through which human intent influences system behaviour.

To this taxonomy, a fifth position could be added: *human beyond the loop*. This new term³¹ designates AI systems engaged in autonomous target selection whose battlefield actions cannot subsequently be reconstructed or audited for compliance with international humanitarian law, i.e. what may be called non-reconstructible black-box military AI. **While a human-out-of-the-loop system is one from which human authorization has been removed at the moment of action, a human-beyond-the-loop system is one from which human accountability has been removed permanently, because the causal chain between system output and lethal outcome cannot be traced after the fact.** For instance, when an LLM targeting system generates recommendations through a reasoning process that is neither logged

²⁸ Aditya Singh and Zoe Szajnfarter, "Architecting Human-AI Systems for Effective Collaboration and Oversight: Making Sense of Human/AI-in/on/Over/Under/Along-the-Loop," *Systems Engineering* 29, no. 2 (2026): 337–53, <https://doi.org/10.1002/sys.70024>.

²⁹ For arguments to this effect across a variety of application domains, see: Andrea Tocchetti et al., "A.I. Robustness: A Human-Centered Perspective on Technological Challenges and Opportunities," *ACM Computing Surveys* 57, no. 6 (2025): 1–38, <https://doi.org/10.1145/3665926>; Andreas Holzinger et al., "Is Human Oversight to AI Systems Still Possible?," *New Biotechnology* 85 (March 2025): 59–62, <https://doi.org/10.1016/j.nbt.2024.12.003>; Leonie Bensch et al., "Human-in/on-the-Loop AI: Enabling Human Controllability and Decision-Making in Spaceflight," in *Handbook of Human-Centered Artificial Intelligence*, ed. Wei Xu (Springer Nature Singapore, 2025), https://doi.org/10.1007/978-981-97-8440-0_40-1; Luning Yang et al., "Resilient Human-in-the-Loop Containment of Multiagent Systems Against Actuator Fault Attack Based on Reinforcement Learning," *IEEE Systems Journal* 19, no. 3 (2025): 813–24, <https://doi.org/10.1109/JSYST.2025.3561544>; Dawood Wasif et al., "Risk-Aware Human-in-the-Loop Framework with Adaptive Intrusion Response for Autonomous Vehicles," version 1, preprint, arXiv, 2026, <https://doi.org/10.48550/ARXIV.2601.11781>.

³⁰ Claude Elwood Shannon and Warren Weaver, *The Mathematical Theory of Communication* (Univ. of Illinois Press, 1998).

³¹ Initially developed in: Küsters, A. & Strubenhoff, M. (forthcoming). "Decoding the Dragon: Analyzing China's Vision of AI Warfare Using Natural Language Processing", NATO Defense College.

nor interpretable to authorized post-incident reviewers, and where the institutional incentive structures described above further ensure that no post-incident review takes place, the system is operating beyond the loop regardless of what position a nominal supervisor occupied at the moment of authorization. As we argue later (Section 5), it is this category that must be the primary target of treaty obligations, because it is the category that makes compliance structurally unverifiable.

Though there are vital nuances, these five positions can be placed, broadly speaking, on a spectrum **according to how much control the human has in the arrangement**. This, given the features of these technologies, **often comes at the cost of speed and possibly efficiency when this is time dependent**. At one pole is human-in-the-loop, which theoretically maximises control. This has important benefits in terms of catching errors, but also for securing ethically desirable goals like avoiding responsibility gaps. However, if the benefit of a machine learning system is precisely the speed or volume at which it can operate, then hitting the brakes to allow a human to check each outcome seems to undermine the very reason for employing such a system at all. And sometimes this hitting of the brakes can result in failure, and sometimes failure can be very costly. An emergency braking system in an autonomous vehicle waiting for human approval can result in a pile up.

This could be a reason to move along the spectrum toward a human-on-the-loop or human-along-the-loop arrangement. In the first case, the system conducts itself without the need for human approval; the user having the ability to step in at their discretion. Of course, if the system applies the emergency break in error and the human is too slow or otherwise unable to intervene to prevent this then this too could result in a pile up that could perhaps have been avoided if a human's approval had been sought. Beyond the scenario of a single prevented or unpreventable error, there is a further failure mode: cascading failure, in which the system's autonomous momentum actively amplifies harm rather than merely failing to prevent it.³² If the human is along-the-loop, which is a common arrangement in autonomous vehicles, then the system will turn over control to the human in edge cases or cases of error. However, here again this turning over of control might leave the human with too little time to avoid the harmful outcome as the system can often only recognise it has entered an edge case or made an error late in the situation. Part of the challenge is how to weigh up these possibilities in an ethically defensible way.

