Ethics and Information Technology (2026) 28:3
https://doi.org/10.1007/s10676-025-09876-9

ORIGINAL PAPER

Autonomous and AI-enabled systems: extensions or replacements of human will and control?

Nathan Wood[1,2]

Abstract
Use of autonomous, AI-enabled, or opaque systems raises many concerns, and some argue that for these to be permissibly deployed in high-stakes or critical domains, they must be subject to so-called "meaningful human control" (MHC). In this article, I focus on the military domain and rebut a strong version of this critique, arguing that off-the-loop systems – i.e., those which can select and engage targets without contemporaneous human input or oversight – can be permissibly deployed while retaining clear lines of responsibility and control. I show that *ex ante* operational constraints and targeting parameters can provide combatants and would-be deployers of off-the-loop systems with strong means to ensure that deployed systems are serving as extensions of combatants' wills, establishing the necessary degree of moral and legal responsibility required. I further show that such constraints and parameters represent clear lines of control that deployers have over such systems, even when these systems, during deployments, are utterly outside of human control. I conclude by distinguishing between what I call "will-extending" and "will-offloading" systems, showing that off-the-loop systems can serve to extend users' and deployers' wills, making such systems inherently subject to meaningful human control. Throughout, the discussion focuses on the example of autonomous and AI-enabled systems in the military domain, but the underlying arguments relate to such systems more generally, showing how these, if utilized as "will-extending" systems, may be used in a controlled and responsibility-retaining manner.

Keywords Artificial intelligence · Autonomous systems · Autonomous weapon systems · Meaningful human control · Responsibility

Introduction

In high-stakes domains where life-and-death decisions are made, many concerns surround the use of autonomous systems, especially ones imbued with AI-processes. In particular, some argue that if humans are not present in the decision-making process, either as direct participants or at least in oversight (and possibly intervening) roles, then these systems will not be subject to so-called "meaningful human control" (MHC).[1] Some critics go on to argue that for autonomous systems to be permissibly deployed in such critical environments, they must, at the least, be subject to meaningful human control.[2]

On the face of it, a demand for human control over autonomous systems is reasonable, as is the requirement that such control be "meaningful". However, in the canonical works most ardently arguing for the necessity of such control, it is often not specified exactly what it would look like in practice.[3] Moreover, upon careful examination, it is clear that on some understandings of "meaningful human control", what is presented as a desideratum for permissible use of autonomous systems turns out to be a ban of such

✉ Nathan Wood
nathan.wood@tuhh.de

[1] Hamburg University of Technology, Hamburg, Germany

[2] California Polytechnic State University, San Luis Obispo, United States

[1] For the canonical formulation, see (Article 36, 2013a, b; Roff and Moyes, 2016).

[2] E.g., Roff and Moyes (2016), Human Rights Watch (2018), Amoroso and Tamburrini (2020), Schwarz (2021).

[3] E.g., Article 36 (2013a, 2013b), Human Rights Watch (2016).

Published online: 25 November 2025

systems.[4] This is because some explications of MHC implicitly require that autonomous systems never be deployed in a fully autonomous manner.

In this article, I resist the strong position that MHC demands contemporaneous control or oversight, arguing instead that control may be established by setting parameters of use, limiting deployments of autonomous systems, and linking autonomous systems' deployments to discrete human decisions. In making this argument, I begin by sketching some existing conceptions of meaningful human control, showing how certain of these function as "bans in disguise" for autonomous systems. I then present a positive case for how we can establish MHC without contemporaneous oversight by setting operational constraints and parameters of use for autonomous systems' deployment. I then explore how a more nuanced understanding of autonomy, actions, and will can inform the discussion of MHC, differentiating between what I call "will-extending" systems and "will-offloading" systems, making the case that will-extending systems allow for meaningful human control of autonomous systems, even when these are operating fully outside of contemporaneous human oversight. I conclude by reiterating that while the arguments rebut critics' objection to using fully autonomous systems in morally and legally fraught environments, the conclusion is not that such control is unnecessary. Rather, the arguments developed show that control can be established and be meaningful even for systems which are operating fully outside of a human's *immediate* control. I thus argue for a broader understanding of MHC, one that is coherent with our understandings of human control over simple(r) artifacts with which we engage on a daily basis.

Before moving onto the arguments themselves though, there are a number of points worth addressing. First, it will be useful to clarify certain terms. As the arguments deal with autonomous systems in general, and autonomous weapon systems (AWS) in particular, a conception of degrees of control is in order. Following the military parlance used for autonomous systems, I use the term "in-the-loop" to denote systems where a human has a direct role in the decision-making of a system, "on-the-loop" to indicate systems which themselves carry out missions and engagements without needing human input, but where a human has persistent override power for the machine's action-outputs, and "off-the-loop" to indicate systems where a human has no ability to guide or override the system once it is deployed.[5] Off-the-loop systems are thus those which may be considered "fully autonomous", and it is these which will be our concern. Building on this broad conception of autonomy in systems, I also, following the United States Department of Defense (US DoD) and the International Committee of the Red Cross (ICRC), take autonomous weapon systems to be systems with autonomy in the "critical functions" necessary for selecting and engaging targets[6] and which can therefore select and engage targets without contemporaneous human intervention.[7] These views of autonomous systems and AWS are broad, including systems which are fully mechanical or partially/fully computerized, which are with or without AI, and, with regards to AWS, which are lethal or non-lethal, anti-personnel or anti-materiel, etc.[8]

As autonomous systems are increasingly AI-enabled or designed using AI-techniques, it is also worthwhile to provide a brief working conception of "AI". We may view AI in terms of human-like intelligence across a broad array of tasks, an ability to engage with complex information for practical reasoning purposes, or a capacity to adapt to a given environment despite having insufficient knowledge and resources.[9] For purposes of this article, I will follow the presentation of the highly influential *Artificial Intelligence: A Modern Approach* by Stuart Russell and Peter Norvig, which takes AI to be "*the study and construction of agents that* **do the right thing**", where "[w]hat counts as the right thing is defined by the objective that we provide to the agent", a general paradigm Russell and Norvig take to be "so pervasive that we might call it the **standard model**".[10]

Deviating from the language common in computer science and used by Russell and Norvig, I also take for granted in what follows that AI-enabled and/or autonomous systems, both as they currently exist and are likely to be in the near future, are not properly agents. As such, these systems do not "make decisions" or "carry out actions" in a meaningful way.[11] Rather, they execute processes, sometimes in a more deterministic manner and sometimes with a probabilistic model serving to inform their outputs. Crucially though, it is not philosophically correct (or good practice) to use the agency-implying language of "decisions" or "actions" when speaking about the functions of autonomous and AI-enabled systems. Throughout what follows, I will speak of "action-outputs" for those things an autonomous or AI-enabled system does, but slips to more agential language are likely. Thus, I stipulate now that all references to what autonomous

[4] E.g., Article 36 (2013a, 2013b), Roff and Moyes (2016), Human Rights Watch (2016, 2016, 2018).

[5] See (Bächle and Bareis, 2022).

[6] International Committee of the Red Cross (2014), p. 5.

[7] International Committee of the Red Cross (2021), p. 1; US Department of Defense (2023), p. 21. See also (Williams, 2015; Boothby, 2016; Altmann and Sauer, 2017; Caron, 2020; Taddeo and Blanchard, 2022; Pacholska, 2024).

[8] Boulanin et al. (2020); Wood (2023).

[9] See, respectively, Newell and Simon (1976), p. 116; McCarthy (1988), p. 308; Wang (2019), p. 19.

[10] Russell and Norvig (2021), p. 22, emphasis in original.

[11] Bryson (2017), p. 134; Evans et al. (2023), esp. pp. 1–2.

systems do should be understood through the lens of "processes", "probabilistic modeling outputs", or similar non-agential understandings, notwithstanding that agential terms may find their way into the discussion.

In exploring permissibility and responsibility in the use of autonomous and AI-enabled systems, especially within the military domain, I also assume throughout that we are presuming actors acting in good faith. This implies that researchers, companies, and states developing and deploying autonomous systems are, at a minimum, seeking to comport with the demands of respective domestic and international laws. With regards to AWS, Article 36 of Geneva Protocol I Additional to the Geneva Conventions (AP I) is especially important, that in developing, testing, and fielding weapons, states are making a reasonable effort to "determine whether [a weapon's] employment would, in some or all circumstances, be prohibited by this Protocol or by any other rule of international law". The assumption of good faith actors is critical, as unless we are willing to accept this caveat, we are not doing good research on autonomous systems or AWS compliance with ethics and law, as assuming bad faith actors is to assume that no rules will matter at all (and so, why explore MHC in the first place?).[12]

It is also worth clarifying that while the concept of "meaningful human control" originated in the debates around autonomous systems in the military, it has since been taken up in numerous other domains such as, e.g., medicine,[13] autonomous vehicles,[14] or the use of artificial intelligence in general[15] (to name only some). Exploration of MHC is thus apt to have broad import – subject of course to arguments made from a particular domain of inquiry being modified as befitting separate (sub-)domains. In this article, I focus on the example and discussions of MHC in the military domain, as this is an area where the ethical and legal implications of automation and autonomy have long been discussed, and, more importantly, because "this is a high-risk domain where lack of control can lead to material damage and harm, so acceptable solutions for controlling AI systems in this domain may be at least equally acceptable in other domains, such as healthcare or transportation, that are at most as risky".[16] Thus, while this article will focus on the example of autonomous and AI-enabled systems in the military, the underlying arguments developed are broad and, subject to modifications, likely germane to debates in other domains where meaningful human control has taken root as a value or design desiderata.

