Chapter 4
Warfare at AI Speed and the Extended Meaningful Human Control Problem

Daniele Amoroso, Diego Mauri, and Guglielmo Tamburrini

Abstract As militaries increasingly incorporate Artificial Intelligence (AI) into their decision-making architecture, legal and ethical debates have primarily focused on Autonomous Weapon Systems (AWS). This paper argues that equally urgent challenges concern AI-enabled Decision-Support Systems (DSS-AI), even though these systems formally preserve human authority over the use of force. We develop the concept of extended meaningful human control (MHC) to address normative risks posed by DSS-AI in operational environments where time pressure, cognitive constraints, and system opacity converge to marginalize human agency in the decision-making loop. Our argument draws on two scenarios in which the integration of AI in military decision-making is linked to the scarcity of temporal resources—whether inherent (as in nuclear early warning systems) or induced (as in the alleged use of DSS-AI by the Israeli Defense Forces in Gaza). Both cases illustrate how taking critical military decisions at AI speed raises profound ethical and legal concerns. To resist the erosion of MHC, we propose three implementation pathways: human-centered system design, clearly distributed responsibility within command structures, and training programs to counter cognitive and institutional sources of automation bias. These measures should be codified in a multilateral treaty or, failing that, embedded in national doctrines and military manuals to resist the systemic drift toward "warfare at AI speed".

D. Amoroso
University of Cagliari, Cagliari, Italy
e-mail: daniele.amoroso@unica.it

D. Mauri
University of Palermo, Palermo, Italy
e-mail: diego.mauri@unipa.it

G. Tamburrini (✉)
University of Naples Federico II, Naples, Italy
e-mail: guglielmo.tamburrini@unina.it

P. Cotta-Ramusino et al. (eds.), *Nuclear Weapon Risks in a Problematic Time*, Springer Proceedings in Physics 358, https://doi.org/10.1007/978-3-032-29873-7_4

4.1 Introduction

Military powers are rapidly incorporating Artificial Intelligence (AI) technologies into their defense systems. These AI-enabled developments include Autonomous Weapons Systems (AWS) and systems supporting multiple facets of military decision-making, which range from strategic planning to operational support of battlefield actions.

Decision-support systems enabled by AI technologies (DSS-AI from now on) crucially differ from AWS in the control roles that are assigned to human operators. AWS can select and then proceed to engage targets without requiring any human intervention after their activation. By contrast, DSS-AI provide advice and other relevant information to human decision-makers. The latter are assigned by design the power to accept, revise, or reject the content of these inputs. In view of their autonomy in target selection and engagement, AWS have raised ethically and legally significant human control issues [4, 11, 13]. DSS-AI have been found to raise human control issues of a different kind [21], whose insurgence is not prevented by the fact that DSS-AI, unlike AWS, are subject by design to human judgment and gating functions.

In this contribution, we focus on military decision-making involving DSS-AI at the operational level, identifying major sources of human control hurdles in the narrow temporal resources that are often available in those decision-making contexts and in limitations of our capability to understand, predict precisely, and explain the behavior of AI systems in general, and DSS-AI in particular. This general claim is explored in connection with two different scenarios.

The first scenario (examined in Sect. 4.2) involves the use of DSS-AI in nuclear early warning. A major motivation for this proposal is that sensory processing at AI speed can help one achieve situational awareness more rapidly, thereby making more time available for downstream human decision-making in response to nuclear attack warnings. It is argued, however, that this expectation fails to properly consider the extra time that one must spend to verify the correctness of DSS-AI outputs. This extra time may offset and even reduce the time available for downstream human decision-making.

The second scenario (examined in Sect. 4.3) concerns the alleged use of DSS-AI made by the Israeli Defense Forces (IDF) to identify bombing targets in Gaza. Unlike the nuclear early warning scenario, which is characterized by an inherent scarcity of temporal resources, a problem of induced scarcity of temporal resources is denounced to arise in this case—more specifically, scarcity induced by hierarchical pressure to adapt the pace of human decision-making to the speed of AI information processing. It is argued (Sect. 4.3.1) that this induced scarcity may trigger human cognitive biases, which jeopardize the effectiveness of human control on DSS-AI advice.

