FEATURE

Toward a Normative Model of Meaningful Human Control over Weapons Systems

Daniele Amoroso and Guglielmo Tamburrini

In academic and diplomatic debates about so-called autonomous weapons systems (AWS), a watchword has rapidly gained ground across the opinion spectrum: All weapons systems, including autonomous ones, should remain under human control. While references to the human control element were already present in early documents on AWS,[1] the U.K.-based NGO Article 36 must be credited for putting it in the center of the discussion by circulating, since 2013, a series of reports and policy papers making the case for establishing *meaningful* human control (MHC) over individual attacks as a legal requirement under international law.[2]

Unlike the calls for a preemptive ban on AWS, the notion of MHC has been met with substantial interest by a number of states. This response is explainable by a variety of converging factors. To begin with, human control is an easily *understandable* concept that "is accessible to a broad range of governments and publics regardless of their degree of technical knowledge"; it therefore provides the international community with a "common language for discussion" on AWS.[3] A second feature contributing to the success of the notion of MHC is its

Daniele Amoroso is professor of international law in the Department of Law of the University of Cagliari, located in Cagliari, Italy.

Guglielmo Tamburrini is professor of the philosophy of science and technology at Università di Napoli Federico II, located in Naples, Italy.

Ethics & International Affairs, 35, no. 2 (2021), pp. 245–272.

doi:10.1017/S0892679421000241

"constructive ambiguity,"[4] which may prove helpful in bridging the gap between various positions expressed at the international level on the AWS issue. Third, and finally, the notion allows one to shift the focus of the AWS debate from a recalcitrant definitional problem, requiring one to draw precise boundaries between automation and autonomy in weapons systems, to a normative problem, requiring one to specify what kinds and levels of human control ought to be exerted over weapons systems in general.[5] Unlike the former definitional problem, the latter normative problem appears to be more tractable and more likely to be successfully addressed through negotiations. In this perspective, it was correctly underlined that one should not look at MHC necessarily as a "solution" to the ethical and legal concerns raised by autonomy in weapons systems.[6] Rather, MHC indicates the right "approach" to cope with them.[7]

And indeed, growing attention to the issue of human control has emerged from diplomatic talks that have been taking place in Geneva within the Group of Governmental Experts on Lethal Autonomous Weapons Systems (GGE), established by the State Parties to the Convention on Certain Conventional Weapons (CCW). In addition to those State Parties explicitly endorsing the call for an MHC requirement,[8] most delegations taking part in the CCW proceedings underscored the importance of this issue in both their official speeches and their working papers. The International Committee of the Red Cross (ICRC) expressed concerns for the "loss of human control and judgement in the use of force and weapons" in a ground-breaking position paper from May 12, 2021. These concerns motivated the ICRC's call for new legally binding rules on AWS.[9]

These converging views are coherent with the guiding principles formulated by the GGE, which were endorsed by the High Contracting Parties to the CCW in November 2019. In particular, principle (b) posits that "human responsibility for decisions on the use of lethal force must be retained."[10] This is, however, exactly where the international consensus stops. As many commentators have pointed out, it is far from settled—*even among those favoring an MHC requirement*—what its actual content should be or, to put it more sharply, what is normatively demanded to make human control over weapon systems truly "meaningful." This is exactly the issue that we address in this article, so as to fill the MHC placeholder with more precise contents, grounded in major ethical and legal concerns expressed in scholarly, diplomatic, and political debates about AWS.

First, we briefly review the main stumbling blocks in building a satisfactory definition of AWS; in other words, in building one that is sufficiently precise and that

is neither overly restrictive nor overly permissive. As mentioned above, these persisting impediments speak clearly in favor of shifting the focus of ethical and legal debates away from definitions of AWS and toward a specification of MHC contents. To lay the groundwork for identifying the core components of MHC, we examine ethical and legal arguments from the AWS debate, which selectively concern *jus in bello* principles of distinction, proportionality, and precaution; responsibility ascription; and human dignity protection. We then argue that these ethical and legal arguments concur to pinpoint distinctive human obligations regarding weapons systems control. These obligations constrain human-weapon shared control by retaining for human agents the roles of "fail-safe actor," "accountability attractor," and "moral agency enactor." We maintain that uniform models of human control—that is, those applying one size of human control to all weapons systems and uses thereof—fail to properly accommodate these normative requirements. Hence the need for an MHC framework that is both "differentiated," in rejecting uniform solutions to the issue of human control, and "principled," in favoring solutions that invariably retain the fail-safe, accountability, and moral agency roles for humans in human-weapon interactions. We additionally argue for "prudential" solutions, chiefly by appealing to epistemic uncertainties about AWS behaviors. The prudential solution we advance here imposes by default higher levels of human control of target selection and engagement processes; designated exceptions to this default rule are admitted solely on the basis of an international agreement entered into by states for specific weapons systems and uses thereof, provided that lower levels of human control are by consensus found sufficient to meet the fail-safe actor, accountability attractor, and moral agency enactor requirements. Finally, we suggest that the outlined differentiated, principled, and prudential framework provides a most appropriate normative basis for both national arms review policies and any international legal instrument enshrining the MHC requirement (such as a possible Protocol VI to the CCW).

Escaping the Definitional Conundrum: What Is the "Autonomy" of Weapons Systems?

Despite what extensive diplomatic debates on the legality of AWS might suggest, agreement is still lacking among states as to what qualifies a weapons system as being "autonomous." However, discussion about this core issue is not fragmented into myriad competing notions of "autonomy." Rather, it is polarized around two

basic understandings of autonomy, epitomized by the respective definitions advanced by the Ministry of Defence (MoD) in the U.K. and the U.S. Department of Defense (DoD).

The MoD sets a very demanding requirement for a weapons system to be autonomous. Indeed, the MoD's Joint Doctrine publications devoted to unmanned aircraft systems defines AWS as systems "capable of understanding higher level intent and direction," and thus as being able "to take appropriate action to bring about the desired state."[11] Even though this condition falls short of equating AWS with moral agents that are free and capable of acting on their genuine intentions,[12] the requirement is still a tall order for a machine to satisfy. Neither existing nor foreseeable weapons systems genuinely comprehend the meaning of higher-level goals and intentions. And no educated guess is currently available about the prospect and timeline of their development. Thus, weapons systems that are genuinely autonomous according to the MoD construal are projected into an undetermined technological future. As a consequence, if one were to follow this definition, ethical and legal discussions about AWS would be simply premature, insofar as they would concern technological possibilities that are far removed in time.