Finally, it is worth noting that while the concept of meaningful human control originated around the regulatory debates surrounding AWS, within central international bodies, the term has since been replaced by the language of "context-appropriate human judgment and control",[17] likely due to a view among some states that MHC had become too politicized of a concept.[18] However, it is important to be clear that this does not reflect a rejection of the underlying intuitions or intent of MHC, as "[c]ontext-appropriate human judgment and control reflects the latest iteration of terms stemming from the conceptualization of 'meaningful human control'. Various conceptions have been proposed, alongside efforts for the term to be seen as a normative feature of the military decision-making process".[19] Indeed, the concept (however one labels it) has seen consistent iterative development,[20] being reflected in ongoing discussions on the design of AI-enabled and autonomous military systems,[21] the teaming of such systems with human combatants,[22] and the ethical,[23] legal,[24] and operational aspects of permissibly developing and deploying these.[25] Beyond this, even if MHC is currently an unfavored term in some circles, the long-standing discussions around meaningful human control have left an indelible mark on the normative debates, and while actual state practices around human control may sometimes be wanting,[26] "the need to exercise a sufficient level of human control over use-of-force decision-making

[12] Zając (2022), Ch. 7.5, especially pp. 269–272; Wood (2024).

[13] E.g., Hille et al. (2023), Beck et al. (2024), Hillis et al. (2024).

[14] E.g., Heikoop et al. (2019), Mecacci and Santoni de Sio (2019), Santoni de Sio et al. (2023).

[15] E.g., Cavalcante Siebert et al. (2023), Davidovic (2023), Abbink et al. (2024), Mecacci et al. (2024).

[16] Tsamados et al. (2024), pp. 1–2.

[17] Group of Governmental Experts on Lethal Autonomous Weapons Systems (2025), p. 2.

[18] Early in the debates around MHC, it was already recognized that the concept had legal limits but could be used for political purposes. See, e.g., Marauhn (2018).

[19] Global Commission on Responsible Artificial Intelligence in the Military Domain (2025), p. 37. The recent ICRC *Submission to the United Nations Secretary-General on Artificial Intelligence in the Military Domain* also includes numerous usages of language highly indicative of the ongoing debates around MHC.

[20] See (Kwik, 2022, 2024a) for detailed examination of official statements, policy, and academic papers on MHC from 2013–2021. See (Roff, 2024) for a higher-level overview of the debates from their origin up to 2024.

[21] AutoPractices (2025), Boshuijzen-van Burken et al. (2024), Global Commission on Responsible Artificial Intelligence in the Military Domain (2025), Trabucco (2025a), Veluwenkamp and Buijsman (2025).

[22] van Diggelen et al. (2024b, 2024a), Veluwenkamp and Buijsman (2025).

[23] Amoroso and Tamburrini (2020), Schwarz (2021).

[24] Ekelhof and Paoli (2020), Paoli et al. (2021), Boutin and Woodcock (2024), AutoPractices (2025).

[25] Ekelhof (2019), Amoroso and Tamburrini (2020), Kwik (2022, 2024a), Bailey (2025), Trabucco (2025b, 2025c).

[26] See, e.g., Bode (2023).

has become a recognised governance principle across various international initiatives",[27] a testament to the impact MHC has had. Thus, while there has been a shift in recent terminology used by the Group of Governmental Experts on Lethal Autonomous Weapons Systems (GGE) working under the Convention on Conventional Weapons and the United Nations Office for Disarmament Affairs, it is by no means the case that MHC has been replaced by a new paradigm, nor is it reasonable to eschew discussion of the concept in light of a single body altering its preferred language. In any event, the arguments to follow will, regardless of the precise terminology one prefers, have clear and significant impact on discussions of "meaningful" or "context-appropriate" human judgment and control of autonomous and AI-enabled systems, whether these are deployed in military or civilian contexts.[28]

Meaningful human control

"Meaningful human control" was first coined by Richard Moyes[29] when he called on the U.K. government to establish regulation over AWS, stating that "[i]t is the robust definition of meaningful human control over individual attacks that should be the central focus of efforts to prohibit fully autonomous weapons".[30] However, Moyes and his non-governmental organization (NGO), Article 36, failed to provide more than a passing conception of what MHC means or requires,[31] and five years after the term's introduction it was still the case that "policy-makers and technical designers lack[ed] a detailed theory of what 'meaningful human control' exactly means".[32] Yet despite its early weaknesses, MHC has since seen much exploration and refinement, and has become central in not just discussions of autonomous systems in the military,[33] but also in medicine,[34] driverless vehicles,[35] and research into autonomous and AI-enabled systems more broadly.[36] From these varying disciplines and their attendant discussions of the concept, numerous aspects have come to the fore, but certain key elements may be picked out.

While each term may be distinctly unpacked,[37] there is a central understanding throughout the debates that what is at stake is a certain type of control – i.e., control which is meaningful – which humans are argued to be responsible for exercising over autonomous and AI-enabled systems. The focus is thus not on which humans are involved nor on the machines themselves which are supposed to be subject to meaningful human control. Rather, the emphasis is on *the control*, on the relation between humans and the action-outputs of machines.

This focus on a particular conception of control follows from a recognition that merely having humans in- or on-the-loop does not necessarily imply that humans are really controlling machines, nor that they are doing so in a meaningful way. Operators of (semi-)autonomous systems may have the capacity to intervene during the deployment of AWS or they may be positively required to exercise fire control during engagements (say, by pressing a button for weapons release), but this does not mean they have the requisite understanding of the operational context and the machine (with its limitations) to afford for a truly meaningful form of control to be established. Something more must be guaranteed.

According to the US DoD, in order for "commanders and operators to exercise appropriate levels of human judgment over the use of force", AWS must, among other things, "[b]e readily understandable to *trained* operators", [p]rovide transparent feedback on system status", and "[p]rovide clear procedures for *trained* operators to activate and deactivate system functions".[38] More philosophically, we may, following Fischer and Ravizza (1998), consider MHC as requiring "guidance control" over actions, where this is understood as agents having reason-responsive control over the mechanisms leading to an action-output, and where those mechanisms can be "the agent's own".[39] To use an example, I may set the thermostat in my office to 72° F. After doing so, the thermostat will adjust the heating/air-conditioning systems as needed, keeping my office at 72° F. In that instance, the thermostat makes changes to the systems responsible for temperature regulation, but the thermostat is a mechanism over which I have control, and where my control is such that the adjusting of heating/air-conditioning is "owned" by me

[27] Bode (2025).

[28] Many thanks to an anonymous reviewer for pressing me to clarify these recent terminological shifts and the iterative nature of understandings of control they indicate.

[29] Roff (2024).

[30] Article 36 (2013a), p. 3.

[31] See, e.g., Article 36 (2013a, 2013b). Roff and Moyes (2016) expands these ideas, but the underpinnings of MHC and its relation to other ethical/legal rules of war are not fully explored.

[32] Santoni de Sio and van den Hoven (2018), p. 1.

[33] E.g., Horowitz and Scharre (2015); Crootof (2016), Homayounejad (2018); Ekelhof (2019); Amoroso (2020), Bode and Watts (2021), Umbrello (2021a).

[34] E.g., Hille et al. (2023), Beck et al. (2024), Hillis et al. (2024).

[35] E.g., Heikoop et al. (2019), Mecacci and Santoni de Sio (2019), Santoni de Sio et al. (2023), Calvert et al. (2024).

[36] E.g., Santoni de Sio and van den Hoven (2018), Veluwenkamp (2022); Cavalcante Siebert et al. (2023); Davidovic (2023).

[37] Robbins (2023).

[38] US Department of Defense (2023), pp. 3–4 (emphasis added).

[39] Fischer and Ravizza (1998). See also Santoni de Sio and van den Hoven (2018), pp. 5–6 for succinct exploration of Fischer and Ravizza's view.

in a particular way. Thus, I possess "guidance control" over the thermostat and the resulting temperature.

Filippo Santoni de Sio and Jeroen van den Hoven build on this, developing a view of MHC predicated on "tracking" and "tracing" conditions for reasons and responsibility; "tracking" requires that "a decision-making system should demonstrably and verifiably be responsive to the human moral reasons relevant in the circumstances... [t]hat is, decision-making systems should track (relevant) human moral reasons",[40] and "tracing" holds that "in order for a system to be under meaningful human control, its actions/states should be traceable to a proper moral understanding on the part of one or more relevant human persons".[41] The intuitions are that machines' processes ought to actively follow and respond to the (moral) reasons of humans deploying those machines, and humans deploying machines ought to have sufficient understanding of the moral facts and values at play during a deployment.[42] Importantly, "moral" facts are also not limited solely to things like value judgments, but can also include factual elements such as "this object is a tank".

Connecting their philosophical conception of MHC to the concrete debates on autonomous weapon systems, Santoni de Sio and van den Hoven conclude that AWS will not satisfy the tracking condition for two reasons: "[f]irst, they are likely to fail in tracking the relevant reasons of the human military personnel behind them; in particular, they cannot track the reasoning required by international law" and "[s]econdly, they are not as flexible as to properly adjust their behavior to the many morally relevant features of the environment in which they operate".[43] However, in making these claims, they say too much, and indeed go beyond how we ought to understand MHC.