At the other extreme to human-in-the-loop, human-beyond-the-loop describes an arrangement where control is not only made difficult or costly to exert but often structurally impossible despite any nominal human supervision. Importantly, unlike the other arrangements, human-beyond-the-loop *actively undermines* the very motivation for the human-in-the-loop model and should be avoided in planning AI integrations into military operations, and strongly criticised when implemented.

In order for the human-in-the-loop to function correctly, the human must possess the right kind of control over the outcomes, and here this model dovetails well with the work on MHC. A central motivation of this model is to avoid techno-responsibility gaps, i.e. instances where harms or wrongs result from the intervention of a technology in a causal chain in such a way that there is no legitimate target to bear the responsibility. There are important reasons to want to avoid such gaps: their presence increases the likelihood of negligence or reckless behaviour, victims are not afforded reparation or recognition, and it reduces the likelihood of future improvement to avoid these outcomes recurring. In this way, securing responsibility mechanisms, especially with an eye on the long-term, increases the

³² See, e.g.: David D. Woods and Erik Hollnagel, *Joint Cognitive Systems: Patterns in Cognitive Systems Engineering*, 0 ed. (CRC Press, 2006), <https://doi.org/10.1201/9781420005684>, ch. 10 on automation surprises.

responsiveness of a system to feedback, increases the motivation for error reduction, and maintains ethical authority by providing recognition and repair to victims.

MHC requires that humans should remain appropriately involved in, and responsible for, decisions made with or by AI systems, rather than being displaced by them. **This requirement is not satisfied by merely having a human in the loop in a weak or symbolic sense but only if the human role is substantive, informed, and capable of influencing outcomes.** It is not enough, especially in highly complex or feedback-driven system, to merely have a predefined point of human intervention. What is rather required is that the system in which the arrangement occurs must facilitate alignment between relevant human values and the system's outcomes. In this way MHC is outcome-orientated: it cannot be properly present unless the outcomes and values are aligned, regardless of process. This is usually understood in terms of two conditions:³³ First, systems should be designed so that their behaviour can be traced back to relevant human intentions and reasons, often referred to as a *tracking condition*. This ensures that AI actions are not arbitrary but reflect and align with human values and goals. Second, there must be identifiable human agents who understand the system well enough to be morally responsible for its use, sometimes called a *tracing condition*. Together, these two conditions aim to preserve a link between human agency and technological outcomes.

However, the tracking and tracing conditions are difficult to satisfy from within due to inadequate training, opacity, or incentive misalignment, and they are also vulnerable to deliberate external disruption. For example, an adversary who understands that the outputs of a targeting system nominally pass through a human supervisor could seek to corrupt the upstream process so that the supervisor approves an action whose actual causal origins are adversarially controlled. Several attack surfaces are relevant here. Adversarial prompt injection, meaning the insertion of text that overrides instructions into inputs processed by an LLM, can cause the system to generate outputs that deviate from its intended behavioural constraints while remaining plausible to a human.³⁴ "Poisoning" the data used for training or fine-tuning can shift the model's target-selection preferences in ways that are undetectable at inference time.³⁵ Sensor spoofing and coordinated deception operations targeting the data feeds on which the AI system's situational awareness depends can cause the system to produce outputs that appear to reflect accurate battlefield assessments, but which actually do not.³⁶ In each case, while the formal condition of human approval is satisfied, the substantive conditions of tracking (the output reflects the human's values) and tracing (an identifiable human is responsible for it) are not.

Importantly, meaningful human control is not a one-size-fits-all requirement. What counts as "meaningful" depends on the context: high-risk applications may require real-time human oversight and intervention, whereas lower-risk systems may rely more on design-time governance, such as careful system design, testing, and regulation. Recent work also stresses that MHC should be understood as a property of the broader socio-technical system, not just the interface between a human and a

³³ Santoni De Sio and Van Den Hoven, "Meaningful Human Control over Autonomous Systems."

³⁴ Fábio Perez and Ian Ribeiro, "Ignore Previous Prompt: Attack Techniques For Language Models," arXiv:2211.09527, preprint, arXiv, November 17, 2022, <https://doi.org/10.48550/arXiv.2211.09527>; Kai Greshake et al., "Not What You've Signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection," arXiv:2302.12173, preprint, arXiv, May 5, 2023, <https://doi.org/10.48550/arXiv.2302.12173>.