In Sect. 4.4, it is claimed that in both conditions of inherent and induced scarcity of temporal resources, limitations of present AI technologies and human cognitive biases converge to erode the meaningful human control (MHC) that one should exert

on military DSS-AI used in operational decision-making. And the erosion of MHC on DSS-AI is found to introduce new threats to the respect of international law, including International Humanitarian Law (IHL) and the law against war (*jus ad bellum* or *jus contra bellum*). While existing rules and principles of international law can already be construed as requesting MHC on DSS-AI, it is argued that States and other international actors must urgently countervail the ongoing erosion of MHC on DSS-AI, by sharing best practices based on human-centered control design and training, by introducing appropriate national and international regulation, and ultimately by checking the ongoing military race towards the objective of fighting wars at AI speed. Acting against the hunger for augmented speed in the battlespace is precisely what legal and ethical requirements currently impose. Section 4.5 concludes.

4.2 The Illusion of Buying Time: AI Speed and Human Control in Nuclear Early Warning

The major nuclear powers seem to converge on the position that decisions to use nuclear weapons must firmly remain in human hands. This is not only supported by self-evident political exigencies, but also inferable from the rules on the use of force in international relations and self-defense (*jus ad bellum*) and those of international humanitarian law (IHL, *jus in bello*), which severely restrict the threat or use of nuclear weapons. As the International Court of Justice famously held, indeed, the "use of nuclear weapons would generally be contrary to the rules of international law applicable in armed conflict", the only exception being their use "in an extreme circumstance of self-defence, in which the very survival of a State would be at stake" (ICJ, *Legality of the Threat or Use of Nuclear Weapons*, Advisory Opinion of 8 July 1996, Dispositif, para. 2D).[1]

However, nuclear states have not excluded several ways in which DSS-AI may contribute to nuclear command, control, and communication (NC3), including the identification of potential missile launches through sensor data processing, prelaunch activity detection, and the selection of actions to undertake in response to nuclear threats [22, 8]. In this section, we focus on the proposal to use DSS-AI for potential missile launch identification. The main rationale for this proposal to use DSS-AI in nuclear early warning is that AI processing is expected to reduce potential errors, and its speed is expected to buy more time for nuclear decision-makers within the narrow temporal windows that are available to them [19, 104, note 22]. These temporal windows are *inherently* narrow insofar as the time span typically elapsing from missile launch to impact is less than 30 min.

It is undoubtedly of paramount importance to make as much time as possible available to nuclear decision-makers within that inherently restricted temporal window; what is more, this should be 'quality' time, in which decision-makers may carefully

[1] It should be noted that the Court was deeply divided on this point, which was included in the Dispositif only thanks to the President's casting vote.

evaluate the real nature of the incoming attack and whether a nuclear response is due and in line with *jus ad bellum* requirements of immediacy, necessity and proportionality [20]. However, the relevant question is whether AI provides effective means to pursue this goal. Significant doubts about a positive answer to this question are fueled by reflecting on current limitations of AI technologies based on machine learning and on related human-AI interactions. Let us briefly recall here some of these limitations, which have played a central role already in the framework of ethical and legal debates about the human control of AWS [4]. These limitations notably include (i) the epistemic difficulty to go beyond statistical assessments of an AI system's accuracy and to predict its pointwise behaviors; (ii) AI errors that adversarial machine learning investigations have highlighted, and that the human cognitive system is unlikely to commit; (iii) the opacity of much AI information processing.

Consider item (i) under the assumption that a DSS-AI system was already tested and found to achieve highly accurate results in nuclear launch detection. These accuracy results do not exclude the occurrence of an infrequent mistake, such as a false positive detection of incoming nuclear missiles. Since the occurrence of such infrequent mistakes may trigger an unjustified use of nuclear weapons, the ensuing high risk for civilian populations and human civilizations demands systematic probing of the veracity of DSS-AI perceptual responses. But the time spent on this task may offset any temporal gain for human decision-makers that one may expect to flow from this use of DSS-AI in nuclear early warning.