Upon closer examination, the MoD construal is affected by some conceptual flaws. First, the MoD's Joint Doctrine documents propose a problematic distinction between autonomous and automated/automatic weapons systems. The existing weapons systems that perform some given tasks without human control are described as "automated" or "automatic" (but *not* autonomous), in that they are "programmed to logically follow a predefined set of rules in order to provide an outcome" as a "response to inputs from one or more sensors."[13] This distinction is blind to the fact that a variety of existing weapons systems act in response to sensor inputs without having been programmed to logically follow a predefined set of rules. Indeed, a major goal of machine-learning techniques is to develop systems performing perception-action cycles without explicitly programming their rules of behavior.

Equally problematic in the MoD's Joint Doctrine documents are the alleged differences between the respective predictability profiles of autonomous and automatic weapons. The outputs of an automatic weapons system are claimed to be entirely foreseeable once "the set of rules under which it is operating" are known. In contrast with this, in an autonomous system only "the overall activity . . . will be predictable," while "individual actions may not be."[14] This proposed

distinction belittles the difficulty of predicting precisely what computational or robotic systems do on the basis of their programmed rules of behavior, without taking into account how the environment contributes to shaping their course of action.[15] In particular, unstructured and surprise-inducing warfare scenarios—think, for instance, of urban warfare—may challenge predictions that are exclusively based on knowledge of the weapons system's set of rules. We elaborate further on this point in later sections, in connection with both human control issues and the prudential MHC approach that we advocate.

Turning now to the second definitional pole, the DoD Directive 3000.09, "Autonomy in Weapon Systems," is remarkably absent of any attempt to capture what autonomy is in terms of high-level intent or comprehension, an unpredictability system profile, or programmed or learned sets of rules. This directive instead introduces a *functional*, task-oriented criterion by which weapons systems count as autonomous: a weapons system must be able, after activation, to "select and engage targets without further intervention by a human operator."[16] It is of the utmost relevance for the international debate on AWS that the DoD definition was embraced by the ICRC, which refers to AWS as "weapons that can independently select and attack targets,"[17] and by Human Rights Watch, which describes AWS as "weapons that could select and engage targets without human intervention."[18] Thus, this formulation of the central properties of autonomy has the support of crucial international stakeholders: the most powerful and technologically advanced military power, the foremost international humanitarian organization, and the global coordinator of the coalition of NGOs advocating a ban on AWS (the Campaign to Stop Killer Robots). Against this backdrop, the more restrictive requirement adopted by the MoD was even challenged domestically in the U.K. by the House of Lords Select Committee on Artificial Intelligence as being "out of step" with the broad understanding of "autonomy" generally agreed upon at the international level.[19]

Unlike the MoD definition, which projects AWS into an undetermined technological future, the DoD definition accommodates a number of presently operating weapons systems. These include antimateriel defense systems, like Germany's Nächstbereichschutzsystem (NBS) MANTIS and Israel's Iron Dome; active protective systems for vehicles, like the South African–Swedish LEDS-150 Land-Electronic Defense System; loitering weapons, like Israel's antiradiation Harpy NG (New Generation) system; a variety of offensive fire-and-forget munitions, like the U.K.'s Brimstone missile in one of its operative modes; and

stationary robotic sentinels, like South Korea's Super aEgis II, which patrols the border between North and South Korea. Arguably, even landmines satisfy the DoD definition.[20]

By the yardstick of the DoD definition, ethical and legal issues surrounding AWS concern weapons systems that are under development or already exist. An exemplary case is provided by Israel's Harpy NG loitering munition system, which is able to overfly assigned areas for up to nine hours in search of enemy radar sources to shoot down without human operators having to intervene after activation. The Harpy NG's extended time window for loitering, along with its limited perceptual-discrimination capacities, raises ethical and legal concerns about its autonomous target selection and engagement functions, especially in dynamic warfare scenarios where civilians and civilian objects may suddenly come into sight. At the same time, however, one should carefully note that other systems satisfying the DoD definition are not equally problematic from a normative perspective—such as the NBS MANTIS and the Iron Dome—given their intended use as antimateriel defense systems.

From an ethical and legal standpoint, one has so far been confronted with an overly restrictive and overly inclusive definition. The overly restrictive MoD definition defers to an undetermined future in which "genuine" AWS will be developed, thereby neglecting the ethical and legal concerns about the autonomous targeting activities of the various existing systems. The overly inclusive DoD definition puts in the same basket antimateriel defense systems, loitering munitions like the Harpy NG, and any future offense system that one may imagine operating autonomously in the fog of war. Hence, this definition does not distinguish existing weapons systems endowed with ethically and legally problematic sorts of autonomy (such as the Harpy) from the less problematic systems, such as the NBS MANTIS and Iron Dome, whose autonomy in identifying and engaging targets has gone unchallenged in international fora.

Robert Sparrow has advanced a definition of AWS that occupies a middle ground between the above overly restrictive and overly inclusive definitions:

> I understand an "autonomous" weapon as one that is capable of being tasked with identifying possible targets and choosing which to attack without human oversight, and that is sufficiently complex such that, even when it is functioning perfectly, there remains some uncertainty about which objects and/or persons it will attack and why. This admittedly rough-and-ready definition represents my attempt to single out an

interesting category of systems while avoiding entering into an extended and difficult argument about the precise nature of machine autonomy.[21]

This definition is intuitively appealing, as it points to a normatively interesting category of existing or technologically plausible AWS. However, by relying on the idea of a system that is "sufficiently complex," this definition introduces additional issues of precision and usability. To make this definition more precise, one should identify which uses of the word "complexity" (for example, physical, computational, informational, cognitive, relating to the system's operational environment or even of its internal functional organization) are relevant here. And to make the definition usable to discriminate between AWS and other weapons systems, one has to set out an appropriate measure of complexity.

Let us take stock. From a normative standpoint, the DoD approach must be preferred over the MoD approach, as the latter sweeps under the rug substantive ethical and legal issues about the autonomy of existing weapons systems. However, as shown by Sparrow's discussion, one cannot take the DoD's functional description as a normatively interesting *definition* of AWS. Hence, we take it to express a *necessary* (but not sufficient) condition for the kind of autonomy one is interested in examining from a distinctively normative perspective.

In the absence of precise and conceptually adequate definitions, the necessary condition expressed by the DoD definition affords a sensible starting point for ethical and legal discussions of autonomy in weapons systems. Indeed, this condition enables one to focus on machines controlling exactly those functions that constitute the main source of ethical and legal concerns, namely the replacement of human decision-makers in the performance of morally sensitive tasks governed by international (humanitarian) law. The wording chosen by the ICRC captures exactly this source of concern, as it refers to "the critical functions of selecting and attacking targets."[22] It is indeed the permission to let machines perform those *critical* functions without human intervention and the proposal of preserving MHC over those critical functions that the whole normative debate on AWS ultimately revolves around. However, as intimated by Sparrow's attempt to find a more stringent definition of an autonomous weapon, due attention must be paid to the fact that ethical and legal concerns are mostly about AWS that satisfy the DoD condition *and* are based on advanced technologies for artificial perception and decision-making.[23] At the same time, however, the relatively simple Harpy NG is a stark reminder that this is not necessarily the case.