There are many systems over which we have meaningful control but where the systems do not track the relevant moral reasons of agents using them. Often, this is simply because systems cannot track *any* reasons, moral or otherwise. Returning to the thermostat, that is a system that functions according to my intent, effectively extending my will, but it does not track any moral reasons I may have for wanting the temperature at a certain level. And while temperature is not the sort of thing we would normally see as being governed by moral reasons, we can imagine cases where it may well be; suppose a colleague with a medical condition is coming over for a meeting, and if the room is too cold he may become ill. That gives me a moral reason to raise the temperature, but my thermostat does not know that reason.

One may object that the thermostat is an *automatic* system rather than an *autonomous* one, and so the distinction does not hold,[44] but we can slightly modify the case to account for this. Rather than a "dumb" thermostat that simply tracks a temperature I give as an input, suppose the thermostat is connected to sensors, detecting what *my* temperature is and regulating the office so that I *perceive it to be* 72° F. It may also be designed to use machine learning (ML) to improve its ability to reliably detect my perceived temperature (due to clothing choices, whether sun is shining onto me through the window, etc.), rendering the system rather "smart" and clearly autonomous, insofar as it can alter states on its own, in response to complex factors, and in ways that are not immediately transparent to the deployer of the system (me). With such a system, this sometimes entails that the thermostat raises/lowers the temperature to meet my changing state (say, I am wearing a thick sweater, so the thermostat lowers the temperature so that, *for me*, it is 72° F). Supposing I am wearing a thick sweater on the day my colleague visits, the thermostat will then "decide" to lower the temperature. This will have negative consequences for my colleague, and fail to track my moral reasons that would mitigate against allowing the temperature to be so lowered, but the thermostat is still very much under my control. Even supposing that I can only change the thermostat early in the morning (for whatever reason), if I do not consider the consequences of my failure to accommodate my colleague's needs, it is not the case that my thermostat fails to be subject to MHC. Rather, it is subject to MHC, and I fail to dispense with all of my moral obligations, possibly simply because I am forgetful.

Beyond this, Santoni de Sio and van den Hoven's position runs afoul of a clarifying remark we made at the outset. They state that "decision-making systems should *track* (relevant) human moral reasons"[45] and "every action of a decision-making system should be traceable",[46] objecting that AWS "cannot track the reasoning required by international law" and "are not as flexible as to properly adjust their behavior to the many morally relevant features of the

[40] Santoni de Sio and van den Hoven (2018), p. 7.

[41] Santoni de Sio and van den Hoven (2018), p. 9.

[42] Though the GGE no longer uses the language of MHC, instead speaking of "context-appropriate human judgment and control", its recent *Rolling Text* includes key statements which directly hearken to the philosophical discussions of MHC. In particular, III.6.A. demands that the effects of AWS be "adequately predictable, reliable, *traceable* and explainable to those responsible for their use", and III.6.B. adds that AWS must be "operated under a responsible chain of command" (Group of Governmental Experts on Lethal Autonomous Weapons Systems (2025), p. 2, emphasis added). Both of these follow the exploration of MHC just presented or implications of that presentation.

[43] Santoni de Sio and van den Hoven (2018), p. 9.

[44] For deeper discussion of this distinction, see, e.g., Roff (2013), pp. 353–354; Williams (2015); Boothby (2016), pp. 247–252.

[45] Santoni de Sio and van den Hoven (2018), p. 7, emphasis in original.

[46] Santoni de Sio and van den Hoven (2018), p. 10.

environment".[47] However, nowhere in international law is it required that *weapons* be able to track reasons or have the flexibility to alter behavior in response to new moral information. And this is because weapons are not required (or expected) to make decisions or carry out actions. And indeed, they do not and cannot.

Autonomous and AI-enabled systems execute processes and carry out programs. Those programs may be open-ended,[48] but they are still programs over which humans have ultimate control.[49] And if humans do not exercise that control responsibly, this does not imply that control is inherently lost, only that it is negligently applied. For example, if I blindly fire a weapon up into the air, I do not have meaningful human control over the exact points where each bullet may land. However, I very clearly have meaningful human control over my weapon. I am just exercising that control in grossly negligent, dangerous, and criminal ways. Moreover, if one wishes to "bite the bullet" and say that one does not have MHC when one blindly fires into the air, then this would further imply that MHC is lost any time environmental conditions impede on the trajectory of kinetic weapons. But wind does not undermine meaningful human control, or if it does, then meaningful human control is a rather odd and unhelpful concept which seemingly requires sunny skies and fair weather. I find it fair to assume this is not the case.

These objections notwithstanding, the work of Santoni de Sio and van den Hoven provides a solid philosophical grounding for MHC, and one which supports more practical views presented by states, NGOs, and other groups. And it is also surely the case that no matter how one defines MHC, the concept has legal, moral, and philosophical value, even if only to serve as a floor condition for permissible use of autonomous and AI-enabled systems;[50] regardless of one's views of AWS in particular, it seems plausible that all would agree that potentially lethal weapon systems should be tightly controlled by and connected to a human's agency. MHC also creates a space for clear design requirements,[51] doctrinal requirements,[52] and discrete requirements in use,[53] all of which are good things to promote. However, for MHC to do so, it must be a pragmatic forward-looking principle, and not simply a ban in disguise.

MHC: A ban in disguise?

Despite disagreements surrounding MHC, "[t]his 'intuitively appealing' principle is immensely popular",[54] finding voice in the works of AWS skeptics[55] and arms regulation groups,[56] and in the documents of state militaries[57] and AWS proponents.[58] Yet some treat MHC not as a condition for guiding development and potential use of autonomous systems, but rather as a silver bullet for justifying a ban of AWS. In particular, by focusing on a narrow conception of MHC as requiring direct control over "individual attacks", groups such as Article 36 and Human Rights Watch set a standard which would rule out use of off-the-loop systems. Moreover, their conception of MHC rules out many unobjectionable self-defensive AWS, such as anti-missile and active protection systems, both of which often do not require humans to be in- or on-the-loop.[59]

Returning to our definition of AWS from Sect. 1, clearly any view of MHC which requires humans to be involved during all parts of an engagement or during "individual attacks"[60] would amount to a ban, especially when MHC is argued to also be a necessary requirement for autonomous systems to be permissibly used (a condition accepted by many on both sides of the debate). This is also not lost for all critics, and groups like Human Rights Watch will sometimes state clearly that "[h]umans should exercise control over individual attacks, not simply overall operations" and "[o]nly a ban on fully autonomous weapons can effectively guarantee such meaningful control by humans".[61] What critics do not show appreciation for, however, is the fact that a requirement of direct human control over *individual attacks*[62] would render numerous existing autonomous

[47] Santoni de Sio and van den Hoven (2018), p. 9.

[48] Though, I would argue, they ought not be; responsible use will demand narrow geographical and temporal limits to deployment. More on this in the following section.

[49] That humans have such control is a crucial assumption, and rules out the use of systems which possess the capacity for *in situ* learning. Such systems would also raise a host of significant moral, legal, and operational/doctrinal challenges. See, e.g., Blanchard and Taddeo (2022), Haugh et al. (2018), Verbruggen (2022), McFarland and Assaad (2023).

[50] See Crootof (2016) for discussion of both a floor for understanding MHC and MHC as a floor in regulatory debates.

[51] Umbrello (2021a, 2021b), Cavalcante Siebert et al. (2023).

[52] Horowitz and Scharre (2015); Bode and Watts (2021).

[53] Roorda (2015), Ekelhof (2019).

[54] Crootof (2016), p. 53.

[55] Roff and Moyes (2016); Amoroso (2020); Schwarz (2021).

[56] Article 36 (2013b); Human Rights Watch (2016, 2018, 2020), Boulanin et al. (2020, 2021), Bo et al. (2022); Stewart and Hinds (2023).

[57] US Department of Defense (2023).

[58] Horowitz and Scharre (2015), Roorda (2015), Scharre and Sayler (2016), Friedrich (2024).

[59] To be sure, these groups also have explicitly called for the ban of AWS, but beyond such overt ban advocacy, they have also put forward principles or conceptions of principles which, while clothed in intuitive language, amount to bans of AWS. Some presentations of MHC fit this categorization.

[60] E.g., Article 36 (2013a, 2013b), Roff and Moyes (2016), Human Rights Watch (2016, 2016, 2018).

[61] Human Rights Watch (2016), p. 38.

[62] For useful discussion around the scope of AWS attacks, see Kwik (2024b), Wood (2025). I would also note that direct control need not

systems impermissible, such as the above-mentioned anti-missile and active protection systems, but also many other systems used for decades, as well as the host of cyber capabilities being developed for defensive purposes, and which by necessity must be able to act independently of human oversight to be effective. The standard of MHC thus cannot generally be one of contemporaneous input or oversight, but must allow for meaningful control to be established without the necessity of having humans in- or on-the-loop.[63] And beyond the realm of autonomous systems, a requirement of direct control over individual attacks could also cut against semi-autonomous and wholly deterministic systems, if those include capabilities of dispersing ordnance over an area or fragmenting a payload into multiple discrete warheads that have distinct targets. This is because each discrete weapon that strikes could be considered as an "individual attack",[64] and the humans deploying such area weapons cannot foresee exactly what conditions will obtain at the moment of payload dispersal. Understanding MHC as pertaining to "individual attacks" is thus untenable, a point which even critics of AWS are coming to appreciate.[65]

Yet none of this is to say that MHC should be abandoned or that it is not required; arguably all systems should be used only on condition that they are being meaningfully controlled by human operators and handlers. However, such control does not require contemporaneous input or oversight of attacks, and can indeed be maintained even for fully autonomous off-the-loop systems, a contention to which we now turn.