Similar doubts are strengthened by considering item (ii), that is, the discovery of unexpected fragilities in AI information processing, leading to mistakes that AI systems may commit, and that the human perceptual system would unproblematically detect and avoid. These fragilities of AI information processing have been uncovered by means of adversarial machine learning techniques [7], thereby revealing perceptual situations in which human information processing is more robust than AI information processing. One cannot presently exclude that AI systems for nuclear early warning will make counterintuitive and potentially catastrophic errors of the same sort that adversarial machine learning highlighted in other critical application domains. Bringing these reflections to bear on international law, the issue emerges—which is admittedly little explored in doctrine—of the so-called "putative" self-defense, that is, the mistake of fact committed by a State that in good faith, but mistakenly, believes an armed attack is occurring against it, and thus reacts by resorting to force [15]. While the debate on whether genuine mistakes of fact may justify States' reactions concerns mostly cyberattacks, it seems to us that it affects also—and to a more troubling extent—DSS-AI in nuclear early warning, due to the catastrophic consequences that such mistakes of fact are likely to generate.

Let us finally consider item (iii). Arguably, assessing the veracity of DSS-AI responses may require human decision-makers to get some understanding of the reasons supporting those responses. But this task introduces additional temporal delays, amplified by the information processing opacity presently affecting DSS-AI and other AI systems based on machine learning and Deep Neural Networks (DNNs).

In conclusion, a reflection on limitations (i)–(iii) raises legitimate doubts on the prospects of leveraging AI processing speed to ease inherently scarce temporal

resources in nuclear decision-making, insofar as this expectation overlooks that extra time must be spent to verify the correctness of DSS-AI outputs in this high-risk military domain. Such 'extra' time stands not only as an ethical requirement (making 'correct' choices in a highly critical domain) but also as a key tool to ensure respect for *jus ad bellum* requirements.

4.3 Inducing Time Pressure Through AI: IDF Uses of DSS-AI in Gaza

In this section, we turn to consider the military drive to leverage AI speed from a different perspective: from the proposal of buying more time for human decision-makers with the help of AI speed, we now examine proposals to leverage AI speed to increase the pace of operational decisions concerning actions in conventional battlefields.

A suitable scenario that enables one to explore these proposals is afforded by news reports concerning DSS-AI systems, known as *Habsora* (Gospel), *Where's Daddy* and *Lavender*, employed by the Israeli Defense Forces (IDF) in the armed conflict in Palestinian territories. Israel has resorted heavily to those systems to identify bombing targets in Gaza and its surroundings, in the aftermath of the massacre of innocent civilians perpetrated by Hamas within Israeli territory on October 7, 2023.

According to investigations carried out by the + 972/Local Call magazines and published between 2023 and 2024 [1, 2], the IDF employed these DSS-AI to generate lists of potential bombing targets in the Gaza Strip. A team of specialists in the IDF targeting division was appointed with the task of reviewing downstream the items included in AI-generated lists of potential targets, eventually approving or rejecting AI recommendations based on a legitimacy check. However, undisclosed sources in the IDF targeting division stated that the automation of the target generation task increased dramatically the targeting division's productivity in the way of potential target generation, moving "from 50 targets a year to 100 targets a day". While this stunning increase does not per se run in contravention with existing international rules, it must be assessed whether this drive to augmented speed in the battlespace risks fueling behaviors that are not in line with legal and ethical requirements.

The increased machine productivity in target generation gave rise to a decision-making bottleneck. According to an anonymous source, human operators perceived psychological pressure to align their filtering decisions to the pace of machine targeting suggestions. Indeed, this source complains about the assembly line character imposed on the filtering task ("it really is like a factory"), about the lack of sufficiently extended time frames to overview machine suggestions and to provide considered judgments about legal compliance ("we work quickly and there is no time to delve deep into the target"). Related to this, the same anonymous source mentions the introduction, higher up in the military hierarchy, of new criteria to evaluate the performances of targeting division members, which prize increased output volumes

of approved targeting suggestions ("the view is that we are judged according to how many targets we manage to generate") [1]. Consistently with this view, target selection procedures were drastically simplified: "One source stated that human personnel often served only as a 'rubber stamp' for the machine's decisions, adding that, normally, they would personally devote only about '20s' to each target before authorizing a bombing—just to make sure the Lavender-marked target is male." [2].