The Ethical and Legal Case for Meaningful Human Control

In this section, we argue that distinctive MHC contents emerge from an effort to make operational various reflections on AWS that are carried out in the light of just war theory, ICL, and concerns about the protection of human dignity.[24]

Just War Principles and Humans as Fail-Safe Actors

Just war theory distinguishes between just and unjust ways for soldiers to fight in armed conflicts and provides moral criteria to judge warfare actions on this basis. These criteria prominently include the noncombatant's right to immunity and the proper moral balances between the means and the ends of military action.[25] These pillars of just war theory found their way into international humanitarian law (IHL)—notably in the principles of distinction, proportionality, and precaution that are enshrined in the 1977 Additional Protocols to the Geneva Convention.

Compliance with IHL is a focal point of ethical and legal debates about AWS. In the early days of this debate, Ronald Arkin expressed his concern for the poor record of human compliance with norms governing the warfare conduct of belligerent parties. He suggested that "ethically restrained" AWS might abide "by the internationally agreed upon Laws of War" better than human war fighters have, as these machines may be programmed with more conservative engagement policies; may reduce battlefield victims by more precise targeting; and may dispense with fear, anger, frustration, and other mental strains conducive to human violations of IHL. However, Arkin signaled "daunting problems" that needed to be addressed before developing IHL-compliant AWS (such as "the development of effective perceptual algorithms capable of superior target discrimination capabilities"). And he did not take it for granted that "this venture will be successful."[26] Arkin's technology assessment entails that the current and foreseeable AWS fail to meet, in many warfare situations, the benchmark represented by properly trained and conscientious human soldiers.[27]

One cannot exclude, therefore, the idea that present and foreseeable AWS may incur violations of the IHL principle of distinction that trained and conscientious soldiers would hardly be prone to. Notably, systems developed by means of current machine-learning technologies were demonstrated by adversarial testing to be prone to unexpected, counterintuitive, and potentially catastrophic mistakes that a human operator would easily detect and avoid.[28] Vivid examples are learning

perceptual systems that were found by adversarial testing to mistake school buses (protected by the distinction principle) for ostriches and turtles for rifles.[29] Contrary to what widespread recourse to anthropomorphic language may suggest, human and autonomous decision-making processes indeed remain qualitatively different and are likely to err in qualitatively different ways. The source of such qualitative differences lies in what AI experts call the "semantic gap." This expression indicates the fact that machines do not perceive the world in the same way as humans.[30] Accordingly, granting (for the sake of argument) that the overall performance of AWS will come to match or even statistically surpass the performance of human combatants still does not entail that this will occur *in each and every situation*. Even the most sophisticated weapon may commit (or be induced to commit) disastrous mistakes from an IHL perspective—mistakes that might have been avoided if a human operator had been substantively involved in the decision-making loop. In light of complementary strengths and weaknesses of human and machine capabilities, it is thus questionable whether the elimination of human judgment and supervision is compatible with the obligation to take all feasible precautions to prevent disproportionate damage to the civilian population.

The belief that AWS may violate the IHL principles of distinction and proportionality is regarded as a "serious possibility" across the wide spectrum of positions emerging from the AWS debate.[31] Indeed, its negation is not included, to the best of our knowledge, in the belief set of any qualified participant in this ethical and legal debate. This serious possibility in itself speaks in favor of maintaining MHC over AWS in order to avoid the occurrence of such IHL violations. Thus, humans involved in human-weapon shared control ought to play the role of fail-safe actors who take due care of situations that engineered fail-safe systems may not adequately deal with in the foreseeable technological future.[32]

Epistemic Uncertainties and War Crimes

A second cluster of issues in the AWS debate revolves around the so-called accountability (or responsibility) gap problem; in particular, around the question of whether the removal of human operators from the targeting decision-making processes would hinder responsibility ascriptions when AWS violate IHL.[33] This concern, however, is not unanimously shared. Susanne Burri, while acknowledging as a serious possibility that AWS can cause violations of just war and IHL principles, suggests that this worry is properly defused when "a conscientious

human commanding officer deploys [AWS] only in contexts where they are able to identify sufficient conditions for the morally permissible infliction of lethal harm."[34] Should a commander deploy an autonomous weapons system in a context where such a sufficiency condition is not fulfilled, she would be held responsible for unacceptable damages. Fulfilling this sufficiency condition presupposes, however, an epistemic assessment of AWS capabilities and action outcomes in each given battlefield situation on the part of commanding officers.

Clearly, the epistemic assessment that commanders must come up with need not exclude altogether the risk of undesired outcomes. As Johannes Himmelreich points out, a commander may have control only over a probabilistic outcome of her orders in general and of AWS actions in particular.[35] Accordingly, Himmelreich introduces the notion of "robust tracking control" to capture the commander's control duties with their attending epistemic risks and responsibilities, suggesting that

> an *a* has control over whether an outcome *x* occurs if [1] there is an order *a* can give, such that [2] if *a* were to give this order, then *x* would occur (in all relevantly similar situations), and [3] if *a* were not to give this order, then *x* would not occur (in all relevantly similar situations).[36]

One should be careful to note that in order to infer that outcome *x* of an AWS action is under their robust tracking control, commanders must be in the position to assert with a high level of confidence that the actual battlefield situation is *similar* in all relevant aspects to their own cognitive model of the battlefield situation and its predicted outcomes. Confidence in this similarity judgment is bound to decrease as the incompleteness and uncertainty characterizing the commander's cognitive model of the battlefield increases. In particular, cluttered and unstructured battlefield situations—involving, for example, sustained interactions among opposing groups of soldiers, AWS, and other artificial agents—are likely to admit only incomplete and uncertain representations.

In these scenarios, the critical autonomy of weapons systems becomes incompatible with the preservation of robust tracking control. Commanders may not be in the position to license the required epistemic assessment about fast interactions between complex AWS or rapid environmental changes impinging on even relatively simple loitering systems such as the Harpy NG. As noted earlier, the Harpy's extended loitering time and limited perceptual discrimination capacities raise substantive concerns about the sustained persistence in dynamic and unstructured

warfare scenarios of conditions for morally permissible infliction of lethal harm. In such scenarios, civilians and civilian objects may suddenly appear and enemies may rapidly move mobile sources of radar signals from otherwise uninhabited locations to the vicinity of vulnerable locations like schools or hospitals.