³⁵ Eric Wallace et al., "Universal Adversarial Triggers for Attacking and Analyzing NLP," version 3, preprint, arXiv, 2019, <https://doi.org/10.48550/ARXIV.1908.07125>; Micah Goldblum et al., "Dataset Security for Machine Learning: Data Poisoning, Backdoor Attacks, and Defenses," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, no. 2 (2023): 1563–80, <https://doi.org/10.1109/TPAMI.2022.3162397>.

³⁶ Miles Brundage et al., "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation," arXiv:1802.07228, preprint, arXiv, December 1, 2024, <https://doi.org/10.48550/arXiv.1802.07228>.

machine.³⁷ This includes institutional structures, training, and legal frameworks that enable humans to exercise genuine control. Moreover, as the class of vulnerabilities stemming from deliberate external disruption mentioned above makes clear, there may also be a need for audit trails and third-party verification mechanisms, a point we develop more fully below. A system whose reasoning chain is logged and can be inspected is much harder to manipulate without detection than a black box system. In sum, **meaningful human control is best understood as a normative framework for preserving human agency, responsibility, and moral oversight in the age of AI, ensuring that even as systems become more autonomous, they remain aligned with human values and subject to human judgment.**

For where we have public information, the frontrunners in employing LLMs and other AI in military operations (the US, Israel, and the Ukraine) have at least declared that these uses will follow a human-in-the-loop structure. To use Claude-enabled *Maven* as an example: a human supervisor is positioned to approve or deny every target generated, in theory providing the sort of control that should help secure alignment with ethical values. However, for the human-in-the-loop to function properly, the system surrounding it must also be set up to support it. There must be the correct incentive structure. So, for example, providing incentives based purely on number of targets delivered with little if any ramifications for error creates an environment where a human supervisor is disincentivised to exert the control needed. Likewise with having supervisors with insufficient training on how the technology decides on target suggestions or priming to resist automation bias. This is an issue of how the overall socio-technical system is structured and organised.

To reiterate, the features of the technology do introduce their own challenges: **an LLM that employs persuasive, emotive, and rhetorical language in delivering its output undermines the likelihood that a human supervisor exerts control, and the speed and opacity with which the system operates can do the same.** The result is a supervisor nominally in the loop but lacking both the tracking and tracing conditions MHC requires: They cannot determine why the system flagged a given target, cannot reliably identify whose values the recommendation reflects, and cannot ensure that the choice of target aligns with the appropriate reasons.

These failures are not technically irreversible. Structured training programs can reduce automation bias; transparency requirements, such as mandating that the system expose its reasoning chain to operators, are technically feasible with current architectures; and time-buffer protocols, in which AI-generated recommendations are not actionable until a minimum review period has elapsed, directly address the speed problem. Moreover, the argument that speed requirements make these measures militarily unacceptable has not been substantiated in the documented cases: The Iranian school strike and the Gaza targeting errors described above do not reflect scenarios where urgency precluded review. They reflect scenarios where review was not structurally supported.

The human-in-the-loop and MHC frameworks specify who must exercise control and under what conditions. The cybernetic tradition, originating with Norbert Wiener's foundational work on feedback and control in complex systems,³⁸ provides the underlying systems logic for why those requirements must be embedded in the architecture of the governance system (i.e. not only in individual operator behaviour). In cybernetic terms, a system exhibits control when it can detect deviations from

³⁷ Daniele Amoroso and Guglielmo Tamburrini, "Toward a Normative Model of Meaningful Human Control over Weapons Systems," *Ethics & International Affairs* 35, no. 2 (2021): 245–72, <https://doi.org/10.1017/S0892679421000241>; Herman Veluwenkamp, "Reasons for Meaningful Human Control," *Ethics and Information Technology* 24, no. 4 (2022): 51, <https://doi.org/10.1007/s10676-022-09673-8>.