In the scenario described in this investigative news report, military commanders expect human control to be exerted under more demanding temporal constraints, intimate their subordinates to adapt the pace of human decision-making to increased machine productivity, and provide incentives when the acceptance rate on DSS-AI suggestions goes up.

Let us now examine this scenario from the perspective of MHC and the related normative requirement of protecting MHC functions from situations in which human control boils down to little more than clerical rubber-stamping of machine suggestions.

The described circumstances suffice to contend that the practice of DSS-AI in the Israel-Palestine conflict fails to comply with the principle of precautions in attacks, one of the cornerstones of the law of armed conflict. What is more, it is known that those systems make what are regarded as "errors" in approximately 10% of cases, occasionally marking individuals who have merely a loose connection to militant groups, or no connection at all [2]. It follows that Israel accepts not only the hazard, but even the virtual certainty of engaging impermissible targets—a conduct that breaches existing rules on distinction and proportionality in the conduct of hostilities, giving rise to both international responsibility of the State and, to the extent that such conduct qualifies as war crimes such as attacks against the civilian population and excessive collateral damage under Art. 8(2)(b)(i) and (iv) of the ICC Statute, international criminal responsibility for individuals involved [5, 14].

4.3.1 Automation Bias and Related Human Decision-Making Heuristics

It was noted in Sect. 4.2 that in situations brought out by adversarial machine learning, human and AI perceptual processing show some complementary strengths and weaknesses. In some circumstances, the human perceptual system appears to be more robust than AI perceptual processing. This observation supports the claim that MHC must play the prospective responsibility role of a fail-safe agent, contributing to some extent to counteracting possible effects of failure anticipated by adversarial machine learning. However, the performance of this fail-safe function is likely to be jeopardized by the automation bias onset. The expression "automation bias" designates a propensity of human decision-makers to defer to recommendations advanced by DSS systems despite warning signals or retrievable information to the contrary.

In the above IDF scenario, the onset of the automation bias may be facilitated by the so-called "authority bias", that is, the tendency to attribute greater value to the opinion of an authority figure and to be more influenced by that opinion [16]. Here, the authority bias may invite subordinates to emulate behaviors of lenient attitudes towards DSS advice adopted by role models in the targeting division and incentivized by hierarchical superiors. The reluctance to question "algorithmic authority" may additionally stem from a different, albeit related phenomenon: a DSS-AI operator would indeed be strongly discouraged from overriding the system's recommendation by the fact that he or she "will be personally blamed if the override goes wrong" with likely negative career consequences, while the failure to reject a harmful recommendation leaves open the possibility "to blame the AI for any difficult decision with a negative outcome" [8, 317].

Moreover, automation bias was experimentally probed in aviation and other time-critical control domains [10]. Anecdotal evidence of automation bias in time-critical military situations of DSS use has been widely reported, including friendly fire episodes in the first Iraqi war induced by wrong DSS detection of approaching enemy fighter jets. A psychological mechanism that has been hypothesized to pave the way to automation bias in time-critical domains has to do with the dual nature of human decision-making.

Dual models of human decision-making identify a significant source of human judgment perturbation in the psychological pressure to meet tight deadlines under severe cognitive load [25]. In this regard, Daniel Kahneman introduced an admittedly simplifying distinction between two distinct cognitive "systems", respectively subsuming heuristic and more analytic decision-making [12]. System 1 is a blanket term collecting heuristic decision-making processes that tend to be faster, mostly automatic, default and often emotionally laden. System 2 embraces more reflective and analytical decision-making processes that are usually slower and less readily activated. The two systems operate concurrently, but not always cooperatively. Options that are more readily selected by System 1 are often accepted by default, and concurrent efforts by System 2 are disregarded, especially when there is limited time to decide what to do.