The realities of AWS deployment

The terminology of "meaningful human control" was born from discussions surrounding autonomous weapons, but the underlying idea of "[e]nsuring appropriate human control over weapons, such that they are used effectively and as intended, has been a critical task for militaries around the world since the invention of the bow and arrow".[66] By the same token, while one may argue that AWS present some new element in need of critical examination, that "[a]utonomy, by its very nature, entails programming machines to perform some tasks or functions that would ordinarily be performed by humans",[67] this also is nothing new; the development of advanced stabilizers for artillery removed some of the need for humans to carefully recalibrate guns, GPS removed the need for bombers to manually sight in on their targets using maps, compasses, and visual cues, and a host of other sensing and information processing technologies have, bit by bit, removed the need for humans to carry out numerous tasks formerly done by us.

It has also always been the case that standard operating procedures, constraints on deployment, targeting parameters, and other factors inform and entrench complex and sometimes distributed forms of control over weapon systems.[68] Those rules and procedures also impact on how narrowly systems are being controlled, how broad the set of potential outcomes is when using (autonomous) systems, and how closely the action-outputs of systems follow the aims and intent of operators or handlers. Following Sean Welsh, we may thus distinguish between the "policy loop"—"the definition of the policy rules that the autonomous weapon mechanically follows"—and the "firing loop"—"the execution of those rules when firing".[69] As will be argued for through the rest of this article, what matters for MHC is this policy loop, the setting of rules which AWS follow.

Operational constraints

AWS, like all weapons, are subject to limitations. But limitations do not imply that using certain systems is impermissible, only that permissible use requires that deployers know the limitations of systems and respond accordingly. Furthermore, in addition to systems' own inbuilt limitations, operators may actively limit systems in a variety of ways. This provides the first and foremost means for one to circumscribe the possible outcome set AWS may bring about, narrowing the action-outputs realizable and making systems track ever closer to human deployers' aims and intent. Such constraining of systems is most straightforwardly achieved by simply altering the hardware or loadout of AWS, which can by necessity set hard limits on lethality, range, endurance, and other factors; an AWS equipped with a single grenade cannot bring down a multi-story building, and an AWS with a range of five miles will not attack a city fifty miles away. And if its sensors do not include cameras in the visual spectrum, it cannot discriminate on the basis of race, simply because it will not be capable of doing this. Similarly, if an AWS is programmed to deactivate itself an hour after launch, we may assume it will do so (absent some

imply that such control is meaningful.

[63] See Crootof (2016), p. 56, and citations therein.

[64] Note this view put forward by critics goes against international law as well, as the technical term of "attack" can refer to larger concerted efforts and need not be limited to each individual strike. See International Committee of the Red Cross (1987), pp. 601–603, especially paragraph 1880; Kwik (2024b).

[65] E.g., Sauer (2024).

[66] Horowitz and Scharre (2015), p. 5.

[67] Horowitz and Scharre (2015), p. 5.

[68] Ekelhof (2019).

[69] Welsh (2016).

mechanical malfunction). All of these factors serve as fundamental means for limiting a system and ensuring it is less capable of executing its processes in unwanted ways.

Beyond such brute methods for limiting systems, setting general operational constraints on deployments provides an additional means for one to further narrow the possible outcome space achievable by systems without fundamentally limiting what systems can do. This allows deployers to make AWS more tightly follow the human's aims and intent by simply ruling out large sets of variables humans cannot adequately control. For example, the most basic operational constraints which may be set are geographic and temporal ones; by setting a (fairly) precise time and place of attack, deployers can limit what might possibly be struck, in virtue of choosing times/places of attack which are known to (not) have certain types of persons or objects (e.g., military vs. civilian ones).

Furthermore, operational constraints allow one to respond to limitations of AWS in thoughtful and responsible ways, without demanding that one simply forgo a militarily advantageous mission because an AWS suffers from some deficit which would make its use impermissible in certain cases. For example, consider an anti-tank AWS which can reliably determine whether potential targets match the profile of tanks and have powerful engines (in virtue of millimetric wave radar, acoustic, and seismic readings).[70] Such a system may be able to reliably distinguish between tanks and cars, but not between tanks and heavy construction or farming machinery. Thus, using these in agricultural areas may be too risky, in terms of the likelihood that civilians will be mistakenly struck. However, knowing this limitation and wishing to narrow the outcome space associated with potential deployments, deployers may simply limit the operating environment of the AWS to exclude agricultural areas where heavy farming vehicles are expected, or such AWS might be deployed only at times when farmers are extremely unlikely to be out on their vehicles but when enemy tanks would still be in the field (i.e., at night). Simply put, limitations of a system are apt to demand limitations on operations—i.e., operational constraints—but limited systems may still find a host of acceptable use cases, provided deployers are acting in good faith and responding to limitations with sensible precautions.

In sum, by choosing the timing and location of attacks, and the spatio-temporal boundaries of AWS' operations, operators and handlers can control the moral complexity of operational environments, matching these to AWS' capabilities *and* limitations. Moreover, this method of control is not unique to AWS, as it pertains to all kinds of weapons, from grenades to air-to-air missiles; a grenade has no in-built capacity for target discrimination, so a soldier throwing a grenade into a building must consider in advance whether it is morally acceptable to use grenades there. Similarly, an AWS handler who knows that an anti-tank AWS can differentiate between tanks and cars, but not between, say, tanks and civilian bulldozers, has to decide whether deploying it in a specific area is acceptable. The AWS will do well in bulldozer-scarce environments, but its use near an active construction site will probably be unacceptable.[71] It is thus the responsibility of deployers to set constraints on operations which ensure that an AWS' action-outputs maximally comport with the aims and intent of those deployers, and such constraints, if thoughtfully implemented, can shear off whole sets of possible outcomes by literally keeping the system away from non-targets, either by deploying at a distance from protected persons/objects or deploying at times when those persons/objects are not expected to be present at the target location.

Targeting parameters

Once broad operational constraints have been set, there will be many additional means for deployers to constrain the possible action-outputs realizable by AWS. Most importantly, deployers and handlers will determine and set what parameters are used by the AWS for making target selections.[72] Thus, while many critics lament machines "deciding who lives and dies", the reality is much less dystopian, and should be understood not in terms of machines "making decisions" about who or what to target, but rather in terms of machines executing processes that lead to engagement of one or another type of object. Those process executions and subsequent engagements, however, will follow from the targeting parameters and engagement protocols set by humans deploying AWS.

In keeping with the limitations of a particular system, deployers ought first and foremost to set target parameters that feed to the strengths of an AWS and offset or mitigate its weaknesses. Thus, going back to the example of an anti-tank AWS discussed immediately above, if the AWS uses millimetric wave radar, acoustic signatures, and seismic readings, then that AWS could only plausibly engage armored vehicles or civilian heavy construction or farming equipment. If deployers have set operational constraints (geographic and/or temporal) which make it unlikely or

[70] This is how, respectively, the Brimstone missile and Brilliant Anti-Tank submunition (BAT) function.

[71] Similar arguments are made in Wood (2025), focusing there on AWS and the principle of distinction. See also (Heller, 2023), especially pp. 49–57. Useful discussions of proportionality and AWS can be found in Homayounnejad (2018), Ch. 7; Zając (2023); Cooper (2024).

[72] The context-dependent nature of MHC and its sensitivity to system limitations and parameters is also explored in, e.g., Amoroso and Tamburrini (2021), Kwik (2024a), Trabucco (2025b, 2025c).

implausible for such civilian vehicles to be present at the time of the mission, the target parameters could thus simply be the positive identification of objects which match some profile and are loud and heavy. Such parameters play off the strengths of the multi-modal sensor apparatus of the imagined AWS, and the operational constraints put in place offset the weaknesses and potential for targeting mistakes. If we were imagining an AWS that selected targets based on object identification from visual sensors though, then deployers would have to contend with the environmental factors at play during the planned time of engagement; does one expect there to be clouds, fog, smoke, or other obstructing factors that may undermine the possibility of reliably classifying objects?

Beyond this, while any single program trying to identify targets (and protected persons/objects) based on single-spectrum sensor information (say, video footage in the visible light spectrum) may be unreliable, reliability increases when the analysis is based on multi-spectral sensory data.[73] Similarly, multiple different programs may be tasked with concurrent object identification, with each sensor apparatus and connected program having potential override capability for engagements in the case of failures to gain positive ID of a target. We can also imagine cases where deployers may sensibly toggle whole sensor arrays on or off in response to environmental factors. For example, for tanks traveling through soft terrain, there may be little to no seismic signature attending their movement, and so there will be little to no possibility to get "positive ID" based on seismic sensors. Requiring that *all* sensors gain positive ID could thus be prohibitive in such environments, giving deployers grounds for toggling off the seismic sensors. There may be other environments where one reasonably requires more sensors to be online and gaining positive ID, say because there is a greater diversity of objects which are expected to be encountered. Thus, an anti-tank AWS which can use millimetric wave radar, acoustic sensors, seismic sensors, *and* visual sensors may allow for that AWS to be more readily used in environments with mixes of heavy military and civilian vehicles (e.g., tanks and bulldozers). By incorporating an additional sensor and feeding that into the AWS' targeting parameters, deployers can be more certain of what will (not) be targeted by the AWS, allowing deployers to more narrowly circumscribe the action-outputs of the systems.