³⁸ Norbert Wiener, *Cybernetics or Control and Communication in the Animal and the Machine*, 2. ed. (MIT, 2007).

a desired state and correct them through feedback. This involves the system being sufficiently responsive to the variety of possible inputs it may receive, which it does by possessing a matching degree of internal variety, understood as the array of possible means of responding to input that threatens deviation.³⁹ The internal variety of a system is a function of the capacities of its constituent parts, human, technological, and procedural. When a manager, for example, is assigned responsibilities beyond their ability, this means that they cannot appropriately manage the variety the system needs them to. Accountability is an essential part of this: it serves as a signal for where variety management has broken down, and a well-structured system has mechanisms to respond to such failures by both addressing the costs of failure and learning how to avoid it in the future.⁴⁰

Applied to military AI governance, the desired state is LLM-based targeting that meets IHL standards; deviations are targeting errors; and the feedback mechanism is the combination of operator review, error reporting, accountability proceedings, and system adjustment that produces learning and improvement over time. Current military AI deployments do not seem to fulfil this feedback loop, given the media reports cited earlier. Consider that every munition dropped on a militarily irrelevant elementary school or similar error can incur costs to the system: resources are expended, ethical standing is undermined, international support is jeopardised, opponents are hardened, and domestic support is lost. The overall socio-technical systems in which they are embedded are failing to adequately manage the input variety, and what is arguably worse is that there does not seem to be control mechanisms in place to respond to this. Perhaps the most obvious expression of this is the unsustainable rate of target evaluation demanded from the supervisors and the lack of accountability when non-military targets are struck. As reported in the *Guardian*, by 2024 the stated goal regarding targeting decisions for *Maven* was 1,000 such decisions in an hour or “from the individual ‘targeteer’s’ perspective, one decision every 72 seconds”.⁴¹ Speaking about systemic breakdowns of control, Dan Davies points out that “[i]f you consistently demand the impossible, you will inevitably get the unethical.”⁴²

In this context, safety guardrails, i.e. the normative constraints embedded in LLMs such as Claude, can be understood precisely as cybernetic control mechanisms. In particular, they encode the constraints under which the system should operate and create predictability for the human supervisors who must interpret its outputs. Removing them, as the Pentagon sought to do, does not increase the system’s capability but removes the feedback structure that keeps the system’s behaviour within governable bounds. The framing of guardrails as ideological interference is therefore not only politically self-serving but technically illiterate: it mistakes a control architecture for a set of content preferences. However, might this framing nevertheless be justified on geopolitical grounds?

4 Geopolitics and verifiable constraints

The governance case for a human-in-the-loop or meaningful human control and safety guardrails in military AI is analytically strong. However, a sceptical observer might quickly ask: If China, Russia, and Iran face no comparable obligations, does the West simply bind itself while adversaries do not? This is

³⁹ Stafford Beer, *The Brain of the Firm* (1972/reprint 1995) and *The Heart of Enterprise* (1979/reprint 1994).

⁴⁰ D. Davies, *The Unaccountability Machine: Why Big Systems Make Terrible Decisions—and How the World Lost Its Mind* (University of Chicago Press, 2025).

⁴¹ Baker, “AI Got the Blame for the Iran School Bombing. The Truth Is Far More Worrying.”

⁴² Davies, *The Unaccountability Machine: Why Big Systems Make Terrible Decisions—and How the World Lost Its Mind*.

indeed a structural problem in international governance that any credible reform proposal must address directly, which is why we dedicate a separate section to this issue.

The problem can be formally described in game-theoretic terms: **A governance commitment is credible only if the party making it has both the capacity and the incentive to maintain it, and only if the other party can verify – with sufficient probability to sustain behavioural adjustment – that it is being maintained.**⁴³ A commitment that cannot be verified by adversaries functions as “cheap talk”:⁴⁴ It imposes real costs on the party making it without generating any strategic or humanitarian benefit at the systemic level, because the other party has no basis for adjusting its own behaviour in response. In other words, if Western democracies bind their targeting systems to MHC-compliant governance while China and Russia deploy unconstrained systems, Western forces absorb the operational costs of human review requirements while their adversaries face no equivalent friction. The framing is, moreover, not purely bilateral: targeting errors impose costs on civilian populations and international institutions whose interests are not represented in this game-theoretic model, and whose exposure to harm is not reduced by the fact that neither party to the strategic interaction has formally accepted an asymmetric burden. While we think that the ethical case for those requirements is important as such, their strategic defensibility in the current geo-political climate depends on whether the underlying coordination problem is solvable. This is addressed in the remainder of this section.