To illustrate, consider the so-called WYSIATI (What-You-See-Is-All-There-Is) heuristics, identified and experimentally probed by Daniel Kahneman and Amos Tversky. WYSIATI is a System 1 heuristics reflecting, according to Kahneman, a "remarkable asymmetry between the ways our mind treats information that is currently available and information we do not have". People tend to focus on what is clearly visible to them, neglecting information that is not known or not presently retrieved from memory: "Information that is not retrieved (even unconsciously) from memory—Kahneman observed—might as well not exist." One jumps to conclusions based on this asymmetry and goes on to construe "the best possible story that incorporates ideas currently activated", endorsing by default the (often but not invariably correct) belief that only this information is relevant [12, 85]. The WYSIATI heuristics induces people to focus on what is clearly visible to them and to treat as non-existent information that is not known or not presently retrieved from memory. WYSIATI may contribute to predict or explain tendencies to conform to DSS-AI advice. In

particular, the lists of targets generated by the DSS-AI allegedly used by the IDF in Gaza are readily visible to human controllers, whereas considerable cognitive effort might be needed on their part to carefully assess (and, in case, reject) them on rational grounds and to propose different courses of action. The anchoring heuristic strategy attributed to System 1 may further contribute to explain the tendency to conform to DSS-AI suggestions, insofar as an individual's decisions turn out to be influenced by an initial reference point or anchor. In this case, the anchor is provided by DSS-AI advice and the human controller's subsequent evaluations and conclusions tend to stay closer to the anchor than they would have been otherwise.

4.4 The Why and How of Extended MHC on AI-Enabled Warfare

A major source of what we call here the extended MHC problem is AI speed and its influence on human decision-making supported by DSS-AI. It has been suggested that both *inherent* and *induced* scarcity of temporal resources in operational decision-making involving DSS-AI significantly hinder the ability of human operators to intervene in the process.

Military deployment of these systems, in other terms, tends towards "a dramatically reduced decision-making cycle as the capabilities in this area improve" [18]. In this context, the operational tempo of decision-making is no longer dictated by human limits. It is rather adapted to the accelerated outputs of machines, with the human role increasingly being reactive rather than deliberative. Commanders who once had the luxury of measured deliberation will now be "forced to [make decisions] in seconds," particularly when facing adversaries leveraging AI to accelerate battlefield tempos [9].

The fear of lagging behind in the arms race toward hyper-accelerated decision-making is already influencing strategic doctrines among the world's leading military powers [23]. The competitive pressure to match or surpass the adversary's DSS-AI capabilities, as well as the ever-lasting doctrine of nuclear deterrence (now entering an unprecedented phase due to its combination with AI systems), lead to a global landscape situation where deliberation time itself becomes weaponized.

These developments point to a growing and pervasive normative challenge in contemporary military practice: the need to ensure meaningful human control in the use of DSS-AI. Unlike traditional discussions focused on autonomous weapon systems, this problem arises even when humans remain formally in the loop but operate under conditions that may undermine deliberation and oversight. Such extended MHC problem can be distinguished into two parts. The first concerns the *why*, that is, the extant motivations for requiring MHC on DSS-AI. The second part of the problem has to do with the *how* of such MHC, that is, with the ways military uses of DSS-AI must be harnessed to ensure MHC.

To begin with the *why*, one can draw on previous analyses of the ethical and legal motivations for requiring MHC in the use of Autonomous Weapon Systems (AWS). As suggested elsewhere, human agents are normatively required to play three key roles in weapon system operations: as fail-safe actors, accountability attractors, and moral agency enactors [4, 24].

Among these roles, the fail-safe function stands out as particularly crucial in the case of DSS-AI. It posits that human oversight is necessary to detect and correct failures that would otherwise go unaddressed by the system itself, especially when these failures stem from adversarial misclassifications, erroneous correlations, or incomprehensible black-box outputs.

One might argue that this role is redundant in the DSS-AI context, since a human is always present to take the final decision. Yet, as discussed in Sect. 4.3.1, the practical conditions of decision-making—marked by extreme time pressure, information overload, and hierarchical incentives—often preclude meaningful verification or dissent. What eventually results from the present analysis is a procedural façade of control, in which the presence of a human decision-maker cannot fulfill the underlying moral and legal function of serving as a check on system failure.