Even more specifically, deployers may require that AWS identify very specific markings, visual cues, or other data points before selecting targets. Lists of "no-strike" markings or identifiers may also be compiled and uploaded to AWS, further allowing for refinement of the target selection process. None of this would place a human in- or on-the-loop, but it would allow humans to make off-the-loop systems far more predictable and far less likely to behave in novel ways, even supposing these were outfitted with some opaque AI systems for certain processes. And the more these parameters are fleshed out, made mission-, operation-, and theater-specific, and tailored to the exact limitations of an AWS and the exact needs of a mission, the more deployers can constrain the processes AWS will execute. All of this narrows the set of possible action-outputs, in turn making autonomous systems more and more a clear extension of a combatant's will.

Put simply, control over weapons does not have to be direct or contemporaneous to be meaningful; if it was so, planting anti-tank mines, launching fire-and-forget missiles, or even firing long-range artillery would be inherently morally problematic, which clearly is not the case. When proper precautions are taken, one can be reasonably sure using these weapons can be compliant with the ethics and laws of war. As the spatio-temporal distance between weapons' release and the hitting of targets increases, precautions will likely need to be more and more extensive, and the same is true with AWS, a new generation of stand-off weapons. But nothing about distance, be it physical, temporal, emotional, or otherwise will inherently undermine one's ability to exercise meaningful control.

Autonomy, actions, and will

So long as operators and handlers are competent in the use of systems, knowing the "dos and don'ts" implied by systems' limitations and responding to these accordingly, they are exercising meaningful human control over those systems. And the more aspects of autonomous systems' target selection a human can impact on and foresee, the more tightly a human can be in control of systems' action-outputs, even when not directly overseeing the systems during operation. Thus, technical limitations of systems, operational constraints on deployment, and targeting parameters set by operators all allow control to be made more precise and the possible set of action-outputs from AWS to be further narrowed. And by including multi-modal sensor apparatuses in AWS and feeding the inputs of these through target parameter selections, human deployers can make AWS increasingly predictable in their operation, even though AWS will still be making ultimate target selections and engagements on their own, a point even critics are beginning to accept.[74] Meaningful control thus does not demand contemporaneous oversight, and the critical point is not whether humans have

[73] Sauer (2024).

[74] E.g, the discussion of "ethical mines" in Sauer (2024). Sauer's discussion, while limited to anti-tank mines, closely follows earlier arguments made by AWS proponents such as (Wood, 2022; Zając, 2022).

direct control over the applications of force, but whether applications of force flow from the choices of humans in a meaningful way, where humans can meaningfully and foreseeably alter subsequent applications of force. Recalling Sean Welsh's terminology, the question is whether humans are in the policy loop, not whether they are in the fire loop.[75] And considering the many ways a human can exercise control over a system and how that control may be deemed meaningful, we arrive at a distinction which is worth exploring, namely the distinction between what I call "will-extending systems" versus "will-offloading systems".[76][77]

Will-extending systems

Machines do not have will and machines do not carry out "actions". Rather, they execute processes, sometimes in a fully deterministic way, sometimes with probabilistic processes involved which lead to a stochastic spread of potential action-outputs for a given input. But at no point does a machine "decide" or "choose" some outcome. It receives input and gives output.

Given the input-output nature of machines, AI included, the critical question when discussing meaningful control of machines is whether the humans deploying, operating, or handling them understand the machines well enough to responsibly and reliably use them for a given purpose. Put differently, humans must be able to use machines to extend their own wills and intent. If this is so, and an AWS, AI-enabled system, or other machine can be seen as a will-extending system for the human making use of it, then that system will be under MHC for that user. Beyond this, the more the human is able to *constrain* the potential action-outputs of the autonomous system, the more tightly can they extend their will (i.e., they can extend their will to a more precise outcome).

Going back to the discussions of Section 3, there are many ways humans can constrain the possible set of action-outputs for AWS; by choosing a system with limited hardware, giving operational constraints, and setting narrow targeting parameters, certain outcomes can be expressly ruled out, unpredictable factors can be eliminated, and possible outcome spaces can be circumscribed. The setting of targeting parameters presents the one element where risks may be unavoidable; depending on how the AWS selects targets, i.e., what sensor apparatuses are used, it may be more or less likely that an illegitimate target could be struck. By choosing what sensors are used though – either in virtue of being able to toggle on/off installed units – or being able to choose how the targeting processes handle sensor data – e.g., by setting confidence thresholds that must be met before targeting, setting priorities on certain sensor data over others, giving pre-set target lists, etc. – humans may make the set of possible action-outputs very narrow. That narrow set may still include outcomes where civilians (or civilian objects) are struck, but that is the case for all weapons; combatants may do their best (and ought to), but circumstances can change, even in the space of time between firing a bullet and it hitting its intended target. The simple fact is that we can never have full control over outcomes, only potentially over actions and processes that we set in motion.

The question is thus not whether the machine is "autonomously deciding" which targets to engage (it can't "decide" anything), but rather whether it is extending a user's will. As an example, if a combatant wants to destroy every armored vehicle in grid square D-14 and deploys an AWS that can carry out that task, then every armored vehicle targeted is targeted as a result of that willing, and in accordance with it. If we suppose that the AWS deployed cannot distinguish well between tanks and bulldozers and the combatant deploys it anyway, pursuing some precautions to reduce the likelihood of mis-targeting (e.g., deploying at night) but still knowing that this could occur, then this is still under MHC. And even assuming that a targeting mistake is made and a bulldozer fired upon, this also does not undermine the attribution of MHC, as the combatant, knowing that a bulldozer could be engaged, extended his or her will intending to destroy tanks, but knowing that it could result in civilian deaths. It is analogically the same as when a grenadier throws a grenade into a building from which enemies are firing but where the grenadier cannot reliably fire back; the grenadier knows there are legitimate targets there, using the grenade is the best (perhaps only) method for eliminating those enemies, eliminating the enemies is significant enough to merit the imposition of some risk to civilians, and knowing these things the grenadier throws the grenade, willing the outcome that the enemies and only the enemies are struck. But the grenadier knows that there is some likelihood, however small, that civilians might also be in the building, and that knowledge implies there is some likelihood of a mistaken strike against those civilians. That mistaken strike does not imply that the grenadier has no control over their weapon though. The weapon is meaningfully

[75] Indeed, even some of AWS' staunchest critics clearly state that "MHC is not, and has never, been an argument for direct physical control over weapons systems", but is rather "about *processes* and *rules* created, instituted, and governed by humans" (Roff (2024), emphasis in original).

[76] Due to space constraints, the presentation of the will-extending vs. will-offloading distinction here represents only a brief first look. Full exploration of the distinction and its implication remains as further work to be done.

[77] This also links to the debates on centaur vs. minotaur couplings of humans with machines. See respectively, e.g., Macak (2005), Scharre (2016); Neads et al. (2021); Johnson (2023) and Sparrow and Henschke (2023), Hasselberger (2024), Lushenko and Sparrow (2024).

controlled and is directed against a legitimate enemy, and bad luck has it that a certain randomness in how this weapon causes harm (i.e., an expanding explosive blast) means that innocents may be caught up in the harm.[78] But in knowing how the weapon ought to be used and in using it to extend their will, the grenade remains firmly under MHC, despite any probabilistic spread in ensuing outcomes.

It is worth reiterating that humans merely being the ones deploying systems does not amount to MHC; competence is key. In order for a machine to extend one's will, one must know enough about how to properly use it for it to extend willing (rather than potentially undermining one's will). But so long as one knows how to use a system well enough that one can reasonably foresee the outcomes it will bring about and knows how to set constraints on the system or its deployment which would narrow the outcome set, then it is under MHC. Its use may still lead to unwanted outcomes, but choosing a tool which brings with it some risk does not imply that one does not control that tool, only that one does not control outcomes. To argue otherwise would require arguing that humans lack control any time they take any actions which include risk of unwanted outcomes outside their direct control. But virtually all actions are of that type, and so the argument is untenable.

Will-offloading systems

In addition to will-extending systems, there are also systems currently in development and increasingly in use which, rather than allowing a combatant to decide on some course of action and then empower themselves to achieve it more fully, will instead undermine combatants' agency, leading them to offload their will to autonomous AI-enabled systems. One prominent example of such systems are AI-enabled decision support systems (AI-DSS).

Generally speaking, AI-DSS are a response to the massive increases in data and intelligence that can now be gathered, and the inability of states (or other organizations) to adequately assess that trove of information.[79] This fact, coupled with a basic intuition that more information is better, demands that AI-enabled systems be developed which can aid in parsing, labeling, and evaluating this wealth of data. Yet the outputs of AI-DSS do not necessarily empower agents to better extend their will in the world (though they may do this). Rather, AI-DSS, by their design and in virtue of how humans think (or don't) often lead individuals to offload their will onto them.