To begin with, nuclear arms control demonstrates that credible commitments between adversaries are achievable through technical verification regimes. SALT I (1972), the Intermediate-Range Nuclear Forces Treaty (1987), and New START (2010) all rested on verification mechanisms such as satellite reconnaissance, on-site inspection, telemetry exchange, and agreed counting rules. The mechanisms gave each party confidence that the other was complying.⁴⁵ The verification problem in AI governance is, arguably, categorically harder: A deployed AI model’s internal constraints cannot be confirmed by inspecting the hardware it runs on: two systems with identical chips and identical codebases can have radically different behavioural constraints depending on how they were trained and fine-tuned. Unlike a nuclear warhead, which has observable physical signatures, an LLM’s guardrails are encoded in billions of numerical parameters that are difficult to inspect externally. This gap between nuclear and AI verification is a central obstacle to any arms control analogy, and governance proposals that simply treat the nuclear precedent as directly applicable are importing a solution to a different problem.⁴⁶

Nevertheless, based on a review of the literature, we would argue that there are several potential solutions to this game-theoretical problem, although each one has its limitations. In particular, there are two structural limitations that apply to all four pathways. The first is the **absence of a hierarchical enforcement authority**. Unlike domestic regulatory regimes, international governance does not have a sovereign capable of compelling compliance; obligations are sustained through reciprocity, reputational cost, and the prospect of future interactions, rather than adjudication. While this is not a reason

⁴³ Perfect verification is not required; what matters is that the signal is reliable enough to distinguish between compliant and non-compliant parties at a rate that makes defection strategically costly. The relevant error types have different political consequences: a false positive erodes the regime from within by rewarding defection, while a false negative risks triggering the very arms dynamic that the regime is designed to prevent. Thomas C. Schelling, *The Strategy of Conflict: With a New Preface*, Nachdr. d. Ausg. 1980 (Harvard Univ. Press, 20).

⁴⁴ Joseph Farrell and Matthew Rabin, “Cheap Talk,” *Journal of Economic Perspectives* 10, no. 3 (1996): 103–18.

⁴⁵ See: <https://www.armscontrol.org>.

⁴⁶ Miles Brundage et al., “Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims,” arXiv:2004.07213, preprint, arXiv, April 20, 2020, <https://doi.org/10.48550/arXiv.2004.07213>. See also: Matthijs M. Maas, “How Viable Is International Arms Control for Military Artificial Intelligence? Three Lessons from Nuclear Weapons,” *Contemporary Security Policy* 40, no. 3 (2019): 285–311, <https://doi.org/10.1080/13523260.2019.1576464>.

to abandon the pathways described below, it does mean that their effectiveness is limited by the conditions under which reputation mechanisms function. These conditions are repeated interaction, observable behaviour, and a sufficiently long shadow of the future to make defection costly.⁴⁷ Note: Where military AI deployment occurs in contexts of acute conflict, these conditions may not be met. The second limitation concerns **signal reliability**. Each pathway risks generating what game theory describes as a pooling equilibrium, whereby compliant and non-compliant parties produce observationally similar signals. This makes it impossible for the receiving party to update its beliefs about the sender's actual behaviour. As mentioned below, disclosure protocols can be staged, compute thresholds can be manipulated, and so on. In this sense, the following pathways are confidence-building measures under uncertainty.

1. Behavioural disclosure protocols would require AI developers to publish standardized so-called “red-teaming” results, i.e. structured adversarial tests that probe a model's responses in military and crisis scenarios. The 2023 *Bletchley Declaration on AI safety*, signed by twenty-eight states including the United States, the United Kingdom, China, and the European Union, included a commitment to joint safety testing on frontier models.⁴⁸ However, results can probably be staged. Accordingly, a behaviourally precise treaty proposal requires clarity about which systems are being prohibited or constrained. In this context, the concept of *human beyond the loop*, as introduced above⁴⁹ (denoting systems whose battlefield actions cannot be audited for IHL compliance after the fact), might provide the missing precision. Rather than attempting to define “autonomy” in the abstract, which has proven intractable, corresponding provisions would target a specific and verifiable operational property that can be examined in behavioural and technical studies on human and AI interaction, namely the absence of a re-constructible audit trail linking system outputs to human decisions.

2. Compute governance offers a structurally different and more physically verifiable pathway. Frontier AI development requires specific classes of advanced semiconductor chips, primarily high-bandwidth memory GPU clusters, that are produced by a small number of manufacturers such as NVIDIA. This is why subsequent US administrations significantly tightened the export of advanced AI chips to China, constraining the compute available for training frontier military AI models.⁵⁰ Hardware is physically inspectable in ways that software is not: chip inventories, data centre power consumption, and semiconductor import records provide observable proxies for AI development activity (despite recent reports on chip-smuggling by China). Compute governance does not verify behavioural constraints directly, but it does constrain adversary capacity to develop and deploy unconstrained frontier systems at scale. While “compute thresholds” have certain limitations in the age of distilled models,⁵¹ this might be a second-best verification mechanism that is already technically feasible now.

