Without a Human-in-the-Loop: On Mitigating the Ethical Risks of AI-Enabled and Autonomous Weapons Systems

Jovana Davidovic *

Policy conversations about AI-enabled and autonomous weapons systems have focused on the issue of meaningful human control.[1] In fact, meaningful human control has been at the center of most policy solutions for risks of AI-enabled weapons systems and autonomous weapons.[2] These risks commonly include worries about brittleness—or the inability to predict when AI might fail, bias—or the differential performance over protected categories of people, transparency, and explainability, among others.[3] These risks have long led scholars, policy makers, and the public to be worried about deploying AI-enabled weapons without so-called humans-in-the-loop. The worry has been that, without a human-in-the-loop, the use of AI weapons would pose serious risks to safety, as well as contribute to the inability to assign responsibility for harm

---

* Prof. Davidovic is an Associate Professor of Philosophy at the University of Iowa, where she also holds a secondary appointment at the Center for Human Rights and the Law School. Prof. Jovana Davidovic is also a Senior Researcher at the Peace Research Institute Oslo (PRIO), a Senior Research Fellow at Stockdale Center for Ethical Leadership at the United States Naval Academy, as well as the Chief Ethics Officer for BABL AI, an AI and algorithm research consultancy focusing on AI risk assessments and AI auditing. Professor Davidovic's research focuses on ethics of AI and algorithms, with a focus on ethics of AI-enabled weapons systems. Davidovic is currently leading a 3-year Research Council of Norway funded project "Ethical Risk Management of AI-enabled Weapon Systems," focusing on getting a clear picture of lifecycle of AI weapons from cradle-to-grave, for purposes of developing holistic risk management practices. Prof. Davidovic has audited and consulted on numerous projects assessing ethical and legal risks of AI tools across a range of industries including banking, professional services, insurance, defense, HR, and education.

[1] *See* Rolling text of the 2025 December Session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, at IV, ¶3, https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2026)/CCW_GGE_LAWS_Rolling_Text_-_status_18_December_2025.pdf; Noel Sharkey, *Saying 'No!' to Lethal Autonomous Targeting*, 9 J. MIL. ETHICS, 369, 370-71 (2010); Peter Asaro, *On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-Making*, 94 INT'L REV. RED CROSS 687, 687 (2012), https://international-review.icrc.org/sites/default/files/irrc-886-asaro.pdf; Christof Heyns (Special Rapporteur), Rep. of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions, 16 U.N. Doc No. A/HRC/23/47 (Apr. 9, 2013).

[2] For example, the U.S. Department of Defense ("DoD") requires that autonomous and semi-autonomous weapon systems be designed to allow commanders and operators to exercise appropriate levels of human judgment over the use of force This policy underscores the DoD's commitment to ensuring that human operators maintain control and oversight in the deployment of autonomous weapon systems. *See* U.S. DEP'T OF DEF., OFF. OF THE UNDER SEC'Y OF DEF. FOR POL'Y, DOD DIRECTIVE 3000.09, AUTONOMY IN WEAPON SYSTEMS (Jan. 25, 2023) ("Autonomous and semi-autonomous weapon systems will be designed to allow commanders and operators to exercise appropriate levels of human judgment over the use of force.").

[3] Jimena Sofía Viveros Álvarez, *The Risks and Inefficacies of AI Systems in Military Targeting Support*, INT'L COMM. RED CROSS: HUMANITARIAN L. & POL'Y (Sep. 4, 2024), https://blogs.icrc.org/law-and-policy/2024/09/04/the-risks-and-inefficacies-of-ai-systems-in-military-targeting-support [https://perma.cc/LTW2-NUBW] (last visited Nov. 6, 2025); Jovana Davidovic & Milton Regan, *AI-Enabled Weapons and Just Preparation for War*, BABL A.I. & STOCKDALE CTR. ETHICAL LEADERSHIP (Nov. 9, 2023), https://babl.ai/wp-content/uploads/2023/10/AI-Enabled-Weapons-and-Just-Preparation-for-War.pdf [https://perma.cc/XLB4-8JPU] [hereinafter *Just Preparation for War*].

caused by such weapons.[4] In addition, many have worried that deploying such weapons without a human-in-the-loop would violate the dignity of victims, as well as complicate or undermine the implementation of international humanitarian law.[5]

This paper discusses some of the main problems with over-reliance on meaningful human control for assuring safety and attributions of responsibility. In Section I, the paper surveys some of the recent accounts that stress the importance of meaningful human control. This is done, in part, to clarify the *common risks* that AI-enabled weapons systems and autonomous weapons pose. This section also differentiates between the *common purposes and types* of human control that might be called for. Section II argues that having humans-in-the-loop in the simple sense of having a human be the final or near final decision maker cannot solve most of the problems we have identified as the primary reasons for having the human-in-the-loop in the first place. Section III shortly lay out other potential ways to mitigate some of the risks we started from.

I. ON THE RISKS OF AI-ENABLED WEAPONS AND THE PURPOSES AND TYPES OF HUMAN ENGAGEMENT

Literature is replete with discussions about the importance of humans-in-the-loop and human control for autonomous weapons systems ("AWS") as well as for any other kinds of AI-enabled weapons systems.[6] To start, it is worthwhile distinguishing between AWS and other kinds of AI-enabled weapon systems.

Lethal autonomous weapon systems ("LAWS") are commonly defined as systems that are a "functionally integrated combination of one or more weapons and technological components, that can identify, select, and engage a target, without intervention by a human operator in the execution of these tasks."[7] Essential features of such weapons include the role that AI plays for them, the lack of human oversight, their lethal capability, and their autonomy in target selection. The importance of meaningful human control or the so-called humans-in-the-loop is not however only stressed in discussions of full lethal autonomy, but also more generally in conversations about the justifiability of deploying

---

[4] *Mind the Gap: The Lack of Accountability for Killer Robots,* HUM. RTS. WATCH (Apr. 9, 2015), https://www.hrw.org/report/2015/04/09/mind-gap/lack-accountability-killer-robots [https://perma.cc/C9SR-C2R5] [hereinafter MIND THE GAP]; Peter Königs, *Artificial Intelligence and Responsibility Gaps: What Is the Problem?*, 24 ETHICS & INFO. TECH. 35, 35 (2022).