In this sense, the core concern articulated by Article36 with regard to AWS in its ground-breaking analyses on MHC—namely, the distinction between meaningful and perfunctory human control—resurfaces with full force in the DSS-AI domain [6]. If the human role does not include the capacity and conditions to act as a fail-safe, then the ethical and legal demands associated with human control remain unmet, even if a human technically signs off on each action.

Having outlined the *why* of MHC over DSS-AI, we now turn to the more challenging question of *how* such control might be practically implemented in AI-enabled military contexts. A valuable starting point in this regard is offered by the ICRC's 2021 position paper on autonomous weapon systems [11]. While the document does not explicitly use MHC terminology, it provides one of the more authoritative attempts to translate normative concerns for human control into concrete operational criteria. As such, it offers not only a regulatory benchmark for AWS, but also a conceptual scaffold for designing control mechanisms in the case of DSS-AI.

In particular, the ICRC articulates the following requirement for ensuring appropriate human oversight:

> (MHC_1) Human decision-makers must be in a position to assess whether AWS operation is constrained – by design or situation of use – to objects that are legitimate military targets, and to intervene as required. This presupposes limits on duration, scope, and scale of use, and requires provisions for proper human supervision, judgement, timely intervention, or deactivation.

Although this condition is formulated in relation to systems that autonomously apply force, its underlying structure is directly relevant to the challenges posed by DSS-AI. Here, the primary risk is not kinetic action, but epistemic distortion—namely, the influence of machine-generated inferences on human situational awareness and decision-making. Extending the ICRC's approach to the DSS-AI domain, one may envisage a parallel requirement:

(MHC_2) Human decision-makers must be in a position to assess the veracity of DSS-AI perceptual inferences, to form pondered and well-informed judgments about their recommendations, to evaluate the likely real-world effects of actions undertaken on the basis of such inferences and recommendations, and to filter DSS-AI information and advice.

Translating the condition expressed in (MHC_2) into concrete operational measures requires addressing the interplay of cognitive, institutional, and procedural factors that determine whether human control is merely formal or functionally effective. The following elements are proposed as essential to the implementation of extended MHC in the DSS-AI domain:

- *Adopting a human-centered approach grounded in cognitive ergonomics.* The design of DSS-AI interfaces and workflows must take into account human cognitive limitations, especially under conditions of stress and time pressure. Insights from cognitive ergonomics highlight the importance of mitigating cognitive biases that are exacerbated by the accelerated tempo of AI-supported decision-making. In particular, systems should be designed to slow down decision points where high-stakes judgments are required, or to flag situations where overreliance on machine inferences is likely. This includes ensuring that output is presented in interpretable formats and that users are not cognitively overloaded by irrelevant or excessive information [17].
- *Ensuring a clear distribution of responsibilities along the chain of command and control.* The allocation of roles and responsibilities in the use of DSS-AI must be explicitly defined, including at the design, deployment, and operational stages. This requirement is rooted in both ethical imperatives and established legal frameworks, particularly those of international humanitarian law (IHL) and international criminal law (ICL), let alone *jus ad bellum* rules that are of paramount importance in scenarios involving AI-DSS nuclear warning systems. Discussions around human responsibilities for attacks against civilians allegedly carried out by the IDF on the advice of DSS-AI illustrate the legal and moral necessity of clarifying who is responsible for what in AI-assisted decision-making processes.
- *Promoting suitable training programs.* One key challenge to MHC is the formation of prior beliefs about AI capabilities, which are more and more often shaped by technological hype and related marketing narratives. The prevailing discourse around DSS-AI tends to emphasize superhuman accuracy and efficiency, while obscuring system limitations such as brittleness and lack of contextual understanding. Moreover, pre-conceptions formed through experience may also result in selective adherence, whereby users accept DSS recommendations only when they confirm preexisting intuitions or stereotypes [3]. Training should aim to correct these fallacies and prepare human operators to critically engage with DSS outputs. In other terms, it must focus on fostering reflective skepticism, not just "procedural familiarity" with the system.