The epistemic limitations on the commander's capability to make robust predictions about the behaviors of AWS are what drive the "accountability gap" problem, especially from the perspective of ICL.[37] Suppose that an autonomous weapons system commits a material act that would be equivalent to a war crime should this act have been performed by human beings. In other words, the AWS targeting decisions, were they taken by human agents, would trigger individual criminal responsibility. Since the direct targeting decision was taken by the AWS, who should be held responsible for its conduct? If sufficiently stringent MHC conditions on the release and subsequent action of the AWS were not in place, commanders might complain that an interruption of their robust tracking control condition occurred due to unexpected epistemic predicaments, such as bizarre perceptual errors of the AI learning system embedded in the AWS or hostile interactions on the battleground. Under these circumstances, commanders could plausibly claim that their responsibility was mitigated, or altogether excluded, "due to epistemic constraints."[38] The list of potentially responsible persons in the decision-making chain includes those overseeing the AWS operation, manufacturers, robotics engineers, software programmers, and those who conducted the AWS weapons review. These persons may cast their defense against responsibility charges on account of their limited decision-making roles, the complexities of the AWS, or the difficulty of anticipating and testing weapons against all possible battlefield scenarios prior to their actual deployment. Cases may therefore occur where one cannot ascertain the existence of the mental element (intent or knowledge), which is required under ICL to ascribe and punish criminal responsibilities. Consequently, no one would be held *criminally* responsible, notwithstanding that the conduct at stake materially amounts to an international crime.

As weapons systems become more autonomous in their targeting decisions, the role of individual criminal responsibility (and ICL more generally) becomes smaller in governing the use of armed violence. Despite what some have argued,[39] one cannot balance waning international criminal responsibility by introducing some kind of collective strict liability that would oblige the deploying state to pay for all damages to civilians caused by AWS, independent of the fault of any commanders. Proposals of this kind rest on the incorrect assumption that the

various responsibility regimes embraced by international law are ultimately replaceable among one another. This assumption neglects the "complementarity" among responsibility regimes, which is rooted in the distinction between the "predominantly reparational aspect of state responsibility and the punitive character of criminal law proceedings against individuals."[40] The latter, indeed, makes ICL unique in playing the crucial function of "pronounc[ing] the wrongfulness of actions that harm the interests of the international community as a whole."[41] This is for the simple reason that, as famously affirmed by the Nuremberg Tribunal, international crimes "are committed by men, not by abstract entities, and only by punishing individuals who commit such crimes can the provisions of international law be enforced."[42]

A sufficiently stringent MHC would enable one to meet this ICL normative requirement, so as to preserve the kind and degree of human control over the critical functions of selecting and engaging targets that is needed to *justly* ascribe individual criminal responsibility, instead of thwarting it. In this sense, shared control between humans and autonomous weapons should be conceived so as to ensure that humans always play the role of accountability attractors.

AWS as an Affront to Human Dignity

A third cluster of arguments against autonomy in weapons systems is grounded in the principle of human dignity protection, insofar as this principle dictates that decisions affecting the life, physical integrity, and property of human beings involved in an armed conflict should be reserved entirely to human operators and cannot be entrusted to an autonomous artificial agent.[43]

Burri, on the other hand, contends by analogy that "there is nothing unduly selfish or inconsiderate about a lion that hunts down its prey, just as there is nothing disrespectful or insensitive about an LAR (Lethal Autonomous Robot) that has decided to lethally engage an enemy combatant based on the algorithms that were programmed into it."[44] On our view, this stance belittles the moral implications of life-taking or life-sparing decisions, and reduces these decisions to verifications of legitimacy. Furthermore, Burri's analogy overlooks the difference between legitimate target assessment and moral evaluations prizing the value of human life. Elaborating on views originally advanced by Thomas Nagel,[45] Sparrow tackles the question of AWS and respect for human dignity. Even in wartime, writes Sparrow, "it is essential that we acknowledge the personhood of those with whom we interact." And in particular, he continues, "the decision to take another person's life must be compatible

with such a relationship." Hence, "when AWS decide to launch an attack the relevant interpersonal relationship is missing. Indeed, in some fundamental sense there is no one who decides whether the target of the attack should live or die. The absence of human intention here appears profoundly disrespectful."[46]

From the MHC viewpoint, this argument implies that human control should ensure that AWS life-or-death decisions are traceable to human intentions. In other words, humans involved in shared control with autonomous weapons must play the role of moral agency enactors, ensuring that decisions affecting the life, physical integrity, and property of people involved in armed conflicts are not taken by artifacts that fail to qualify as moral agents. Establishing this form of control is likely to conflict with some forms of autonomy in weapons systems, such as those involving autonomous-targeting decisions taken by complex AWS operating in unstructured warfare scenarios and unchecked by human beings over extended temporal and spatial windows. Not invariably so, however, there are limited forms of autonomous targeting in weapons systems that appear to be consistent with human dignity protection requirements.

Each of the arguments examined here so far regarding the capabilities of AWS to select and engage targets is deontological in character, insofar as it makes an appeal to distinctive moral duties (such as the IHL-embedded duty to protect noncombatants, the ICL-embedded duty to preserve criminal responsibility ascriptions in warfare, and the duty to respect human dignity). In addition to these types of deontological arguments, consequentialist arguments have also played a central role in the debate concerning the ethical and legal acceptability of AWS. As noted earlier, Arkin emphasized possible consequentialist advantages of future AWS, such as fewer victims on the battlefield. Others have cautioned instead that AWS make wars easier to wage, thus leading to negative consequentialist appraisals in the wider context of preserving global peace and stability.[47] Such appraisals, however, are comparatively less relevant than the deontological arguments we have outlined in this section in connection with the main problem that we are concerned with—*the problem of shaping the contents of MHC*. This relationship will be addressed in the concluding section. First, we turn to a normative framework for MHC.

A Normative Framework for MHC

The foregoing ethical and legal reasoning goes a long way in shaping the content of MHC, by pinpointing functions prescriptively assigned to human control, and

by providing criteria with which to distinguish perfunctory from truly meaningful human control. In particular, the arguments above suggest a threefold role that must be fulfilled in order for human control over weapon systems to be considered "meaningful." First, the obligation to comply with just war principles and IHL in warfare operations entails that human operators must play the role of fail-safe actor, preventing malfunctioning weapons from resulting in direct attacks against the civilian population or excessive collateral damages.[48] Second, to preserve ICL-embedded criminal responsibility ascription, human control must function as an accountability attractor, securing legal conditions for criminal responsibility ascription in case a weapon follows a course of action that is in breach of international law.[49] Third, respect for human dignity demands that human control operates as a moral agency enactor, ensuring that decisions affecting the life, physical integrity, and property of people involved in armed conflicts, including combatants, are not taken by artificial agents.[50] To be ethically and legally sound, therefore, rules aimed at determining MHC obligations should guarantee that these functions are jointly and invariably fulfilled.[51] Let us call "principled" any solution to the MHC problem that meets these conditions for the human element in a scenario of shared control between humans and weapons. Against this backdrop, we will now examine whether major proposed solutions for maintaining MHC qualify as principled solutions in this sense.