[5] Asaro, *supra* note 1, at 689; Adriadna Pop, *Autonomous Weapon Systems: A Threat to Human Dignity?*, INT'L COMM. RED CROSS: HUMANITARIAN L. & POL'Y (Apr. 10, 2018), https://blogs.icrc.org/law-and-policy/2018/04/10/autonomous-weapon-systems-a-threat-to-human-dignity [https://perma.cc/GUK8-XRYN].

[6] *See, e.g.*, Rep. of the 2017 Grp. of Governmental Experts on Lethal Autonomous Weapons Sys. (LAWS), annex II, U.N. Doc. CCW/GGE.1/2017/3 (Dec. 22, 2017); INT'L COMM. OF THE RED CROSS, AUTONOMOUS WEAPON SYSTEMS: IMPLICATIONS OF INCREASING AUTONOMY IN THE CRITICAL FUNCTIONS OF WEAPONS (2016).; *Autonomous Weapons Must Remain Under Human Control, Mogherini Says at European Parliament*, EUR. UNION EXTERNAL ACTION (Sep. 14, 2018), https://www.eeas.europa.eu/eeas/autonomous-weapons-must-remain-under-human-control-mogherini-says-european-parliament_en [https://perma.cc/NT47-V38F].

[7] CCW-GGE Rolling text Dec 2025, supra note 1, at I, ¶1.

almost any kind of AI-enabled weapons. AI-enabled weapon systems are commonly defined as weapons that utilize machine learning ("ML") algorithms for some central or ethically salient function. For example, they include weapons that use AI for object recognition, targeting, or decision-support. These systems can operate in autonomous or semi-autonomous modes.[8]

Since AI-enabled weapon systems include any system that utilizes ML for key ethically salient functions, more recently conversations have shifted from ordinary examples of AI tools using neural networks and built via, for example, supervised learning for object recognition and target selection, to potential use of LLMs, chatbots, and agentic AI in weapon systems.[9]

A. *Risks of AI-enabled Weapons*

All of the above – autonomous weapon systems as well as simply AI-enabled weapons systems, and generative AI-driven or agentic AI-driven or ML-driven AI weapons all *share certain features in warfighting contexts that make them ethically risky* (in the same or similar way) and it is those risks that have most commonly given rise to the insistence on having a human-in-the-loop or meaningful human control as a way to mitigate those ethical risks.

These features include the fact that AI tools (of all of the above types) can be brittle, opaque, and biased.[10] And it is those shared features of brittleness, opaqueness, and bias that raise worries around the safety, responsibility attribution, and dignity of and for such tools.

Brittleness refers to the tendency of AI systems to fail or behave unpredictably when faced with scenarios outside of their training data or narrowly defined operational environments.[11] For example, a weapon system trained to recognize tanks in desert environments may misidentify a civilian truck in a snowy setting as a target due to environmental differences between the training data and the use case data. Brittleness is inversely related to the concept of robustness- namely the system's ability to continue operating reliably across a range of conditions, including unexpected, adversarial, or complex

---

[8] Jovana Davidovic, *The Ethics of AI-Enabled Weapon Systems: Testing and Evaluating,* PONTIFICAL ACAD. SOC. SCI., https://www.pass.va/en/publications/studia-selecta/studia_selecta_11_pass/davidovic.html [https://perma.cc/ZX9Y-ZK53] (last visited Nov. 8, 2025).

[9] *See generally* LEWIS HAMMOND ET AL., MULTI-AGENT RISKS FROM ADVANCED AI (2025), https://arxiv.org/abs/2502.14143 [https://perma.cc/AMZ7-6GUJ] (this report discusses an entire taxonomy of potential risks of LLM-based and ordinary multi-agents, not just in the military domain but across industries).

[10] This list of risks is not meant to be exhaustive, instead it is meant to be illustrative of key types of ethical risks of Ai-enabled weapons and specifically illustrative of the types of risks that advocates of "meaningful human control" worry about. *See, e.g.*, FORREST E. MORGAN ET AL., MILITARY APPLICATIONS OF ARTIFICIAL INTELLIGENCE: ETHICAL CONCERNS IN AN UNCERTAIN WORLD 24 (2020), https://www.rand.org/pubs/research_reports/RR3139-1.html [https://perma.cc/N74F-7AW4]; EMMANUEL BLOCH ET AL., ETHICAL AND TECHNICAL CHALLENGES IN THE DEVELOPMENT, USE, AND GOVERNANCE OF AUTONOMOUS WEAPONS SYSTEMS 20, https://standards.ieee.org/wp-content/uploads/import/documents/other/ethical-technical-challenges-autonomous-weapons-systems.pdf [https://perma.cc/S5V5-A9FP] (last visited Nov. 8, 2025); Davidovic, *supra* note 8.

[11] *Just Preparation for War*, supra note 3, at 7.

environments. Brittleness and robustness are counterparts- robust systems perform as expected even when inputs or circumstances deviate from the ideal or training examples, and brittle systems do not. A lack of robustness can lead to unintended escalation or unlawful harm, especially in dynamic combat situations. The brittleness of AI systems- and in general their unexpected performance over varied conditions, is one of the biggest obstacles to creating trustworthy autonomous systems.

Opaqueness refers to the lack of transparency or explainability in how an AI-enabled system makes its decisions. Opaqueness, inexplicability, or unintelligibility are particularly important in those cases that involve life-and-death judgments. Opaqueness creates problems directly and indirectly for accountability and traceability. It creates direct problems for accountability and responsibility attributions because the inability to understand (in highly complex systems) how the final decisions was achieved makes it hard to assign blame, or liability for bad outcomes. It creates indirect problems for accountability and responsibility attribution because even when we chose to place a human-in-the-loop or make a human be in some sense the "final decision maker" their inability to fully understand why the machine provided the output makes it harder to assess whether that person was justified in acting on that output. Simply put, many AI systems in use in weapons and targeting are sufficiently opaque or black box to raise concerns about traceability and post-hoc explanations thus undermining our ability to assure compliance with just war principles and international humanitarian law.[12]

Bias might or might not be present in AI tools used in warfighting, but it is particularly worrisome in cases that might include targeting decisions. Bias (in the ethically salient sense) is the worry that there will be unjustified and unfair differences in the way the AI tool performs across varying protected classes, including race, gender, ethnic, or religious affiliation.[13] The causes for this can be varied, but they commonly include: flawed training data, training data that appropriately captures social biases existent in society, unjustified or unconsidered choices on preferences between false positives and false negatives, etc. The worries around biased tools are numerous, including that biased tools might perform well on aggregate and only get tested on aggregate, but that they might impose all or most risks and harms on a minority, or that such tools will be intentionally misused, and similar.