Together, these three elements—human-centered design, clearly distributed responsibilities, and reflective training—constitute the operational core of extended MHC in DSS-AI environments. Their effectiveness will depend on their incorporation into formal regulations, standard operating procedures, and international best

practices. Codifying these measures at both national and international levels is essential to resisting the accelerationist competition that currently governs much of AI adoption efforts in military contexts. Without such regulatory embedding, the very speed that gives DSS-AI its tactical appeal will continue to erode the conditions that are necessary to preserve and enhance human control.

4.5 Conclusions

This paper has explored the emerging challenge of extending the scope of the MHC requirement to cover AI-enabled decision-support systems in military operations. It has been argued that the problem of MHC is not confined to the automation of lethal force, but applies also to decision-making processes increasingly shaped by DSS-AI. We thus proposed an extended MHC model that identifies human control not simply with presence in the decision loop, but with the capacity to exercise judgment under conditions that are cognitively and institutionally conducive to meaningfully perform that task.

We argued that DSS-AI—despite being functionally different from AWS—raise similar normative concerns. The issue is not whether force is applied autonomously or manually, but whether critical functions of target identification and engagement remain under conditions that allow for genuine human deliberation. As Article36 foresightedly put it more than a decade ago, what matters is not who *presses the button*, but whether the decision-making context permits meaningful human agency in interpreting the information that leads to that action.

To address the *how* of extended MHC in AI-supported military operations, we outlined a set of operational recommendations. These include adopting a human-centered design approach informed by cognitive ergonomics; ensuring clearly distributed responsibility throughout the chain of command; and implementing training programs that promote critical engagement with AI systems.

Looking forward, a number of steps might profitably be taken to bring the extended MHC problem into the fold of global governance.

First, the issue should be integrated into key arms control and disarmament forums, including the Non-Proliferation Treaty (NPT) review process and high-level initiatives such as the Summit on Responsible AI in the Military Domain (REAIM) [22].

Second, a comprehensive and empirically grounded inquiry is needed to identify where, how, and under what conditions AI systems are already being deployed in military contexts that challenge the feasibility of extended MHC. This diagnostic effort should inform regulatory priorities and guide the development of accountability structures.

Third and finally, international legal instruments should be considered to regulate military applications of AI beyond (and *before*) AWS, using extended MHC as a foundational principle. Such an initiative might take the form of a legally binding treaty dedicated to arms control in the AI domain.

One must be realistic, though, about the geopolitical headwinds facing multilateral disarmament efforts today, ranging from the deterioration of relations between military powers (most notably the US and China), to renewed military investments around the world (including in the EU under the banner of "Readiness2030") and the worrisome withdrawal of some States from cornerstone treaties such as the Anti-Personnel Mine Ban Convention.

A second-best pathway lies in embedding extended MHC principles within national doctrines, especially IHL manuals and operational protocols. This could lay the groundwork for the eventual emergence of customary international norms or help coalesce, at least in the medium term, the consensus required for future treaty negotiation.