Uniform Solutions

Several attempts have been made—by scholars, states, and NGOs—to define human-weapon shared control policies dictated by the MHC requirement. While significantly different from one another in many respects, these various proposals generally share a common feature: they aspire to capture optimal partnership using a one-size-fits-all formula that is supposed to apply *uniformly* to all kinds of weapons systems and all of their possible uses.

This feature is particularly evident in the so-called wider loop approach, advocated by the Dutch government: On this approach, MHC is regarded as having been satisfactorily exercised by human commanders at the planning stage of the targeting process.[52] This approach may have some limited applicability and relevance with regard to the deliberate targeting of military objectives, as long as these are known in advance to exist and can be mapped with reasonable certainty. It is, however, largely unhelpful with regard to dynamic targeting, which pursues targets of opportunity. Moreover, it allows for weapons to have unrestrained

autonomy after deployment. In this way, the Dutch approach appears to be deeply problematic in scenarios populated by civilians: It drives a wedge between the state owing a duty of care to the civilian population and the actual ability to reliably comply with that duty by influencing the course of events through its agents.[53] This is especially true for AWS endowed with the capability of loitering for sustained periods of time in search of enemy targets (like the above-mentioned Harpy NG). After all, the conditions licensing the activation of a loitering AWS by human operators may rapidly change in warfare scenarios characterized by erratic dynamics and surprise-seeking behaviors.

The uniform and overly permissive approach sketched out and advocated by the Dutch government is located at one end of the spectrum of MHC constructs. At the other end of the spectrum, one finds similarly uniform but overly restrictive approaches to defining MHC, whereby no autonomy whatsoever in weapons systems is permitted.[54] While undoubtedly praiseworthy for their attention to humanitarian concerns, these attempts run the risks of (1) banning some weapons whose *jus in bello* admissibility has so far gone undisputed on ethical or legal grounds, and (2) requiring milder forms of human control to remove the troubling ethical and legal implications potentially ensuing from the autonomy of these weapons. Extant cases in point are the United States' Phalanx Defense Systems, Israel's Iron Dome, and the German NBS MANTIS, when they are used as intended—that is, as protective shields from incoming shells and missiles.

A reflection on these systems suggests that a "supervised autonomy," or "human-on-the-loop," policy that occupies a middle ground between the two extremes analyzed above might suffice to strip their autonomy of ethically and legally troubling traits.

As defined in the DoD directive, human-supervised AWS are designed "to provide human operators with the ability to intervene and terminate engagements, including in the event of a weapon system failure, before unacceptable levels of damage occur."[55] Notably, one may use these systems for defending manned installations and platforms from "attempted time-critical or saturation attacks," provided that they do not select "humans as targets,"[56] which is exactly the case for the defense systems mentioned earlier. While effective for these and similar warfare scenarios, supervised autonomy is not the silver bullet for every ethical and legal concern raised by AWS. To begin with, keeping humans on the loop does not prevent faster and faster offensive AWS from being developed, which

will eventually reduce the role of human operators to a perfunctory supervision of decisions taken at superhuman speed, leaving only the illusion of meaningful human control. Moreover, "automation bias"—the human tendency to over trust machine decision-making—is demonstrably exacerbated when the human role consists solely of the ability to override decisions that have already been autonomously made by machines.[57]

In brief, any desire to grant even limited forms of critical autonomy to weapons systems[58] must be coupled with a principled approach to MHC that includes ethical and legal considerations for the fail-safe actor, accountability attractor, and moral agency enactor roles identified above. To this end, we suggest giving up the quest for a one-size-fits-all solution[59] to the concept of MHC in favor of a suitably differentiated approach, without relinquishing the demand that this solution be nonetheless principled.

A Differentiated, Principled, and Prudential Framework for MHC

In order to actually implement the various elements that constitute MHC, we need *rules* to bridge the gap between ethical and legal principles, on the one hand, and specific weapon systems and uses thereof, on the other. These "bridge rules" should be able to express the fail-safe, accountability, and moral agency conditions for exercising MHC over weapons systems *in context*.

Schematically, these rules may be conceived of as "if-then" statements. Their "if" part should include properties concerning *what* mission the weapon system is involved with, *where* the system will be deployed, and *how* it will perform its tasks. The "what properties" must specify operational goals (defensive vs. offensive); the targeting modes (deliberate vs. dynamic); and the nature of targets to be engaged (human combatants, human-occupied vehicles, and/or inhabited military objects vs. uninhabited vehicles and military objects). The "where properties" must concern dynamic features of the environment, including interactions with hostile autonomous artificial agents, and have special regard for the presence of civilians, civilian objects, and friendly forces. The "how properties" must concern information-processing and sensory-motor capabilities that the system puts to work in its mission and that affect its overall controllability and predictability. Learned decision-making and "swarm intelligence" abilities (which may increasingly characterize future AWS), jointly with loitering capabilities, are significant cases of "how properties" raising concerns from an MHC perspective.[60]

The "then" part of the bridge rules should establish what kind of human-machine shared control would be legally required on each use of a given weapons system. Following (and only slightly modifying) a taxonomy proposed by Noel Sharkey,[61] one may consider five basic levels (Ls) of human-machine interactions to fill in the then part of the bridge rules. These are ordered according to decreasing levels of human control and increasing levels of machine control in connection with target selection and engagement tasks.

L1: A human engages with and selects targets, and initiates any attack.
L2: A program suggests alternative targets and a human chooses which to attack.
L3: A program selects targets and a human must approve them before the attack.
L4: A program selects and engages targets, but is supervised by a human who retains the power to override its choices and abort the attack.
L5: A program selects targets and initiates an attack on the basis of the mission goals as defined at the planning/activation stage, without further human involvement.

On these assumptions, an if-then rule may—for instance—assume the following form: "IF the weapons system is programmed to perform an exclusively antimateriel defensive function (what property) AND is deployed in a sufficiently structured scenario (where property), THEN (L4) human operators must be put in charge of supervising the weapon's selection of targets and be given the power to override its choices."