While these are not all the ethically salient features that AI tools of the above types share, these three common features of complex AI tools are the

---

[12] Nathan Gabriel Wood, *Explainable AI in the Military Domain*, 26 ETHICS & INFO. TECH., 2024, at 28; Scott Sullivan, *Targeting in the Black Box: The Need to Reprioritize AI Explainability*, LIEBER INST. WEST POINT (Aug. 28, 2024), https://lieber.westpoint.edu/targeting-black-box-need-reprioritize-ai-explainability/ [https://perma.cc/Y28M-G8E7]; *XAI: Explainable Artificial Intelligence*, DEF. ADVANCED RSCH. PROJECTS AGENCY, https://www.darpa.mil/research/programs/explainable-artificial-intelligence [https://perma.cc/7C45-L7B8] (last visited Nov. 8, 2025).

[13] Ali Hasan et al., *Algorithmic Bias and Risk Assessments: Lessons from Practice*, DIGITAL SOC'Y, Aug. 2022, at 8; Shea Brown, Jovana Davidovic & Ali Hasan, *The Algorithm Audit: Scoring the Algorithms That Score Us*, BIG DATA & SOC'Y, Jan.—June 2021, at 4.

central reasons that give rise to the calls for meaningful human control. The basic idea is that meaningful human control or having a human in the loop can assure safety and responsibility attribution as well as dignity in cases when we chose to deploy AI tools that we know *have the potential* for brittleness, opaqueness, and bias.[14] Many scholars and policy makers have made exactly such arguments.

Specifically, with respect to policy and international organizations, there has been many that have emphasized the importance of maintaining meaningful human control over autonomous weapons systems. Within United Nations, for example, the Convention on Certain Conventional Weapons ("CCW") has been the main forum for these discussions. CCW discussions have consistently underscored the necessity of ensuring that humans remain directly involved in decisions involving the use of lethal force.[15] In addition, civil society groups, like the Campaign to Stop Killer Robots—a coalition that includes prominent non-governmental organizations such as Human Rights Watch and Amnesty International—has advocated for a ban on fully autonomous weapons. These organizations argue that meaningful human control is essential to maintaining legal accountability and ethical responsibility in armed conflict. Human Rights Watch, in particular, has warned that the delegation of life-and-death decisions to machines risks undermines both international humanitarian law and moral norms.[16] The Future of Life Institute has similarly called for the establishment of international norms governing the use of AI in warfare, emphasizing human oversight as a core principle of responsible AI development.[17] In 2018, the European Parliament passed a resolution calling for an international ban on lethal autonomous weapons systems that function without meaningful human control, asserting that the decision to take a human life should never be entrusted to a machine.[18] The International Committee of the Red Cross ("ICRC") has also reiterated that human control over the use of force is essential to ensure compliance with the laws of armed conflict and to uphold humanitarian principles.[19] Taken together, these positions reflect a growing international consensus among policy makers on the need to ensure that AI-enabled weapons systems always rely on human control or judgment.

---

[14] In the sense of the narrow OODA loop where a soldier like a pilot might engage in observations, orientation, decision, and an action.

[15] Report of the 2015 Informal Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS), Submitted by the Chairperson of the Informal Meeting of Experts, at IV, U.N. Doc. CCW/MSP/2015/3 (June 2, 2015).

[16] *Losing Humanity: The Case Against Killer Robots*, HUM. RTS. WATCH (Nov. 19, 2012), https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots [https://perma.cc/98DJ-PAVG].

[17] *Autonomous Weapons Open Letter: AI & Robotics Researchers*, FUTURE LIFE INST. (Feb. 9, 2016), https://futureoflife.org/open-letter-autonomous-weapons [https://perma.cc/MGJ4-PKLV].

[18] Resolution on Autonomous Weapon Systems, EUR. PARL. DOC. C 433/86 (2018).

[19] INT'L COMM. RED CROSS, AUTONOMOUS WEAPON SYSTEMS: IMPLICATIONS OF INCREASING AUTONOMY IN THE CRITICAL FUNCTIONS OF WEAPONS 46 (2016), https://icrcndresourcecentre.org/wp-content/uploads/2017/11/4283_002_Autonomus-Weapon-Systems_WEB.pdf [https://perma.cc/8WEP-UR6N].

On the side of academia, literature has reflected this same sentiment. For example, in their well-regarded work Santoni de Sio and van der Hoven have argued that all autonomous weapon systems must have meaningful human control and that the two conditions for such control include that the system must track the relevant moral reasons of human operators and that the system must allow outcomes to be traced back to human decisions.[20] While this argument extends the concept of meaningful human control in such a way so as not to oversimplify it to mean human-in-the-OODA-loop or human making the "final" decision, it does still stress that human control in a robust sense is necessary for a justifiable deployment and use of AI-enabled weapons system. In addition to scholars who argue in favor of human control for purposes of accountability and safety, many also have argued for it for purposes of compliance with international law and moral norms. For example, Amoroso and Tamburinni have argued that meaningful human control is necessary to ensure compliance with international humanitarian law and that such control can be maintained through appropriate design and operational procedures.[21] Again, the authors here extend the idea of human control more broadly not just to include *the use* of AI weapons, but also its development and design, but ultimately stress that that design must require human control at the point of use. Ultimately, they argue for a high level of human control, with some potential for exceptions that are agreed upon. Finally, other authors stress the importance of human control for its relevance to presentation of human dignity. Sharkey for example discusses the implications of deploying autonomous weapons systems, often referred to as 'killer robots,' on human dignity. She argues that removing human decision-making from lethal operations raises significant ethical concerns and challenges for the existing laws of war.[22]

B. *Purpose of Human Control of AI-enabled Weapons*[23]

All in all, there seems to be broad agreement among scholars, policy makers, and the public that meaningful human control is a good thing, and a needed thing. These policy makers and scholars, however, often confuse the various aims that such human control is meant to achieve. If we are to decide on whether to support such policies that require human control and if we are to decide what the implementation of such requirements would look like, we need to be clear on: *what the purpose of calls for meaningful human control is all about and what types of human control we are after*. In other words, we need to be explicit about why we have called for meaningful human involvement or engagement with AI

---

[20] Filippo Santoni de Sio & Jeroen van den Hoven, *Meaningful Human Control Over Autonomous Systems: A Philosophical Account*, Feb. 2018, at 1.