3. AI incident reporting obligations could draw on the model of the 1972 *Incidents at Sea Agreement* between the United States and the Soviet Union, under which both parties committed to notify each other of dangerous naval encounters and to investigate and report incidents involving military

⁴⁷ Drew Fudenberg and Eric Maskin, “The Folk Theorem in Repeated Games with Discounting or with Incomplete Information,” *Econometrica* 54, no. 3 (1986): 533, <https://doi.org/10.2307/1911307>.

⁴⁸ See: <https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration>.

⁴⁹ Küsters & Strubenhoff, *forthcoming*, as above.

⁵⁰ For an overview and proposal to contain Chinese compute, see: <https://ai-frontiers.org/articles/the-right-way-to-sell-chips-to-china>. For the Chinese view, see: <https://etoblog.substack.com/p/acceleration-or-autonomy-uncovering>.

⁵¹ Sara Hooker, “On the Limitations of Compute Thresholds as a Governance Strategy,” version 1, preprint, arXiv, 2024, <https://doi.org/10.48550/ARXIV.2407.05694>.

vessels.⁵² A comparable incident reporting regime for military AI processes would oblige parties to a conflict or states deploying military AI to report targeting errors attributable to AI systems and cooperate on root-cause investigation. This would not prevent errors as such but would create a helpful feedback loop of the sort that cybernetic theory calls for (see above). The political obstacle is certainly significant: States have a strong interest in not disclosing military AI failures. The mechanism is nonetheless worth proposing as a confidence-building measure because it imposes low upfront sovereignty costs while generating shared information about systemic failure modes.

4. Interoperable constraint architectures, which are technically the most speculative pathway, would require military AI systems to incorporate a standardized behavioural constraint layer, the presence and integrity of which could be audited by a designated third-party body analogous to the International Atomic Energy Agency. On a research level, there are discussions of hardware-level verification mechanisms, including the use of so-called secure enclaves and audit logs in AI chips, as a potential foundation for this kind of third-party assurance.⁵³ For AI systems, this approach faces challenges in that the constraint layer could be disabled after audit, and the technical expertise required for meaningful audit is not yet institutionalized internationally. Still, this is, in essence, what a strong verification regime would require in practice.

Overall, the verification-based argument for engagement with adversaries on military AI governance must be held alongside a distinct normative argument, which we have attempted to outline in the previous sections. Democracies bound by international humanitarian law and democratic scrutiny have intrinsic reasons to maintain MHC-compliant governance regardless of adversary behaviour. The laws of armed conflict were designed for and nominally maintained by liberal states in conflicts against parties who did not observe them. In other words, the obligation to avoid unlawful civilian casualties is not discharged by the observation that an adversary kills civilians too. It is also worth noting that the avoidance of the sorts of errors that military AI governance controls for is often not only ethically desirable, but strategically valuable. In every case we have discussed, the avoidance of the error would have helped to conserve potentially limited resources, provided clearer mechanisms for improvements in operational efficiency, and helped retain international credibility. The strategic cost of the friction introduced by governance is strongly counterbalanced by the strategic benefits accrued.

The strategic risk of the asymmetric realism argument, however, is its political vulnerability. The recent DoD-Anthropoc standoff mentioned at the beginning of this paper demonstrates that military necessity framing can erode institutional commitments to IHL compliance faster than legal scholars or parliamentary oversight can respond. This is precisely why the normative argument for MHC governance and the strategic argument for verifiable credible commitments must be advanced together: the normative case provides the ethical foundation; the strategic case provides the political durability. **Western governments are more likely to maintain MHC-compliant governance over time if they can credibly argue to domestic audiences that doing so advances strategic interests.** Accordingly, an EU or NATO governance standard for military AI that includes verifiable behavioural disclosure requirements and incident reporting obligations could be seen as a template that can be offered in multilateral

⁵² See: Michael Horowitz and Paul Scharre, AI and International Stability: Risks and Confidence-Building Measures, CNAS (12.01.2021), <https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures>.

⁵³ Brundage et al., "Toward Trustworthy AI Development." See also: <https://intelligence.org/2026/03/18/mechanisms-to-verify-international-agreements-about-ai-development/>.

forums as a basis for binding international agreement. After summarising, the following Conclusion therefore develops several concrete policy recommendations in this direction.