As noted by Heather Roff and Richard Moyes, it "may be difficult in detailed terms" to specify what is the appropriate level of control needed to establish MHC for each use of a weapons system, as too many contextual factors may come into play.[62] However, within the wide space of "differentiated" and "principled" solutions to the MHC problem, we argue for a solution that is also "prudential," in the sense of taking every reasonable precaution to preserve the fail-safe, accountability, and moral agency contents of MHC, thereby minimizing the risk of IHL breaches, accountability gaps, or affronts to human dignity, described in "The Ethical and Legal Case for Meaningful Human Control" section of this article.[63] Against the background of L1–L5, the gist of a differentiated, principled, and prudential solution to MHC is specifiable by means of (1) a general default policy and (2) exceptions formulated as specific bridge rules. The default policy would require that higher levels of human control (L1 and L2) should be applied, unless the given autonomous weapons system is an exception that the international community of states has agreed to handle by means of specific bridge rules allowing for lower levels of human control.

Bridge rules for internationally agreed-upon exceptions should specify what level of interaction (L1–L5) is required to ensure the threefold role of MHC (fail-safe actor, accountability attractor, moral agency enactor) as a function of suitable combinations of the what, where, and how properties. Deviations from the general default policy should be crafted by taking into account the following observations:

1. The L4 human supervision and veto level might be deemed acceptable only for AWS with exclusively anti-materiel defensive functions (what property).
2. Deliberate targeting (what property) by AWS may be pursued at L3. Since targeting decisions have actually been taken by humans at the planning stage, operators have only to confirm that no changes occurred in the battle space that may affect the lawfulness of the operation.
3. One may allow human control at L3 for AWS programmed to engage human or humanly inhabited targets (what property) when activated in fully structured scenarios (where property), such as in the high seas or deserts where civilians and civilian objects are not present. Unlike L4, L3 ensures that a human on the attacking end can verify whether there are persons hors de combat and take appropriate denial or granting measures. An example of AWS potentially operating in a fully structured environment is the legacy South Korean robotic sentry Super aEgis II, which patrols the Korean demilitarized zone and is reportedly able to function in full autonomous mode.[64]
4. Capabilities that may reduce the predictability of weapon systems' behavior, such as loitering, learned decision-making, and swarming (how properties), should in principle be treated as factors pushing toward the application of higher levels of human control (L1 and L2).
5. The full autonomy of L5 should be considered incompatible with the MHC requirement. While it is true that operational constraints set at the planning or activation stages may play an important role in limiting weapons' autonomy, this "boxed autonomy" alone is insufficient to ensure MHC,[65] unless operational space and time frames are so severely circumscribed as to make targeting decisions reliably traceable to human operators.

The latter exclusion is suitably illustrated by reference to homing fire-and-forget munitions; that is, precision-guided munitions that use passive sensors or active

seekers to track onto moving targets. Most fire-and-forget munitions simply lock on targets preselected by human operators, thereby meeting the highest human control level (L1). Those involving mutual coordination to avoid hitting the same target can be fired in a salvo to simultaneously attack multiple targets that are close to one another and previously identified by human operators. In this case, each munition might be said, prima facie, to select and engage targets on its own based on mission goals defined by human operators, which would be an L5 degree of human control. On closer scrutiny, however, the mission is so precisely defined (for example, "Destroy *that specific* line of enemy tanks") that nobody would question the conclusion that those targeting decisions are collectively attributable to the human launching the attack and thus fall under L1 human control.

The same conclusion does not hold for fire-and-forget munitions, like the Brimstone, enabled to search for and hit targets within a "kill box" designated by a human operator. Although the lack of loitering capabilities requires the operator to be convinced that there are valid targets within the box (otherwise the missile would be wasted),[66] a missing link is detectable in the decision-making chain, with the consequence that targeting decisions appear in fact to be taken by the weapon system, not by the human user. While this is admittedly a borderline case to be treated with caution, the existence of this missing link speaks against including this functionality of fire-and-forget munitions among candidate exceptions to the general default policy.

In this respect, the history of the Brimstone is particularly instructive. A previous version of the Brimstone, solely operating in the "autonomous" mode, was deemed incompatible with the rules of engagement of the Afghanistan campaign, which prompted the U.K.'s Royal Air Force to issue an urgent operation requirement in 2007, aimed at modifying the existing missiles and enabling human selection of targets through laser guidance ("a man-in-the-loop capability," in the manufacturer's own wording).[67] Contrary to what is sometimes maintained, therefore, the ethical and legal acceptability of offensive autonomy—even in such a limited form—is far from undisputed, even among the primary users of these weapons systems.

The Quality of Human Involvement: Design and Training

The MHC requirement is compatible, as the above observations suggest, with various forms of human-weapon partnerships in critical target selection and

engagement functions. The MHC requirement is also compatible—in limited circumstances—with the preservation of the critical autonomy of some weapons systems. Nevertheless, the incompatibility of MHC with L5 means that human operators retain some exclusive control privileges throughout, notably the power to approve at L3 or veto at L4 machine targeting decisions. In all these cases, it is crucial to ensure a proper quality of human involvement, so that human control privileges are meaningfully exerted. This requires intervention at both the design and training stages of AWS life cycles.[68]

AWS should be *designed* to provide human operators and commanders with sufficient humanly understandable information about machine data processing, thus achieving adequate situational awareness ("interpretability" requirement); and to obtain an account of the reasons why the machine suggests or intends to take a certain targeting decision ("explainability" requirement). Moreover, military personnel *training* should foster awareness of both established and likely limits in the proper functioning of weapons systems and related predicaments in the capability to predict and control their behavior.[69]

Which Legal Regime Should Be Used for Meaningful Human Control over Weapons Systems?

At the August 2018 meeting of the GGE, the delegations of Austria, Brazil, and Chile submitted a joint proposal for a mandate to "negotiate a legally binding instrument to ensure meaningful human control over critical functions in lethal autonomous weapon systems."[70] One may legitimately doubt that this proposal will be swiftly followed up within the CCW institutional framework, as some major military powers, including the United States, have been resisting a solution of this kind, although the recent call by the ICRC for new legally binding rules on AWS may prove to be a game changer. At the same time, however, this article's proposal to relinquish the quest for a one-size-fits-all solution to the MHC issue in favor of a suitably differentiated approach may help sidestep this stumbling block at the CCW. Indeed, diplomatic and political discontent about an MHC requirement that appears to be overly restrictive with respect to the limited autonomy of some weapons systems (such as the Phalanx, Iron Dome, and NBS MANTIS) might be mitigated by recognizing the possibility of negotiating exceptions to L1–L2 human control, as long as one is able to identify weapons systems and contexts of use in which milder forms of human control are acceptable. In any

case, one cannot ignore the possibility that concerned states may at some point explore alternative venues to negotiate an international agreement establishing the MHC requirement, as has already occurred with the Anti-Personnel Mine Ban Convention. It seems therefore appropriate to begin thinking about the content of such a—for now wholly hypothetical—treaty. In light of the foregoing, we suggest including the following in any MHC convention or protocol:

1. The MHC requirement for *all* weapon systems must be stated in a provision of general purport. Its specific contents should be clarified in three ensuing parts: "control in use," "training," and "control by design."[71]
2. The control in use part will undoubtedly be the more important and challenging part for states to agree upon. As explained above, our ethically and legally motivated suggestion is to establish higher levels of human control (L1–L2) as a default policy, and to regulate exceptions thereto by way of internationally accepted bridge rules.
3. The provision(s) on training and control by design must spell out state obligations as set out in the section on the quality of human involvement.
4. Crucial to the actual MHC implementation is the introduction of transparency obligations, verification procedures, and confidence-building measures. The analysis carried out here provides some indications as to *what* information State Parties should collectively share. For instance, states should notify other parties of any proposed exception to the default policy, formalize the exception by means of suitable bridge rules, and provide proper ethical and legal motivations as a basis for shared decision-making.

This outline allows one to pinpoint both the common ground between and the distinguishing features of our proposal and the one recently put forth by the ICRC. In a nutshell, the ICRC proposal contains two prohibitions of general character regarding (1) anti-personnel AWS and (2) AWS that are by design unpredictable.[72] Outside these two provisions, autonomy in weapons systems is permitted, albeit subject to robust constraints, including—but not limited to—the requirement of human supervision and veto power (as at L4).[73] A differentiated approach is thus adopted by the ICRC proposal too. Indeed, anti-personnel and anti-materiel uses of weapons systems are subject to different levels of human control: denied autonomy (L1 and L2) for anti-personnel uses and supervised autonomy (L4) for anti-materiel uses. A crucial difference, however, between

the ICRC proposal and our own lies in the content of the default rule. The prudential character of our proposal is expressed by a default rule that is unambiguously restrictive, as it applies throughout in the absence of specific provisions to the contrary. Indeed, the higher levels of human control (L1 and L2) are imposed on all weapons systems that go unmentioned in internationally agreed upon designated exceptions. The ICRC proposal makes no mention of a similarly restrictive default policy, to be applied to all cases not covered by the prohibition of anti-personnel and unpredictable AWS. Admittedly, the constraints on admissible AWS set out by the ICRC suggest the idea that the higher levels of human control (L1 and L2) should be applied in all but a quite limited number of circumstances. Nonetheless, we maintain that the need to minimize the risk of IHL breaches, accountability gaps, or affronts to human dignity is better ensured by a less open-ended formulation. In particular, our restrictive interpretation, rejecting as impermissible what is not explicitly agreed on by states, would offer legal instruments to contrast unilateral, runaway uses of AWS threatening compliance with MHC and yet falling outside the scope of the ICRC requested prohibitions.

During GGE talks, Switzerland alternatively propounded to address the AWS regulation issue through a so-called compliance-based approach.[74] This less ambitious alternative to an MHC convention/protocol gained widespread support among military powers. States would be demanded to implement strict national weapons reviews so as to guarantee that autonomy in weapons systems would comply with international law. National weapons reviews might in principle conform to our differentiated, principled, and prudential MHC framework. The DoD directive "Autonomy in Weapons Systems" is a remarkable case in point, as it envisages training and design guidelines jointly with differentiated policies for human control.[75] Room is made in it for autonomous targeting solely in relation to the application of "non-lethal, non-kinetic force . . . against materiel targets,"[76] and human-supervised autonomy is admitted for the defense of human-inhabited installations or platforms against "time-critical or saturation attacks" and "with the exception of selecting humans as targets."[77] Adopting this policy as a legal yardstick for national weapons reviews would go a long way toward MHC enforcement. It should be noted, however, that the same directive authorizes high-ranking U.S. officials to override this very policy.[78] Accordingly, a weapons system failing a legal review in the United States today may pass it in the future on account of newly introduced policy changes. This possibility exposes a major flaw of national weapons reviews in the absence of internationally shared legal frameworks, for

legal review criteria may radically differ from one state to another (or within the same state over time). This fragmentation risk is hardly avoidable by the mere sharing of national best practices.

Conclusion

We have argued in this article for a principled, differentiated, and prudential MHC framework. Selected ethical and legal reasons advanced in AWS debates buttress a principled stand on indefeasible fail-safe actor, accountability attractor, and moral agency enactor roles for human controllers of weapons systems. Given the wide variety of AWS and the graded and normative concerns they give rise to, these core requirements can be met by adopting suitably differentiated approaches, which neither rule out all autonomy in weapons systems nor relegate human control to a nominal or perfunctory role. And differentiated control levels must be modulated along the "what," "where," and "how" dimensions of weapons systems tasks and capabilities. Finally, our framework is prudential, in that stricter forms of human control are applied by default on precautionary grounds, unless another form is otherwise agreed on by the international community of states.

A final remark is in order, which concerns the relationship between the outlined core contents of MHC and the stability and peace concerns raised by AWS. While ultimately motivated here by deontological reasons, the core contents of MHC would also prove effective from a consequentialist perspective. Crucially, by enforcing the MHC requirement in the ways unfolded here, one connects the tempo of military attacks to human cognitive capacities and reaction times (with the notable exception of certain uses of defensive AWS), thereby mitigating the widespread concern that autonomy in weapon systems might lead to an acceleration in the pace of war that is incompatible with human cognitive and sensory motor limitations.[79] By preserving the MHC communication channel between humans and weapons systems, as even required under the L3 and L4 exceptions, one indeed sets up a bulwark against the risk of runaway interactions between AWS that can lead to unintentional outbreaks of conflict.

NOTES

[1] See, for example, U.S. Department of Defense, "Autonomy in Weapons Systems," Directive 3000.09 (November 21, 2012), p. 2: "Autonomous and semi-autonomous weapon systems shall be designed to allow commanders and operators to exercise *appropriate levels of human judgment* over the use of force" (emphasis added).

[2] See, for example, Article 36, "Killer Robots: UK Government Policy on Fully Autonomous Weapons" (policy paper, April 2013), available at: article36.org/wp-content/uploads/2013/04/Policy_Paper1.pdf.