[21] Daniele Amoroso & Guglielmo Tamburrini, *Autonomous Weapons Systems and Meaningful Human Control: Ethical and Legal Issues*, 1 CURRENT ROBOTICS REPS. 187, 190 (2020).

[22] Amanda Sharkey, *Autonomous Weapons Systems, Killer Robots and Human Dignity*, 21 ETHICS & INFO. TECH. 75, 75 (2019).

[23] Sections b. and c. rehearse arguments I have made elsewhere in literature including Jovana Davidovic, *On the Purpose of Meaningful Human Control of AI*, FRONTIERS BIG DATA, Jan. 2023, at 2–4; Jovana Davidovic, *Rethinking Human Roles in AI Warfare*, 7 NATURE MACH. INTEL. 1593, 1593 (2025).

tools. And then we need to ask what types of human involvement can satisfy that purpose.

Let's start with purpose. As I have argued elsewhere, there are at least 5 common purposes for human control that policy makers and scholars most commonly invoke or imply in their arguments.[24] Each purpose gives rise to different ethical, legal, or institutional concerns and necessitates a different tailored approach to human engagement or oversight of the tool. These five purposes of human control (that get mentioned or implied in literature and among policy makers) include:

1. Safety: Human control is employed to enhance the accuracy and reliability of AI systems. This is crucial in scenarios where AI decisions can lead to significant harm, and human intervention can mitigate risks.

2. Responsibility and Accountability: Meaningful human control ("MHC") serves to establish clear lines of accountability when AI systems are involved in decision-making processes that may result in harm. By ensuring human involvement, it becomes possible to assign responsibility for outcomes, which is essential for legal and ethical accountability.

3. Democratic Engagement and Consent: Incorporating human control facilitates transparency and allows stakeholders to understand, consent to, or contest decisions made by AI systems. This is particularly important in contexts like criminal justice or public services, where individuals affected by AI decisions should have the opportunity to engage with and challenge those decisions (this purpose is less commonly invoked for AI weapons, but could be particularly salient for arguments for the importance of civilian oversight of military).

4. Institutional Stability: Even if human oversight does not significantly alter outcomes, its presence can enhance public trust in institutions deploying AI. The perception of human control can reassure stakeholders that AI systems are being monitored and that there is recourse in case of errors, thereby contributing to the legitimacy and stability of institutions. This purpose also subsumes arguments that claim only humans can discharge or meet the requirements of international humanitarian law. Also, this argument is sometimes made in the

---

[24] Jovana Davidovic, *On the Purpose of Meaningful Human Control of AI*, 5 FRONTIERS BIG DATA 1, 2 (2023); Jovana Davidovic, *Rethinking Human Roles in AI Warfare*, 7 NATURE MACH. INTEL. 1593, 1593 (2025).

following way: laws and rules are made for humans with human cognitive abilities and motivations and thus it is impossible or hard to apply such rules to AI tools.

5. Human Dignity and Morality: Ensuring that AI systems operate under human control helps assure alignment with moral norms and preservation of human dignity of potential victims. Arguments for human dignity usually stress the idea that being the decision to kill a human being must be made by another human and that dignity (inherently, arguments often say) requires human control at the very last stages of weapons deployment.

Even though an attempt has been made here to clearly differentiate between these purposes, these purposes are most commonly *not clearly articulated*. Even when they are articulated to some extent, scholars and policy makers seem to have more than one of these aims in mind.[25] But institutional and technical design of the AI tools that is meant to meet the above aims will greatly differ depending on what is our main reason for wanting a so-called human-in-the-loop. To illustrate, if we are after human control for safety we should be open to empirical arguments about when, where, and whether human control contributes to safety (and to what sense of safety). But if we are after responsibility attributions, we might be able to rely on normative or institutional arguments for where and when we can hold people accountable for actions partly constituted or caused by AI tools. Being clear about purpose of meaningful human control is important to establish institutional and technical design requirements, meaning where and which human with what training need to be given what information etc. Additionally, knowing the purpose is important because it helps specify or guide the *type of human engagement* or human involvement we are after.

C. *Types of Human Engagement*

One of the reasons to be explicit about the type of purpose for which we want human engagement is to answer the difficult question of which type of human engagement could serve that purpose. Early on in conversations about LAWS and AI weapons both scholars and policy makers insisted on meaningful *human control*, or having a human in the loop- usually referring to the kill chain or the OODA loop. Later on, largely led by political reasons, scholars and policy makers started focusing on *appropriate human judgment* and by doing that they expanded what could count as human engagement and where humans can leverage their deliberative and cognitive skills to satisfy some of those purposes or aims mentioned above. More recently, conversations have expanded to include

---

[25] Jovana Davidovic, *On the Purpose of Meaningful Human Control of AI*, 5 FRONTIERS BIG DATA 1, 2 (2023); Jovana Davidovic, *Rethinking Human Roles in AI Warfare*, 7 NATURE MACH. INTEL. 1593, 1593 (2025).

an even more nuanced concept, namely *context-dependent human judgment*.[26] Parallel to this move from meaningful human control to focus on some kind of human judgment, there has been an expansion of landscape where such human judgment can be exercised – moving from focus on the OODA loop, to the entire lifecycle of the algorithm. In addition and largely prompted by similar concerns, there have also been a move towards talking about *preserving human values* and *preserving human intentions*, with a significant focus on the later. Finally, there has also been some conversation about an inherent value of *human deliberation*- especially when it comes to democratic processes or decisions that affect the democratic process. The wide span of various types of human engagements makes it obvious that we need not only to be asking what the purpose is of having a human-in-the-loop, but also what types of human engagement can serve that purpose. *Without those two (purpose and type of engagement) being explicitly stated we will not have a clear picture of which types of technical designs, and institutional structures to require for various AI tools.*

The differences between these five: human control, human judgment, human deliberation, preserving human intent, and preserving human values are not clear cut, and scholars have spent significant effort trying to disambiguate between them, so while we can't exhaustively define these here, it is nonetheless worthwhile shortly illustrating how these are different.