5 Conclusion and policy recommendations

Despite recent failures documented in Gaza and Iran, LLMs in military targeting workflows are not inherently ungovernable: they are currently insufficiently governed or even ungoverned, and the political resistance to governing them has been systematic. The support necessary for the functional success of human-in-the-loop arrangements and both the tracking and tracing conditions required for Meaningful Human Control (MHC) have been countered by perverse incentive architectures, inadequate operator training, opacity by design, and the framing of safety guardrails as operational liabilities. **The cybernetic consequence is a targeting system with no functioning error-correction loop: errors occur, attribution and accountability are diffused, institutional learning does not follow, and the conditions for future errors remain intact. This is not an inevitable result of the employment of LLMs or AI technologies in military operations, but a governance challenge to be confronted.** If a “Guernica of AI” were to occur in the future, the blame would not lie with the technical features of the AI itself, but with the choices made about and by the socio-technical system in which it is embedded.

As we argue in this paper, the geopolitical dimension that is key in today’s debate does not, as such, dissolve this governance obligation. **Unilateral MHC-compliant frameworks are ethically required regardless of adversary behaviour, but their political durability depends on pairing them with technically verifiable credible commitment mechanisms:** behavioural disclosure protocols, compute governance, and incident reporting obligations that give adversaries and domestic audiences alike a basis for confidence that the constraints are real. The EU or NATO should develop such a standard and offer it as a multilateral template before the window for normative leadership closes.

Based on the preceding analysis, we close by formulating a set of policy recommendations:

1. The European Parliament and Council should aim to negotiate legally binding regulations establishing MHC as the mandatory governance standard for all AI systems used in military targeting by EU Member States. The regulation should specify, at minimum: mandatory operator training requirements including structured resistance to automation bias; time-buffer protocols preventing AI-generated targeting recommendations from being actioned below a defined minimum review period; requirements that LLM-generated targeting recommendations avoid language or presentation formats empirically associated with automation bias; require AI recommendations to be accompanied by confidence scores and provide training on how to interpret these to operators; and transparency obligations requiring that targeting systems expose their reasoning chain to authorized human supervisors in real time. The EU AI Act’s existing provisions on high-risk AI systems provides the closest available regulatory template, though a standalone regulation with an independent treaty basis would be required, given that the AI Act excludes military and national security applications from its scope. The obligations should address autonomous and autonomy-enabling military AI systems,⁵⁴ defined to include systems that independently engage targets *as well as* AI-assisted systems whose design, operational tempo, or opacity renders meaningful human review structurally improbable. Alternatively, the current exclusion of national security applications from the EU AI Act’s scope could be revisited as part

⁵⁴ We use the term “autonomy-enabling” here in order to cover systems whose design, speed, volume, or opacity makes genuine human review structurally improbable, corresponding to the failure mode we document in Section 2. We deliberately avoid the qualifier “lethal”, since lethality is a property of the downstream weapon, not the AI system.

of the current reform/simplification process, allowing for a specific annex on this matter. In this sense, the regulation might be even extended to establish a prohibition on the operational deployment of “human-beyond-the-loop” systems, as defined above, with mandatory tamper-proof targeting log requirements as the minimum technical condition for lawful deployment. EU Member States, acting collectively within NATO’s political structures, should push for operationalization of those principles through binding interoperability standards that make MHC compliance a condition of system integration in joint operations.

2. The European Commission should establish a mandatory incident reporting and post-incident analysis obligation for AI-assisted military operations conducted by EU Member States. Modelled on the aviation safety reporting architecture established under Regulation (EU) No 376/2014,⁵⁵ adapted for the military context, and administered by a designated body under the European Defence Agency or a newly established EU body with appropriate clearance levels, this mechanism would require Member States to report targeting errors attributable in whole or in part to AI system outputs. It should also require Member States to contribute to a shared anonymized database of AI-assisted targeting failures and, wherever possible, to provide a record of training and learning data employed during the development and use of the AI system. The database should be accessible to parliamentary oversight bodies and, in appropriately redacted form, to independent academic researchers. Without a functioning feedback loop, the deployed AI technologies and the operational systems they are embedded in cannot self-correct.