The core problem is that by simply deferring to a machine, a human stops acting as an agent, as they are no longer making meaningful moral choices. Their choices may have grave moral import, but the humans themselves are no longer acting as moral beings. Rather, they are cogs in a machine, meat-robots pressing buttons. And the AI-DSS contributes to this by purporting to impart more informed assessments—in virtue of having accounted for more information—which justifies to the human advisee that the system's recommendation is to be treated seriously. Through this, decision support systems can lead to grave mistakes, and to a general undermining of human will and intent in morally and legally fraught domains, which itself can cause just as much or more harm than autonomous systems may.[80] And this may be the case even when AI-DSS themselves can have no impact in the world, when their recommendations are just that, and only that, recommendations.

Thus, decision support systems, while not themselves causing any negative outcomes in the world (or any outcomes at all, for that matter) can create greater moral and legal hazards, for the simple reason that they create opportunities for human agents to offload their will to machines.[81]

Now, one may object that those designing and deploying systems such as AI-DSS are doing qualitatively the same thing as those deploying AWS; an intelligence officer or higher-ranking decision-maker determines that they want some outcomes to be better realized (in terms of strikes carried out, targets engaged, etc.), they set some basic parameters for targets being nominated by the system, and they then task their lower-ranked analysts or soldiers to act on the recommendations of the parameterized (albeit perhaps poorly) AI-DSS. However, there is a difference in kind between using a will-offloading AI-DSS and a will-extending AWS.

For the user of AWS, they must, in advance, consider the full context of the engagement for which they wish to use autonomous systems; what targets should be engaged, what potential bystanders may be present which ought *not* be engaged, what operational/targeting constraints most tightly limit AWS while allowing for effective task execution, etc. Deployers then deploy the system, knowing full well that things may go wrong. Moreover, deployers deploy AWS knowing the systems will be out of their hands, that they have up until the moment of deployment (or moment of "no return") to alter the parameters or deployment. But genuine decision-making is squarely and obviously with the human planning the engagement.

For decision support systems aiding individuals, however, the decisions are given by the AI-enabled system *and*

[78] Similar arguments are developed at length in Wood (2025).

[79] See, e.g., Walrath (1989), Klonowska (2020), Dorsch and Moll (2024), Nadibaidze et al. (2024), Reynolds (2025).

[80] See, e.g., Stewart and Hinds (2023). For recent recommendations on development and deployment of AWS and AI-DSS, see Blanchard and Bruun (2025).

[81] Cf. Fritz (2024) for a defense of deference to highly reliable AI-DSS.

only then do humans come into the picture. And the biases of humans (especially automation bias and confirmation bias)[82] indicate just how much will is lost simply in virtue of the order of communication; by allowing machines to give recommendations first, decision-spaces become tainted, with machine-derived outputs likely creating significant framing effects for all subsequent deliberations humans may undertake.[83] More than this, humans will often see AI-DSS as better judges than they themselves are, and indeed, some argue that humans *should* sometimes defer to machines, on precisely these grounds.[84] And once humans have more faith in systems than they do in themselves, they are apt to increasingly defer to machines' recommendations. At that point, AI-DSS (or any other systems treated in this way) become will-offloading systems. And critically, they are will-offloading systems even when all strike choices are ultimately made by human agents; if humans take their guidance from machines, subverting their own will and deferring to the recommendations of AI-enabled systems, humans are not meaningfully controlling anything. They have become minotaur.

This is not to say that all AI-DSS are will-offloading systems, and design and training criteria can (and will) greatly affect how much an AI-DSS allows a combatant to either be a better judge of information or a cog in a machine. However, AI-DSS represent an archetypal form of will-offloading system, as they are designed precisely to aid humans in some of their decision-making, but this is nothing more than saying that they allow humans to offload some element of assessment or judgment, be it of information, reasoning over the connections between data, or connection of data with other facts or aims already given.

Conclusion

Humans can control all manner of technical artifacts without necessarily having to exercise contemporaneous oversight or override capabilities throughout the deployment of those artifacts. From thermostats to autonomous weapons, the core criterion is that in using automated and autonomous systems, these serve to extend our will rather than offloading it. And when these function as will-extending systems, we retain clear lines of not just control, but also responsibility, as human agents whose will is being so extended clearly and naturally remain the authors of all outcomes brought about through the use of autonomous systems. A re-centering of will and agency in the debate thus clarifies that, far from AWS "acting" in any way or "choosing" or "deciding" who lives and dies, these are instead merely executing processes and delivering action-outputs that are devoid of agency. This is not to say that such systems may not introduce some level of unpredictability, especially when these incorporate increasingly opaque AI systems, but that unpredictability does not imply a lack of control so much as a potential failure of distinction in deploying such systems (in virtue of it becoming less possible to limit the effects of using such a means of combat, as required by international humanitarian law).[85] However, humans making use of autonomous systems, even potentially unpredictable ones, still control those systems. As an analogy, when humans use fragmentation weapons, they can control the deployment of those weapons, where, when, and under what precautions they are used, but the precise scattering of fragments will be somewhat unpredictable to them. That unpredictability, and the potential risks it raises, does not however undermine the core element of control grenadiers or bombers have over their weapons. Such weapons allow warfighters to act in more profound ways, causing more destruction and more greatly impacting enemy forces than a rifleman or gun-based aerial strike craft likely could, but the added unpredictability brought forward by fragmenting ordnance does not impugn such soldiers' control.

I should also clarify that though I have argued that meaningful human control can be retained for off-the-loop fully autonomous weapon systems, this should not necessarily be taken as a positive argument in favor of such systems. Moreover, there are likely to be systems that ought not be deployed at all, such as those which allow for *in situ* or adaptive learning in the field.[86] Aside from this, there are numerous ethical and legal principles that may argue against the use of certain types of AWS or against the use of AWS in certain combat scenarios, and it will certainly be the case that circumstances will be the surest arbiter of permissibility across a wide range of cases. In any event, the arguments show that such permissibility will not be undermined by a "lack of control", as AWS can very much be subjected to MHC, even when operating completely without human oversight. The core question will not be whether humans have direct fire control, but rather whether humans have exercised sufficient policy control to be able to competently assert that AWS are serving as extensions of their will, as

[82] See, e.g., respectively, Cummings (2017) and Nickerson (1998). See Arnott (2006), Phillips-Wren et al. (2019), Phillips-Wren and Adya (2020) for a more general discussion of the various biases potentially attending use of AI-DSS.

[83] See Tversky and Kahneman (1981), Druckman (2001), Sher and McKenzie (2008) for further discussion of framing effects. For recent exploration of a potential design response to these challenges, see Veluwenkamp and Buijsman (2025).

[84] E.g., Fritz (2024), Ross (2024).

[85] See especially AP I, Art. 51.

[86] See footnote 49 above.

systems meaningfully controlled by them. If so, critics' concerns surrounding MHC may be defused.

It is also worth reiterating that while the discussion here has focused on the military domain, the arguments developed relate to autonomous and AI-enabled systems more generally; "deployment parameters" calls to mind a militaristic mode of thinking, but doctors also set parameters and limitations on systems deployed in the medical domain. Likewise, insurance assessors using AI-enabled tools to evaluate claims or applicants will determine how, when, and with what restrictions such tools may be deployed. The same holds for educators bringing AI into the classroom, artists bringing AI into the studio, or anyone else making use of autonomous or AI-enabled systems. Humans making use of technical artifacts, even artifacts with some degree of potential autonomy, determine how, where, when, and under what limitations those systems may operate. And in holding that control, humans deploying these systems have the opportunity to lay out circumstances of use that mitigate toward these systems either serving as extensions and empowerments of humans' will, or as replacements of it.

Author Contributions Nathan Wood is the sole author of this work.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

Abbink, D., Amoroso, D., Cavalcante Siebert, L., van den Hoven, J., Mecacci, G., & Santoni de Sio, F. (2024). Introduction to meaningful human control of artificially intelligent systems. In Mecacci, G., Amoroso, D., Siebert, L. C., Abbink, D., van den Hoven, J., and Santoni de Sio, F., editors, *Research Handbook on Meaningful Human Control of Artificial Intelligence Systems*, 1–11. Edward Elgar Publishing.

Altmann, J., & Sauer, F. (2017). Autonomous weapon systems and strategic stability. *Survival, 59*(5), 117–142.

Amoroso, D. (2020). Autonomous weapons systems and international law.

Amoroso, D., & Tamburrini, G. (2020). Autonomous weapons systems and meaningful human control: Ethical and legal issues. *Current Robotics Reports, 1*(4), 187–194.

Amoroso, D., & Tamburrini, G. (2021). Autonomy in weapons systems and its meaningful human control: A differentiated and prudential approach. In G. Giacomello, F. N. Moro, & M. Valigi (Eds.), *Technology and International Relations.* Edward Elgar Publishing.

Arnott, D. (2006). Cognitive biases and decision support systems development: A design science approach. *Information Systems Journal, 16*(1), 55–78.

Article 36. (2013). *Killer robots: UK government policy on fully autonomous weapons* (p. 36). Article: Technical report.

Article 36. (2013). *Structuring debate on autonomous weapons systems* (p. 36). Article: Technical report.

AutoPractices Project. (2025). *Map of practices: Governing AI technologies in military systems from the bottom up*. Technical report.

Bächle, T. C., & Bareis, J. (2022). Autonomous weapons'' as a geopolitical signifier in a national power play: Analysing AI imaginaries in Chinese and US military policies. *European Journal of Futures Research, 10*(1), 1–18.

Bailey, R. (2025). Killer robots beyond the loop: Autonomy, UAS, and meaningful human control. Master's thesis. https://www.cfc.forces.gc.ca/259/290/351/286/BaileyMDS.pdf.