First, consider human judgement. Usually when we hear talk of the importance of human judgment it is when and because those stressing its importance think that human cognition is superior relative to some purpose. That purpose could be safety, for example. So, if we think of safety as accuracy (which is one way to think about it), we might think there will be times when a tool performs more accurately with human judgment than without it. Coherent policy proposals that insist on human judgment usually do so because they think that relative to some purpose human judgment serves that purpose better and not when empirical evidence doesn't support that claim. The most recent focus has been on appropriate and context-dependent human judgment.

Second, in addition to human judgment, scholars and policy makers have also at times stressed the value of human deliberation and the very act of deliberating. We might think that the act of human deliberation has inherent value, or we might think it has instrumental value. Human deliberation has instrumental value, for example, if it contributes to accuracy, or to institutional stability. It may do so, for example, if we think that human deliberation

---

[26] Global Commission on Responsible AI in the Military Domain (GC REAIM), *Expert Policy Note: Context-Appropriate Human Judgment and International Humanitarian Law*, THE HAUGE CTR. FOR STRATEGIC STUD. (2025), https://hcss.nl/wp-content/uploads/2025/05/Dorsey.pdf [https://perma.cc/HW6U-5XGX]; Global Commission on Responsible AI in the Military Domain (GC REAIM), *Submission to the United Nations First Committee, 80th Session: Responsible Military AI and Context-Appropriate Human Judgement*, THE HAUGE CTR. FOR STRATEGIC STUD. (2025), https://docs-library.unoda.org/General_Assembly_First_Committee_-Eightieth_session_%282025%29/79-239-GC_REAIM-EN.pdf [https://perma.cc/3U2A-NRXY]; CCW GGE, *Rolling text; LAWS, December 2025,* at IV, ¶3, https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2026)/CCW_GGE_LAWS_Rolling_Text_-_status_18_December_2025.pdf [https://perma.cc/Y9GQ-CM72].

contributes to speak out, challenge, and dissent, and if we think those features can contribute to accuracy or institutional stability. Human deliberation has also been seen by some as having inherent value – in as much as it might be a good in itself that increases or changes the morality of actions that come from it regardless of the instrumental contributions such deliberation contributes.

Third, more recently there has been an increased focus on the idea of preserving human intent, as a key thing we should strive for when developing and designing AI-enabled weapons. Often those relying on this concept are interested in preserving is the commander's intent specifically. On this view, we want human control or engagement in the sense that we want to assure that the outcomes of our AI-mediated or AI-augmented actions are in alignment with, or represent, in some sense, the commander's intent. The view is that once a military commander makes a decision about a mission, the aim of our AI tools and institutional policies should be to preserve the substance of their intention all the way through the execution of this mission. This can obviously serve responsibility attributions as well as potentially safety.

Fourth, at other times when scholars talk about the importance of the human engagement for deployment of AI weapons and LAWS, they're really talking about embedding human values into the human machine team, or in into the AI tool. Those values could be the international humanitarian law, they could be the rules of engagement, or they could be some broader moral system of values. They could be embedded in a myriad of ways, sometimes the proposal is that they get embedded through one of the other four types of human engagement, and sometimes the proposal literally is that we code our AI tools with human values in that way (maybe metaphorically, maybe literally) exert human values or human control or exemplify human oversight. The literature around AI alignment often focuses on this type of human engagement.[27]

Finally, there is the original language the community started from, namely the language of meaningful human control. To this day, many, if not most policy makers, insist on human control as the right type of human involvement in deploying AI-enabled weapons. The most commonly stated reason is simply that we want humans to have ultimate control over life-or-death decisions, often in the spirit of dignity, or because we (maybe wrongly) believe that such control maximizes safety. The spirit of such arguments often seems to be that human control in the final stages of a weapon use, simply is what morality requires.

Clarifying the purpose and type of human engagement with AI tools is essential. If our policies to mitigate the risks of deploying AI weapons depend on keeping humans in the loop, we must understand both *why* human involvement is necessary and *what form* it should take. These insights are necessary for guidance on requirements for effective institutional and technical design and our policy proposals are meant to provide clear guidance for both institutional and technical design. Institutional design includes things like where in the loop

---

[27] Iason Gabriel, *Artificial Intelligence, Values, and Alignment*, 30 MINDS & MACHS. 411 (2020) (this paper discusses the main problem of alignment of human values with Ai and centrally focuses on the technological challenges of embedding norms- especially human norms into AI code).

we place the human(s), what sort of training they receive, what hierarchical structures will they fall within, etc. Technical design includes things like: what is the user interface? What type of architecture should we use? What type of architecture is compatible with the purpose and the type of engagement we're after? Is agentic AI appropriate or do we need reinforcement learning? What types of explanations should we give to this human who's supposed to provide that type of human engagement for this specific purpose?

II. WHY HUMAN-IN-THE-FINAL-STAGES-OF-THE-LOOP CANNOT MITIGATE THE KEY RISKS OF DEPLOYING AI-ENABLED WEAPONS

The concept of "meaningful human control" has emerged as a central principle in discussions about the ethics of AI-enabled weapon systems. As mentioned above, advocates often argue that inserting a human in the decision-making loop enhances safety, ensures responsibility assignment, and upholds human dignity. But this assumption warrants scrutiny, not only because calls for human control are not always clear, but also because often human control cannot contribute to the purpose that the advocates are claiming it is needed for. For example, whether human control contributes to safety and accuracy of a tool is an empirical matter. It is obvious that there are times when having a human somewhere in the loop will contribute to safety and there are times when it won't.[28] When this is the case and when it is not will depend on the type of tool we are building and its use cases. For example, if an AI tool is meant to recognize and engage incoming swarms of drones and defend a ship against such overwhelming forces, then it is obvious that a human approving each target (if for example there are 300 incoming attack drones) is not feasible and a ship without a tool to defend itself against such threats ought not to be deployed. So at least in some cases using AI tools without a human providing final approval for targets is necessary. Now of course, one might argue that a human providing a clear description of when such tools may autonomously engage incoming targets is a form of human judgment being implemented to assure safety (to assure for example that the system only gets deployed when it is far from potential shipping lanes, or to assure it only engages drones that show signs of carrying weapons and similar). And that would be right. The point here is simply that human engagement of different types and often far away from the point of target selection and engagement might be what is needed for safety.