References

1. Y. Abraham, A mass assassination factory: inside Israel's calculated bombing of Gaza, in *+972 Magazine* (2023). https://www.972mag.com/mass-assassination-factory-israel-calculated-bombing-gaza
2. Y. Abraham, Lavender: the AI machine directing Israel's bombing spree in Gaza, in *+972 Magazine* (2024). https://www.972mag.com/lavender-ai-israeli-army-gaza/
3. S. Alon-Barkat, M. Busuioc, Human–AI interactions in public sector decision making: "automation bias" and "selective adherence" to algorithmic advice. J. Public Adm. Res. Theory **33**, 153–169 (2023)
4. D. Amoroso, G. Tamburrini, Toward a normative model of meaningful human control over weapons systems. Ethics Int. Aff. **35**(2), 245–272 (2021)
5. D. Amoroso, Sistemi di supporto alle decisioni basati sull'IA e crimini di guerra: alcune riflessioni alla luce di una recente inchiesta giornalistica. Diritti umani diritto internazionale **18**(2), 347–368 (2024)
6. Article36, Structuring debate on autonomous weapons systems, in Briefing Paper, Geneva 14–15 November (2013)
7. B. Biggio, F. Roli, Wild patterns: ten years after the rise of adversarial machine learning. Pattern Recogn. **84**, 317–331 (2018)
8. N. Bostrom, E. Yudkowsky, The ethics of artificial intelligence, in *The Cambridge Handbook of Artificial Intelligence*, eds. by K. Frankish, W.M. Ramsey (Cambridge University Press, Cambridge, 2014), pp. 316–334
9. J. Brendarn, A. Kulkarni, Fighting for seconds: warfare at the speed of artificial intelligence. Modern War Inst. West point (2023). https://mwi.westpoint.edu/fighting-for-seconds-warfare-at-the-speed-of-artificial-intelligence/
10. M.L. Cummings, Automation bias in intelligent time critical decision support systems, in *Decision Making in Aviation*, ed. by D. Harris (Routledge, London, 2015)
11. International Committee of the Red Cross, Position paper on Autonomous Weapons Systems. Geneva (2021)
12. D. Kahneman, *Thinking, Fast and Slow* (Farrar, Straus and Giroux, New York, 2012)
13. D. Mauri, *Autonomous Weapons Systems and the Protection of the Human Person. An International Law Analysis* (Edward Elgar, Cheltenham/Northampton, 2022)
14. D. Mauri, Numeri, persone, umanità. Sistemi di supporto alle decisioni umane in campo militare da parte dell'IDF e diritto internazionale umanitario. Diritti umani diritto internazionale **18**(2), 329–346 (2024)
15. M. Milanović, Mistakes of fact when using lethal force in international law: part II. EJIL:Talk! (2020). https://www.ejiltalk.org/mistakes-of-fact-when-using-lethal-force-in-international-law-part-ii/

16. S. Milgram, Behavioral study of obedience. J. Abnorm. Soc. Psychol. **67**(4), 371–378 (1963)
17. M. Draper et al., Human Systems Integration for Meaningful Human Control over AI-based systems (NATO, STO Technical Report, 2025). https://www.sto.nato.int/document/human-systems-integration-for-meaningful-human-control-over-ai-based-systems/
18. G. Moy et al., *Technical Report. Recent Advances in Artificial Intelligence and Their Impact on Defence* (Defence Science and Technology Group, Canberra, 2020). www.dst.defence.gov.au/sites/default/files/publications/documents/DST-Group-TR-3716_0.pdf
19. NSCAI (National Security Commission on Artificial Intelligence), Final Report (2021). https://assets.foleon.com/eu-central-1/de-uploads-7e3kk3/48187/nscai_full_report_digital.04d6b124173c.pdf
20. C. O'Meara, *Necessity and Proportionality and the Right of Self-Defence in International Law* (Oxford University Press, New York, 2021)
21. E.S. Probasco et al., AI for military decision-making. Harnessing the advantages and avoiding the risks, in *Issue Brief* (Center for Security and Emerging Technology, 2025). https://cset.georgetown.edu/publication/ai-for-military-decision-making/
22. A. Saltini, Assessing the implications of integrating AI in nuclear decision-making systems, in *Policy Brief* (European Leadership Network, 2025). https://europeanleadershipnetwork.org/policy-brief/assessing-the-implications-of-integrating-ai-in-nuclear-decision-making-systems/
23. F. Sauer, The military rationale for AI, in *Armament, Arms Control and Artificial Intelligence. The Janus-faced Nature of Machine learning in the Military Realm*, eds. by T. Reinhold, N. Schörnig (Springer, Cham, 2022), pp. 27–38
24. P. Scharre, Centaur warfighting: the false choice of humans vs. automation. Temple Int. Comp. Law J. **30**(1), 151–165 (2015)
25. F. Strack, R. Deutsch, The duality of everyday life: dual-process and dual system models in social psychology, in *APA Handbook of Personality and Social Psychology, Vol. 1. Attitudes and Social Cognition*, eds. by M. Mikulincer et al. (APA Books Washington DC, 2015), pp. 891–927