Beck, S., Gerndt, S., Samhammer, D., & Dabrock, P. (2024). Meaningful human control in shared medical decision making. In Mecacci, G., Amoroso, D., Siebert, L. C., Abbink, D., van den Hoven, J., and Santoni de Sio, F., editors, *Research Handbook on Meaningful Human Control of Artificial Intelligence Systems*, 131–147. Edward Elgar Publishing.

Blanchard, A., & Bruun, L. (2025). *Autonomous weapon systems and AI-enabled decision support systems in military targeting: A comparison and recommended policy responses*. Technical report.

Blanchard, A. & Taddeo, M. (2022). Predictability, distinction & due care in the use of lethal autonomous weapon systems. *SSRN Electronic Journal*.

Bo, M., Bruun, L., & Boulanin, V. (2022). *Retaining human responsibility in the development and use of autonomous weapon systems: On accountability for violations of international humanitarian law involving AWS*. Technical report.

Bode, I. (2023). Practice-based and public-deliberative normativity: Retaining human control over the use of force. *European Journal of International Relations, 29*(4), 990–1016.

Bode, I. (2025). Emerging norms around military applications of AI: The case of human control. Technical report.

Bode, I., & Watts, T. F. (2021). *Meaning-less human control: Lessons from air defence systems on meaningful human control for the debate on AWS*. Technical report.

Boothby, W. H. (2016). *Weapons and the Law of Armed Conflict* (2 edition). Oxford University Press.

Boshuijzen-van Burken, C., de Vries, M., Allen, J., Spruit, S., Mouter, N., & Munyasya, A. (2024). Autonomous military systems beyond human control: Putting an empirical perspective on value trade-offs for autonomous systems design in the military. *AI & Society, 40*(4), 2507–2523.

Boulanin, V., Bruun, L., & Goussac, N. (2021). *Autounomous weapon systems and international humanitarian law: Identifying limits and the required type and degree of human-machine interaction*. Technical report.

Boulanin, V., Davison, N., Goussac, N., & Carlsson, M. P. (2020). Limits on autonomy in weapon systems: Identifying practical elements of human control. Technical report, Stockholm

International Peace Research Institute and International Committee of the Red Cross.

Boutin, B., & Woodcock, T. (2024). Aspects of realizing (meaningful) human control: A legal perspective. In R. Geiß & H. Lahmann (Eds.), *Research Handbook on Warfare and Artificial Intelligence,* 179–196. Edward Elgar Publishing.

Bryson, J. J. (2017). The meaning of the EPSRC principles of robotics. *Connection Science, 29*(2), 130–136.

Calvert, S., Johnsen, S., & George, A. (2024). Designing automated vehicle and traffic systems towards meaningful human control. In Mecacci, G., Amoroso, D., Siebert, L. C., Abbink, D., van den Hoven, J., and Santoni de Sio, F., editors, *Research Handbook on Meaningful Human Control of Artificial Intelligence Systems*, 162–187. Edward Elgar Publishing.

Caron, J.-F. (2020). Defining semi-autonomous, automated and autonomous weapon systems in order to understand their ethical challenges. *Digital War, 1*(1–3), 173–177.

Cavalcante Siebert, L., Lupetti, M. L., Aizenberg, E., Beckers, N., Zgonnikov, A., Veluwenkamp, H., Abbink, D., Giaccardi, E., Houben, G.-J., Jonker, C. M., van den Hoven, J., Forster, D., & Lagendijk, R. L. (2023). Meaningful human control: Actionable properties for AI system development. *AI and Ethics, 3*(1), 241–255.

Cooper, C. G. (2024). Ensuring lawful use of autonomous weapons: An operational perspective. *Journal of International Humanitarian Legal Studies, 1*, 1–30.

Crootof, R. (2016). A meaningful floor for meaningful human control. *Temple International & Comparative Law Journal, 30*(1), 53–62.

Cummings, M. L. (2017). Automation bias in intelligent time critical decision support systems. In *Decision Making In Aviation*, 289–294. Routledge.

Davidovic, J. (2023). On the purpose of meaningful human control of AI. *Frontiers in Big Data, 5*, 1–5.

Dorsch, J., & Moll, M. (2024). Explainable and human-grounded AI for decision support systems: The theory of epistemic quasi-partnerships. *arXiv*

Druckman, J. N. (2001). Evaluating framing effects. *Journal of Economic Psychology, 22*(1), 91–101.

Ekelhof, M. (2019). Moving beyond semantics on autonomous weapons: Meaningful human control in operation. *Global Policy, 10*(3), 343–348.

Ekelhof, M., & Paoli, G. P. (2020). *The human element in decisions about the use of force*. Technical report.

Evans, K. D., Robbins, S. A., & Bryson, J. J. (2023). Do we collaborate with what we design? *Topics in Cognitive Science*. https://doi.org/10.1111/tops.12682

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and Control: A theory of moral responsibility*. Cambridge University Press.

Friedrich Ebert Stiftung (2024). Disarmament in times of war: Interview with Ryan Gariepy. https://ny.fes.de/article/disarmament-in-times-of-crisis-interview-with-ryan-gariepy.html, Retrieved December 11, 2024.

Fritz, J. (2024). Deference to opaque systems and morally exemplary decisions. *AI & Society*, 1–13.

Global Commission on Responsible Artificial Intelligence in the Military Domain (2025). Responsible by design: Strategic guidance report on the risks, opportunities, and governance of artificial intelligence in the military domain. Technical report.

Group of Governmental Experts on Lethal Autonomous Weapons Systems. (2025). *GGE on LAWS: Rolling text, status date: 12 May 2025*. Technical report.

Hasselberger, W. (2024). Will algorithms win medals of honor? Artificial intelligence, human virtues, and the future of warfare. *Journal of Military Ethics*. https://doi.org/10.1080/15027570.2024.2437920

Haugh, B. A., Sparrow, D. A., & Tate, D. M. (2018). *The status of test, evaluation, verification, and validation (TEV &V) of autonomous systems*. Technical report.

Heikoop, D. D., Hagenzieker, M., Mecacci, G., Calvert, S., Santoni De Sio, F., & van Arem, B. (2019). Human behaviour with automated driving systems: A quantitative framework for meaningful human control. *Theoretical Issues in Ergonomics Science, 20*(6), 711–730.

Heller, K. J. (2023). The concept of "the human" in the critique of autonomous weapons. *Harvard National Security Journal, 15*(1), 1–76.

Hille, E. M., Hummel, P., & Braun, M. (2023). Meaningful human control over AI for health? a review. *Journal of Medical Ethics*.

Hillis, J. M. & Payne, K. (2024). Health AI needs meaningful human involvement: Lessons from war. *Nature Medicine*, 1–2.

Homayounnejad, M. (2018). *Lethal Autonomous Weapons Systems Under the Law of Armed Combat*. PhD thesis, Kings College London.

Horowitz, M. C., & Scharre, P. (2015). *Meaningful human control in weapons systems*. Technical report.

Human Rights Watch. (2016a). *Killer robots and the concept of meaningful human control*. Technical report.

Human Rights Watch. (2016b). *Making the case: The dangers of killer robots and the need for a preemptive ban*. Technical report.

Human Rights Watch. (2018). *Heed the call: A moral and legal imperative to ban killer robots*. Technical report.

Human Rights Watch. (2020). *New weapons, proven precedent: Elements of and modes for a treaty on killer robots*. Technical report.

International Committee of the Red Cross. (1987). *Commentary on the Additional Protocols*. Martinus Nijhoff Publishers.

International Committee of the Red Cross. (2014). *Autonomous weapons systems: Technical, military, legal and humanitarian aspects*. Technical report.

International Committee of the Red Cross. (2021). *ICRC position on autonomous weapons systems*. Technical report.

International Committee of the Red Cross (2025). Submission to the United Nations secretary-general on artificial intelligence in the military domain. Technical report, International Committee of the Red Cross. https://www.icrc.org/sites/default/files/2025-04/ICRC_Report_Submission_to_UNSG_on_AI_in_military_domain.pdf.

Johnson, J. (2023). *The AI Commander: Centaur Teaming, Command, and Ethical Dilemmas*. Oxford University Press.

Klonowska, K. (2020). Article 36: Review of AI decision-support systems and other emerging technologies of warfare. *Yearbook of International Humanitarian Law, 23*, 123–153.

Kwik, J. (2022). A practicable operationalisation of meaningful human control. *Laws, 11*(3), 43.

Kwik, J. (2024). Controlling AWS: A cyclical process. *Lawfully Using Autonomous Weapon Technologies* (pp. 27–47). Springer.

Kwik, J. (2024b). The scope of an autonomous attack. In *2024 16th International Conference on Cyber Conflict: Over the Horizon (CyCon)*, 191–206. IEEE.

Lushenko, P. & Sparrow, R. (2024). Artificial intelligence and U.S. military cadets' attitudes about future war. *Armed Forces & Society*.

Macak, C. A. (2005). *Centaurs for maneuver warfare: Human-machine collaboration and manned-unmanned teaming for the fifth-generation ground combat element*. Technical report.

Marauhn, T. (2018). Meaningful human control - and the politics of international law. In W. H. von Heinegg, R. Frau, & T. Singer (Eds.), *Dehumanization of Warfare,* pp. 207–218. Springer.

McCarthy, J. (1988). Mathematical logic in artificial intelligence. *Daedalus*, 297–311.