Whether and when to utilize humans in the lifecycle of an AI-enabled weapons must be evaluated against real-world performance data and technological realities. There is strong reason to doubt that requiring a human to confirm lethal decisions necessarily improves safety, especially in the case of high-speed, data-intensive environments where AI systems significantly outperform human cognition.

In addition to the fact that it is an open question whether and when human judgment provides for more accuracy and safety, there is also reason to think

---

[28] Jovana Davidovic, *What's Wrong with Wanting a "Human in the Loop"?*, WAR ON THE ROCKS, (June 23, 2022), https://warontherocks.com/2022/06/whats-wrong-with-wanting-a-human-in-the-loop/ [https://perma.cc/A2CC-SYYX].

that often human judgment is at best a mirage. Weapons systems like the U.S. military's Collaborative Operations in Denied Environments ("CODE") illustrate the limitations of human-in-the-loop control.[29] These systems are designed to operate in complex, contested environments with limited communication, where humans can offer only high-level instructions—such as search area boundaries or confirmation of targets—while the system handles most of the tactical execution. The supposed safeguards, like requiring human authorization before firing, often amount to little more than symbolic gestures. If a system has already pre-identified a target, recommended it for engagement, and prepared the action plan, a human "confirming" the target in the final moments often does little to genuinely increase safety.

AI's key advantage lies precisely in its speed and data processing capabilities. It can analyze sensor input, interpret environmental data, and generate action plans far faster than any human operator. In scenarios where milliseconds matter, inserting a human checkpoint risks not only degrading performance, but potentially increasing danger, especially if human hesitation delays an otherwise effective maneuver. Now, advocates for meaningful human control might argue that oversight should occur elsewhere in the AI lifecycle—during design, testing, or rules-of-engagement programming. And they would be right. We ought to ask where in the lifecycle of an AI tool does human moral and legal judgment *truly* shape outcomes- and then we ought to use human judgment there. For some systems that might be before fielding, deployment, and use, and for others it might be in the final stages of use.

In short, the assumption that human-in-the-loop control automatically contributes to *safety* should not be taken for granted. Whether it does so depends on the specific system, the context of its use, and the empirical reality of human cognitive limitations. Policies based on outdated or romanticized notions of human control risk being ineffective—or worse, counterproductive when it comes to safety if they are not grounded in empirical facts about whether or not and when human engagement contributes to safety (whether by safety we mean accuracy or minimizing civilian causalities or something third).

So far, we have focused on the first of the five purposes, namely safety. The reason for this is that almost all calls for meaningful human control of AI weapons explicitly or implicitly seem to aim at safety at least partly. But as we saw above – policy makers and scholars call for human control with other aims in mind as well- the most prominent other aims being responsibility attribution and dignity. Thus, it is worth examining whether similar issues arise with insistence on human control or a human in the loop for purposes of responsibility attribution and dignity.

The worries about responsibility attribution are that the use of AI tools will give rise to the so-called responsibility gap. In recent CCW conversations and scholarship the example of South Korea's SGR-1 Sentry Gun has often been

---

[29] CODE: Collaborative Operations in Denied Environment, DARPA https://www.darpa.mil/research/programs/collaborative-operations-in-denied-environment [https://perma.cc/42Y9-F922] (last visited Nov. 16, 2025).

cited.[30] SGR A1 is a semi-autonomous or potentially autonomous sentry fun what is deployed by South Korea along the DMZ. Using sensors, thermal imaging, and cameras SGR A1 can detect, track, and issue warnings to intruders, but it can also aim and fire a machine gun (although it does need human authorization to engage targets depending on configuration). The worry about the deployment of such tool in our context (other than safety) is what to do if the robot was to autonomously engage a human and if that action turned out to be flawed in some sense- maybe unlawful or excessive. The so-called responsibility gap is the worry that it is not clear who we ought to hold responsible (morally or legally). Is it the developer (Samsung), the procurer and deployer (South Korean military and government) or the user (commander/or tactical officer) that is responsible?[31] The literature is full of worries about cases like this one. Simply put, how do we assign responsibility when AI tools act in unpredictable ways that are not directly traceable to human actions or intentions. Scholars, like Matthias, Nyholm, Sparrow, and others have argued that traditional frameworks for ascribing responsibility may be inadequate for such systems., or that we need to reevaluate our very ideas of responsibility.[32]

Hopefully it is obvious that here too (just like with safety) whether we want to track moral blameworthiness or legal liability and whatever sense of either we have in mind, it is not necessary that we have a human make final or proximate decisions over the tool. On some subset of moral responsibility accounts we can only be held morally responsible for those actions over which we have control (the control principle), or to which we make a substantive difference (the difference principle).[33] But even on those views, it is an open question what would count as control and whether it requires proximity to the outcome and what type of proximity. In other words, no matter what theory of responsibility we rely on it does not necessarily follow that there is a single way to deploy human judgment or control to assure we close the responsibility gap.

In addition, note that any account that requires human engagement with the primary aim of responsibility attribution needs to balance that requirement against the safety of the tool. This is partly because responsibility attributions only matter in cases when the system fails. It would be odd for a policy or a theoretical account to require a human somewhere in the loop for the purposes of making sure we can assign blame when things fail and at the same time for

---

[30] Alexander Velez-Green, *The Foreign Policy Essay: The South Korean Sentry—A "Killer Robot" to Prevent War,* LAWFARE (Mar. 1, 2015, at 10:00 CST), https://www.lawfaremedia.org/article/foreign-policy-essay-south-korean-sentry%E2%80%94-killer-robot-prevent-war [https://perma.cc/WFD4-EMNY].

[31] Similar example is used in Jovana Davidovic, *Bridging the Responsibility Gap for AI-enabled weapons systems*, OXFORD INTERSECTIONS: AI SOCIETY, 2025. Early in the paper (section on Responsibility Gap) the author uses the same example to motivate and explain how a responsibility gap may arise and why it is worrisome to some scholars.