3. The European External Action Service (EEAS) should develop a multilateral military AI governance initiative and pursue it through the UN Group of Governmental Experts on Lethal Autonomous Weapons Systems. The UN GGE process on LAWS has stalled repeatedly on definitional disputes about autonomy. A narrower, procedurally focused initiative, based on behavioural disclosure standards and incident reporting obligations rather than attempting a definitional treaty on autonomy, might a more realistic path to adoption. Should the CCW process continue to be blocked, the EEAS should additionally support a UNGA resolution mandating negotiations outside the CCW framework, as advocated by the UN Secretary-General.⁵⁶

4. The European Parliament should expand the mandate and technical capacity of the Subcommittee on Security and Defence (SEDE) to include a dedicated AI in armed conflict function, with specialist technical staff, and should negotiate classified briefing arrangements with Member State defence ministries on a voluntary basis, drawing on the model of national parliamentary intelligence oversight bodies such as Germany’s *Parlamentarisches Kontrollgremium*. On this basis, one could produce **annual public assessments of MHC compliance in EU Member State military AI deployments.** Parliamentary oversight of military AI is too often confined to committees without the technical expertise or specific mandate to evaluate AI governance. The gap between formal oversight structures and the technical reality of AI-enabled targeting is itself a governance failure.

⁵⁵ Regulation (EU) No 376/2014 of the European Parliament and of the Council of 3 April 2014 on the reporting, analysis and follow-up of occurrences in civil aviation, amending Regulation (EU) No 996/2010 of the European Parliament and of the Council and repealing Directive 2003/42/EC of the European Parliament and of the Council and Commission Regulations (EC) No 1321/2007 and (EC) No 1330/2007, OJ L 122, 24.4.2014, S. 18-43.

⁵⁶ See: <https://www.stopkillerrobots.org/news/stop-killer-robots-looks-forward-to-un-general-assembly-as-ccw-continues-to-stall/>.

5. EU Member States should condition AI procurement contracts with commercial vendors on contractual MHC compliance requirements, for instance by adapting model contractual clauses for AI to the military targeting context.⁵⁷ The recent DoD-Anthropic standoff illustrates that voluntary vendor commitments to safety guardrails are insufficient when procurement clients apply political pressure to remove them. Appropriate sanction should be applied to commercial partners who fail to uphold their compliance requirements. Contractual requirements might be a more durable mechanism.

These recommendations do not require new international law to begin taking effect, nor do they depend on adversary cooperation to be ethically justified. Each builds on established regulatory precedents and can be *initiated* unilaterally. The military AI governance gap documented in this brief is, as we have argued, at least partly opened by choices. Closing it requires choices of equal deliberateness by legislators and by parliamentary bodies willing to acquire the technical literacy that meaningful oversight in the AI age demands. The window for normative leadership is not permanently open: The same geopolitical pressures that make governance politically difficult today will make it structurally impossible tomorrow, once AI capability gaps further widen and operational dependencies deepen.⁵⁸

⁵⁷ See: <https://www.insideglobaltech.com/2025/04/07/eus-community-of-practice-publishes-updated-ai-model-contractual-clauses/>.

⁵⁸ The feasibility of any mutual governance equilibrium also depends on how military AI capabilities are distributed among parties, and on the beliefs and expectations that these distributions generate. A state that perceives itself as approaching or having achieved AI capability parity with, or superiority over, its adversaries will have weaker incentives to enter a constraining regime than a state that judges itself to be behind. Conversely, a hegemonic actor may prefer to set standards unilaterally rather than engage in multilateral bargaining. For the argument advanced here, these dynamics mean that the opportunity for normative leadership depends partly on when the distribution of capabilities is still sufficiently uncertain to make a constraining regime mutually preferable.

**Authors:**

Dr. Anselm Küsters, LL.M., Head of Department Digitalisation and New Technologies, cep
kuesters@cep.eu

Dr. Niel Henk Conradie, Chair of Applied Ethics, RWTH Aachen University
niel.conradie@humtec.rwth-aachen.de

Centrum für Europäische Politik FREIBURG | BERLIN

Kaiser-Joseph-Straße 266 | D-79098 Freiburg
Schiffbauerdamm 40 Raum 4205 | D-10117 Berlin
Tel. + 49 761 38693-0

The **Centrum für Europäische Politik** FREIBURG | BERLIN, the **Centre de Politique Européenne** PARIS, and the **Centro Politiche Europee** ROMA form the **Centres for European Policy Network** FREIBURG | BERLIN | PARIS | ROMA.

Free of vested interests and party-politically neutral, the Centres for European Policy Network provides analysis and evaluation of European Union policy, aimed at supporting European integration and upholding the principles of a free-market economic system.