McFarland, T., & Assaad, Z. (2023). Legal reviews of in situ learning in autonomous weapons. *Ethics and Information Technology, 25*(1), 1–10.

Mecacci, G., Amoroso, D., Siebert, L. C., Abbink, D., van den Hoven, J., & Santoni de Sio, F., editors (2024). *Research Handbook on Meaningful Human Control of Artificial Intelligence Systems*. Edward Elgar Publishing.

Mecacci, G., & Santoni de Sio, F. (2019). Meaningful human control as reason-responsiveness: The case of dual-mode vehicles. *Ethics and Information Technology, 22*(2), 103–115.

Nadibaidze, A., Bode, I., & Zhang, Q. (2024). *AI in military decision support systems: A review of developments and debates*. Technical report.

Neads, A., Galbreath, D. J., & Farrell, T. (2021). *From tools to teammates: Human-machine teaming and the future of command and control in the Australian Army*. Technical report.

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM, 19*(3), 113–126.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175–220.

Pacholska, M. (2024). Autonomous weapons. In B. Brożek, O. Kanevskaia, & P. Pałka (Eds.), *Research Handbook on Law and Technology,* 392–407. Edward Elgar Publishing.

Paoli, G. P., Spazian, A., & Anand, A. (2021). *Table-top exercises on the human element and autonomous weapons systems: Summary report*. Technical report.

Phillips-Wren, G., & Adya, M. (2020). Decision making under stress: The role of information overload, time pressure, complexity, and uncertainty. *Journal of Decision Systems, 29*, 213–225.

Phillips-Wren, G., Power, D. J., & Mora, M. (2019). Cognitive bias, decision styles, and risk attitudes in decision making and DSS. *Journal of Decision Systems, 28*(2), 63–66.

Reynolds, I. J. (2025). Speed and war in US military thought: Mapping the conditions for AI–enabled decision-making. *Millennium: Journal of International Studies*. https://doi.org/10.1177/03058298251317205

Robbins, S. (2023). The many meanings of meaningful human control. *AI and Ethics, 4*(4), 1377–1388.

Roff, H. (2024). Magnifying human confusion: Meaningful human control and the ongoing debate on autonomous weapons. *The Rule of Law Post*. https://www.penncerl.org/the-rule-of-law-post/magnifying-human-confusion-meaningful-human-control-and-the-ongoing-debate-on-autonomous-weapons/.

Roff, H., & Moyes, R. (2016). *Meaningful human control, artificial intelligence and autonomous weapons* (p. 36). Technical report.

Roff, H. M. (2013). Killing in war: Responsibility, liability, and lethal autonomous robots. In F. Allhoff, N. G. Evans, & A. Henschke (Eds.), *Routledge Handbook of Ethics and War: Just War Theory in the* 21^{st} *Century,* 352–364. New York, NY: Routledge.

Roorda, M. (2015). NATO's targeting process: Ensuring human control over and lawful use of 'autonomous' weapons. In A. P. Williams & P. D. Scharre (Eds.), *Autonomous Systems: Issues for Defense Policymakers,* pp. 152–168. The Hague, Netherlands: NATO Communications and Information Agency.

Ross, A. (2024). AI and the expert: A blueprint for the ethical use of opaque AI. *AI & Society, 39*(3), 925–936.

Russell, S. J., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach* (fourth edition edition). Pearson.

de Santoni Sio, F., Mecacci, G., Calvert, S., Heikoop, D., Hagenzieker, M., & van Arem, B. (2023). Realising meaningful human control over automated driving systems: A multidisciplinary approach. *Minds and Machines, 33*(4), 587–611.

de Santoni Sio, F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI, 5*, 1–14.

Sauer, F. (2024). An ethical minefield? on counter-mobility and weapon autonomoy. *War on the Rocks*. https://warontherocks.com/2024/10/an-ethical-mine-field-on-counter-mobility-and-weapon-autonomy/.

Scharre, P. (2016). Centaur warfighting: The false choice of humans vs. automation. *Temple International and Comparative Law Journal, 30*, 151–165.

Scharre, P., & Sayler, K. (2016). *Autonomous weapons and human control*. Technical report.

Schwarz, E. (2021). Autonomous weapons systems, artificial intelligence, and the problem of meaningful human control. *Philosophical Journal of Conflict and Violence*. https://doi.org/10.22618/TP.PJCV.20215.1.139004

Sher, S., & McKenzie, C. R. (2008). Framing effects and rationality. In N. Chater & M. Oaksford (Eds.), *The Probabilistic Mind: Prospects for Bayesian cognitive science,* 79–96. Oxford, UK: Oxford University Press.

Sparrow, R. J., & Henschke, A. (2023). Minotaurs, not centaurs: The future of manned-unmanned teaming. *The US Army War College Quarterly: Parameters, 53*(1), 115–130.

Stewart, R. & Hinds, G. (2023). Algorithms of war: The use of artificial intelligence in decision making in armed conflict. *ICRC Humanitarian Law and Policy Blog*. Retrieved from October 26, 2023 from https://blogs.icrc.org/law-and-policy/2023/10/24/algorithms-of-war-use-of-artificial-intelligence-decision-making-armed-conflict/

Taddeo, M., & Blanchard, A. (2022). A comparative analysis of the definitions of autonomous weapons systems. *Science and Engineering Ethics, 28*(5), 1–22.

Trabucco, L. (2025). AI-enabled autonomous weapons and human control: Part I: Human control and machine learning design and development. *International Law Studies, 106*, 534–578.

Trabucco, L. (2025). AI-enabled autonomous weapons and human control: Part II: Human control and military commanders. *International Law Studies, 106*, 579–615.

Trabucco, L. (2025). AI-enabled autonomous weapons and human control: Part III: Human control and system operators. *International Law Studies, 106*, 616–649.

Tsamados, A., Floridi, L., & Taddeo, M. (2024). Human control of AI systems: From supervision to teaming. *AI and Ethics*. https://doi.org/10.1007/s43681-024-00489-4

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*(4481), 453–458.

Umbrello, S. (2021a). Coupling levels of abstraction in understanding meaningful human control of autonomous weapons: A two-tiered approach. *Ethics and Information Technology*. https://doi.org/10.1007/s10676-021-09588-w

Umbrello, S. (2021b). *Towards a Value Sensitive Design Framework for Attaining Meaningful Human Control over Autonomous Weapons Systems*. PhD thesis, Consortium Fino.

US Department of Defense (2023). DoD Directive 3000.09. Technical report, United States Department of Defense.

van Diggelen, J., Boshuijzen-van Burken, C., & Abbass, H. (2024). Team design patterns for meaningful human control in responsible military artificial intelligence. In B. Steffen (Ed.), *Bridging the Gap Between AI and Reality,* 40–54. Cham, Switzerland: Springer Nature.

van Diggelen, J., van den Bosch, K., Neerincx, M., & Steen, M. (2024b). Designing for meaningful human control in military human-machine teams. In Mecacci, G., Amoroso, D., Siebert, L. C., Abbink, D., van den Hoven, J., and Santoni de Sio, F., editors, *Research Handbook on Meaningful Human Control of Artificial Intelligence Systems*, 232–252. Edward Elgar Publishing.

Veluwenkamp, H. (2022). Reasons for meaningful human control. *Ethics and Information Technology*. https://doi.org/10.1007/s10676-022-09673-8

Veluwenkamp, H., & Buijsman, S. (2025). Design for operator contestability: Control over autonomous systems by introducing defeaters. *AI and Ethics, 5*(4), 3699–3711.

Verbruggen, M. (2022). No, not that verification: Challenges posed by testing, evaluation, validation and verification of artificial intelligence in weapon systems. In T. Reinhold & N. Schörnig (Eds.), *Armament, Arms Control and Artificial Intelligence, Studies in Peace and Security,* 175–192. Springer.

Walrath, J. D. (1989). *Aiding the decision maker: Perceptual and cognitive issues at the human-machine interface*. Technical report.

Wang, P. (2019). On defining artificial intelligence. *Journal of Artificial General Intelligence, 10*(2), 1–37.

Welsh, S. (2016). We need to keep humans in the loop when robots fight wars. *The Conversation*. https://theconversation.com/we-need-to-keep-humans-in-the-loop-when-robots-fight-wars-53641.

Williams, A. P. (2015). Defining autonomy in systems: Challenges and solutions. In A. P. Williams & P. D. Scharre (Eds.), *Autonomous Systems: Issues for Defense Policymakers*, 27–62. The Hague, Netherlands: NATO Communications and Information Agency.

Wood, N. G. (2022). Autonomous weapons systems and force short of war. *Journal of Ethics and Emerging Technologies, 32*(2), 1–16.

Wood, N. G. (2023). Autonomous weapon systems: A clarification. *Journal of Military Ethics, 22*(1), 18–32.

Wood, N. G. (2024). Explainable AI in the military domain. *Ethics and Information Technology, 26*(2), 1–13.

Wood, N.G. (2025). Bombs, bots, and the principle of distinction : The law of armed conflict and contemporary warfare. *Texas National Security Review 9*(1), 52–67.

Zając, M. (2022). *Autonomous Weapon Systems from a Just War Theory Perspective*. PhD thesis, University of Warsaw.

Zając, M. (2023). AWS compliance with the ethical principle of proportionality: Three possible solutions. *Ethics and Information Technology, 25*(1), 1–13.