[32] Andreas Matthias, *The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata*, 6 ETHICS & INFO. TECH. 175, 175 (2024); Sven Nyholm, *Attributing Agency to Automated Systems: Reflections on Human–Robot Collaborations and Responsibility-Loci*, 24 SCI. & ENG'G ETHICS 1201, 1201 (2018); Robert Sparrow, *Killer Robots*, 24 J. APPLIED PHIL. 62, 62 (2007).

[33] CHRISTOPHER KUTZ, COMPLICITY: ETHICS AND LAW FOR A COLLECTIVE AGE (2000), 116.

that account to not consider whether placing a human in that place in the loop increases the chances of the system failing.

All in all, closing the responsibility gap need not require human control, and even if it does require human control, that control need not be exhibited simply proximately to the outcome with the human "pulling the trigger" or "pushing the button," and even when such requirements *are* demanded due to the type of an AI tool and the type of an underlying theory of responsibility, those demands need to be balanced against the primary purposes of the AI tool and its safety.

Finally, let us turn to those accounts that claim that we need human control for the purpose of preserving human dignity.[34] For example, Peter Asaro has argued that "[a]s a matter of the preservation of human morality, dignity, justice, and law we cannot accept an automated system making the decision to take a human life."[35] Asaro, Heyns, and others have argued that permitting AI weapons to autonomously make lethal decisions would treat human beings as mere objects. Dignity based arguments are probably the hardest to make sense of, for several reasons. First, as Sharkey points out there are numerous understandings of what human dignity is and what it requires. Second, almost all the accounts that advocate for meaningful human control in the name of dignity, also argue for the need for meaningful human control for purposes of responsibility attribution and safety and so it is hard to pull apart what are the main reasons these policy makers and scholars have in mind when relying on dignity as grounds for justifying calls for human control of AI-enabled weapons systems. To start then, it is probably sufficient in reality to point to the fact that any account that calls for meaningful human control not solely for purpose of dignity, but also safety or responsibly attribution will have to adjudicate what to do if one of those purposes requires one type of human engagement in one place in the life cycle and the other requires a competing type of human engagement and in a different place in the lifecycle. In other words when we have more than one purpose in mind, we need to first ask whether the institutional and technical design required by that purpose is compatible with the institutional and technical design required by the other purpose. In cases when it is possible to accommodate both purposes and neither purpose suffers because of the institutional and technical decisions made to serve the other, it is perfectly acceptable and desirable (assuming both purposes are justified) to try to accommodate both. However, when such purposes conflict the decision needs to be made which purpose is more important and should thus take precedent. This is a difficult question, especially if the conflict is between safety and dignity, but there are some reasons to think that even those arguing in favor of meaningful human control for the sake of dignity would choose safety. One reason to think that is, as already argued because they often mention both in their arguments. The other reason to think this is because on most accounts of

---

[34] Heyns, *supra* note 1 (while Heyns doesn't specifically use the word dignity he argues in a number of places including, p. 1 that autonomous weapons "deployment may be unacceptable...because robots should not have the power of life and death over human beings"); *see also* Sharkey, *supra* note 24 (while Sharkey doesn't argue for this position, her account is a great survey of views that do and the complexities and problems behind such views).

[35] Asaro, *supra* note 1, at 708.

dignity, respecting human dignity requires that we pay heed to what the potential victims would want for themselves, and we need to give some preference to that. If that is the case, then if it turns out that most potential victims would prefer AI weapons that are safer even if autonomous, we ought to respect that. All this still leaves us with three difficult cases. One, cases when it turns out that potential victims actually care more about not being killed by LAWS than they do about safety, two, cases where safety and dignity are compatible and those insisting on dignity thus can ask for final stage human control, and three, cases when in fact scholars really do not confuse and combine safety with dignity and think only dignity matters ultimately. In those three subsets of cases (which jointly make up a very small percentage of potential arguments) we are faced with a genuine reason to want a human in the loop in the final stages of the deployment of the weapon. To summarize, there are only a few genuine cases where scholars have called for human control in the name of dignity and dignity only, and in those cases, there might be reason to prefer having a human in the loop in the ordinary sense of the word even in cases when that threatens more civilian or combatant lives. It is not clear how seriously we should take these calls, partly because they are rare and partly because arguments such as this apply to a range of weapons and not only AI-enabled tools or LAWS.

All of this is to say that a better framework for human control is one that focuses on all the potential ways to mitigate risks of AI weapons, rather than the comforting—but possibly misguided—image of a person sitting behind a kill-switch.

III. HOW TO MITIGATE RISKS OF AI-ENABLED WEAPONS SYSTEMS

What we have seen so far is that *meaningful human control* need not entail the image of a "button-pusher"—a human operator who makes the final decision about deploying a weapon or selecting a target. Instead, we need to pay close attention to the questions of purpose and type of human engagement we are after. If we are serious about achieving goals such as safety, responsibility, attribution, institutional stability, and even human dignity, we must adopt a more expansive analytical lens. We need to consider the entire life cycle of an AI-enabled weapon system or tool, not merely its moment of use, and we might also need to pay attention to the entire targeting cycle within which dozens of AI tools might be deployed. Therefore, if we are to consider practical alternatives to human-in-the-loop, before diving deeper into this broader framework, it is important to ask a foundational question: What is the correct unit of analysis for our policy proposals and scholarly interventions? Should our focus be on the weapon system itself, the targeting process, or the mission execution? Regardless of which unit we select, what becomes clear is that the potential for meaningful human engagement extends far beyond the narrow confines of a single decision point.

A. *Lifecycle of AI Weapons*

Let us begin with the life cycle of a weapon system. Many scholars and policymakers have acknowledged that human control need not consist merely of

pressing a button. Yet, that is often the focal point of policy conversations. To counter this, we might turn to the lifecycle approach that many scholars and policy makers have mentioned. For example, the U.K. delegation to the Convention on Certain Conventional Weapons breaks down the life cycle of an AI-enabled weapon system into the following stages: i. research and development; ii. testing and evaluation—or, more comprehensively, testing, evaluation, validation, and verification ("TEVV"), iii. Procurement, iv. fielding, v. use, vi. post-use assessment and reflection.[36]

If we start thinking of the entire lifecycle of an AI-enabled weapon system what becomes obvious is that there are many human-machine interactions at every stage and those can all be leverage points. Designers and developers make crucial decisions about training data, user interfaces, and the intended uses of the tool. Ideally, they also engage relevant stakeholders to anticipate common or ethically sensitive scenarios. Simply put, designers and developers make a myriad of ethically salient decisions that can act to assure human judgment or human control. Similarly, testers and evaluators also make all sorts of keys decisions in the process of trying to understand not only the technical functioning of the system, but also the context in which it will be deployed. Their evaluations ensure both compliance with international humanitarian law ("IHL") and provide assurance that the tool will perform as expected under real-world conditions. TEVV teams in other words also make a wide range of choices that feed back to development, but also can build calibrated trust, and can inform compliance with human values (like the IHL embodies). The procurement process also presents a number of places where ethically salient decisions get made. For instance, developers or procurers may determine whether to allow end-users to adjust the sensitivity of a system that nominates or validates targets. If the sensitivity is left entirely to the user, that user may choose to lower the threshold for what counts as a legitimate target—potentially broadening the scope of action in problematic ways. Developers, testers, evaluators, and procurers all bear responsibility for creating appropriate constraints and guidelines to ensure the ethical use of such tools. When these actors fail to provide adequate oversight or safeguards, they—not just the end-user—may bear moral and legal responsibility for failures at the deployment or operational stages. They also can in these and many other ways assure safety, or promote dignity. All of this is simply meant to illustrate the broad range of places where humans can exert control, or judgment, or engage in deliberation, that can have significant and meaningful role in assuring those purposes we started from like safety, responsibility, dignity and can mitigate the risks we started from like brittleness, opaqueness, and bias. Moving into the operational domain, further moments of human-machine interaction arise—from the transmission of a commander's intent to tactical users, to the real-time decisions made by operators during a mission. These points of interaction also are obvious places where human control or human judgment can be exerted and contribute to the underlying risk mitigation we are after.

---

[36] MORGAN et al., *supra* note 10, at 126.

All of this suggests that meaningful human control must be distributed across the entire life cycle of an AI system: from research and development, through TEVV and procurement, to fielding, operational use, and post-use assessment. At each stage, human agents can exert influence that cumulatively serves the goals of safety, accountability, and legal compliance. How this is to be done will likely vary from tool to tool, although there must be some similar points of human engagement that can mitigate risks of almost all AI tools. For example, and to oversimplify, it is hard to imagine an AI tool that is safe that doesn't in the development process engage all kinds of stakeholders to understand a range of way in which the tool might be used. This can then in turn can help us navigate design choices around user interface, training of users, and testing and evaluating to assure that these tools do not fail or if they do fail, that they fail "gracefully."[37]

B. *Targeting Process*

In addition to the fact that the entire lifecycle of an AI-enabled weapon provides a landscape for potential intervention with human engagement, so does the targeting process. The targeting process is of particular interest because this is where most of the life-or-death decisions take place and that is what most scholars and policy makers are worried about. In her work *Lifting the Fog of War*, Merel Ekelhof outlines the NATO's conception of a targeting cycle, which involves six stages:

1. Target development
2. Target identification, nomination, and validation
3. Capabilities analysis and selection
4. Commander's intent and force planning
5. Mission execution
6. Post-mission assessment.[38]

In this framework, we shift from focusing on individual AI tools to looking at a broader process- one that might involve multiple AI-enabled systems, each with their own life-cycles and thus with an even broader potential landscape for human oversight and control. Now we are dealing not only with the internal structure of one system, but with the interactions among multiple systems and with the ways that human operators engage with the outputs and inputs of those systems across the targeting cycle.

---

[37] SYSTEMS ENGINEERING RESEARCH CENTER, *SERC TALKS: Progress in Test and Evaluation of AI-enabled Systems in the DoD - Dr. Pinelis (JAIC)*, at 7:40, 39:35 (YouTube, Nov. 9, 2021) https://www.youtube.com/watch?v=1eSKngsJvvo [https://perma.cc/6XT4-RMC2].

[38] Merel A. C. Ekelhof, *Lifting the Fog of Targeting: "Autonomous Weapons" and Human Control Through the Lens of Military Targeting*, 71 NAVAL WAR C. REV. 61, 66-69 (2018); Human Control in the Targeting Process, Presentation at the U.N. CCW meeting of experts on Lethal Autonomous Weapon Systems, (Apr. 12, 2016), https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Informal_Meeting_of_Experts_(2016)/2016_LAWS%2BMX_presentations_towardsaworkingdefinition_ekelhofnotes.pdf [https://perma.cc/B6DD-454V].

Each of these systems, each of these stages, presents *leverage points*—opportunities to exert human influence in ways that matter ethically and operationally. Thus, meaningful human control should not be viewed as something that merely occurs at the moment of weapon release or in the final stage of a targeting decision. Rather, it is the product of human involvement throughout the entire targeting process and throughout the life-cycle of each AI tool.

In conclusion, if we are to make meaningful policy proposals, we need to be clear about what compliance requirements for institutional and technical design best meet our policy goals. And to do that we need to ask first, what purpose do we want human engagement to serve—safety, responsibility attribution, human dignity? Next, we need to ask what types of human engagement best serve that purpose. Then we need to decide on what exactly we are trying to govern (is it the weapon system or the targeting process or a particular type of missions). Only once we are clear and explicit about all of this can we start thinking about how to balance competing values (e.g. safety vs. dignity) and what institutional and technical design choices are best suited for those purposes. In other words, only then can we know where to put and which human in the lifecycle, what sort of training and information does that human need to have to engage in deliberation or appropriate judgment, what the user interface need to be and how must AI decisions be explained to this user, what types of hierarchical and oversight structures do we need to have in institutions to justifiably deploy these tools, etc. These are the key institutional and technical choices that our policy should strive to guide, but as this paper has argued they cannot do so until the purpose, the type of engagement and the unit of analysis/landscape of analysis are clearly laid out. Only by answering these questions clearly—and designing accordingly—can we ensure that human involvement in AI-enabled military systems is truly meaningful.