[3] United Nations Institute for Disarmament Research, *The Weaponization of Increasingly Autonomous Technologies: Considering How Meaningful Human Control Might Move the Discussion Forward*, UNIDIR Resources no. 2 (Geneva: UNIDIR, 2014), p. 3.
[4] Rebecca Crootof, "A Meaningful Floor for 'Meaningful Human Control,'" *Temple Journal of International & Comparative Law* 30, no. 1 (Spring 2016), pp. 58–60.
[5] Maya Brehm, "Meaningful Human Control" (paper presented at the Informal Meeting of Experts on Lethal Autonomous Weapons Systems, Convention on Certain Conventional Weapons, Geneva, April 14, 2015), p. 5.
[6] United Nations Institute for Disarmament Research, *Weaponization of Increasingly Autonomous Technologies*, p. 4.
[7] Ibid., p. 4.
[8] See, for example, Austria, Brazil, and Chile, "Proposal for a Mandate to Negotiate a Legally-Binding Instrument That Addresses the Legal, Humanitarian and Ethical Concerns Posed by Emerging Technologies in the Area of Lethal Autonomous Weapons Systems (LAWS)" (working paper CCW/GGE.2/2018/WP.7, submitted to the Group of Governmental Experts on Lethal Autonomous Weapons Systems conference, Convention on Certain Conventional Weapons, Geneva, August 29, 2018).
[9] International Committee of the Red Cross, *ICRC Position on Autonomous Weapons Systems* (Geneva: ICRC, May 12, 2021), p. 2.
[10] "Guiding Principles Affirmed by the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems," Annex 4 in *Final Report of the Meeting of the High Contracting Parties to the Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects*, CCW/MSP/2019/CRP.2/Rev.2 (Geneva, November 15, 2019).
[11] Ministry of Defence, *The UK Approach to Unmanned Aircraft Systems*, Joint Doctrine note 2/11 (withdrawn) (Wiltshire, U.K.: Development, Concepts and Doctrine Centre, March 30, 2011), p. 14; and Ministry of Defence, *Unmanned Aircraft Systems*, Joint Doctrine publication 0-30.2 (Wiltshire, U.K.: Development, Concepts and Doctrine Centre, August 2017), p. 21.
[12] Issues about the status of AWS as moral agents, quasi-agents, or mere technological instruments go beyond the scope of this article. For a discussion of these topics, see Michael Robillard, "No Such Thing as Killer Robots," *Journal of Applied Philosophy* 35, no. 4 (November 2018), pp. 705–17, at p. 707.
[13] Ministry of Defence, *Unmanned Aircraft Systems*, p. 21.
[14] Ibid., p. 21.
[15] This fact was duly recognized ever since the early days of artificial intelligence research on intelligent behavior in machines and biological systems. Herbert Simon, for one, characterized an intelligent system as one involving an interface between inner and outer environments. The complexity of its behavior, and the related difficulty of predicting what it will do reflect in significant ways, according to Simon, the complexity of its environment. See Herbert A. Simon, *The Sciences of the Artificial*, 3rd ed. (Cambridge, Mass.: MIT Press, 1996), p. 6.
[16] U.S. Department of Defense, "Autonomy in Weapons Systems," pp. 13–14.
[17] International Committee of the Red Cross, "Views of the International Committee of the Red Cross on Autonomous Weapon Systems" (paper submitted to the Informal Meeting of Experts on Lethal Autonomous Weapons Systems, Convention on Certain Conventional Weapons, Geneva, April 11, 2016), p. 1.
[18] Human Rights Watch, *Losing Humanity: The Case against Killer Robots* (Cambridge, Mass.: International Human Rights Clinic, Harvard Law School, November 19, 2012), p. 1.
[19] Select Committee on Artificial Intelligence, House of Lords, *AI in the UK: Ready, Willing and Able?* (HL Paper 100, Report of Session 2017–19, Authority of the House of Lords, April 16, 2018), paras. 60 and 345.
[20] For an overview of the weapons systems that may be characterized as "autonomous" on the basis of the DoD definition, see Vincent Boulanin and Maaike Verbruggen, *Mapping the Development of Autonomy in Weapon Systems* (Solna, Sweden: Stockholm International Peace Research Institute, November 2017).
[21] Robert Sparrow, "Robots and Respect: Assessing the Case against Autonomous Weapon Systems," *Ethics & International Affairs* 30, no. 1 (2016), pp. 93–116, at p. 95.
[22] International Committee of the Red Cross, "Views of the International Committee of the Red Cross," p. 5.
[23] Sparrow, "Robots and Respect," p. 105; and Heather M. Roff, "Killing in War: Responsibility, Liability, and Lethal Autonomous Robots," in Fritz Allhoff, Nicholas G. Evans, and Adam Henschke, eds., *Routledge Handbook of Ethics and War: Just War Theory in The Twenty-First Century* (New York: Routledge, 2013), p. 355.
[24] A crystal-clear articulation of the problems at stake has been provided by the then special rapporteur on extrajudicial, summary, or arbitrary executions, Christof Heyns, in his groundbreaking 2013 report on lethal autonomous robotics, "Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions" (A/HRC/23/47, April 9, 2013, paras. 57–92).

[25] For discussion, see Michael Walzer, *Just and Unjust Wars: A Moral Argument with Historical Illustrations* (New York: Basic Books, 1977), especially chaps. 8, 9, 14–16.

[26] Ronald C. Arkin, *Governing Lethal Behavior in Autonomous Robots* (Boca Raton, Fla.: CRC Press, 2009), pp. 211–12.

[27] On the opportunity of using this benchmark for AWS compliance with IHL, see Marcus Schulzke, "Robots as Weapons in Just Wars," *Philosophy & Technology* 24, no. 3 (2011), pp. 293–306, at p. 294; and Robert Sparrow, "Twenty Seconds to Comply: Autonomous Weapon Systems and the Recognition of Surrender," *International Law Studies* 91 (2015), pp. 699–728, at pp. 709–10.

[28] International Committee of the Red Cross, *Artificial Intelligence and Machine Learning in Armed Conflict: A Human-Centred Approach* (Geneva: ICRC, June 6, 2019), p. 11.

[29] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus, "Intriguing Properties of Neural Networks," arXiv.org, last updated February 19, 2014; and Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok, "Synthetizing Robust Adversarial Examples," arXiv.org, last updated June 7, 2018.

[30] International Committee of the Red Cross, *Artificial Intelligence and Machine Learning in Armed Conflict*, p. 20.

[31] On the notion of "serious possibility," see Isaac Levi, "Serious Possibility," in Esa Saarinen, Risto Hilpinen, Ilkka Niiniluoto, and Merrill Provence Hintikka, eds., *Essays in Honour of Jaakko Hintikka* (Dordrecht, Netherlands: Springer, 1979), pp. 219–36.

[32] The crucial role of humans as fail-safe actors is shown by the well-known episode involving the lieutenant colonel Stanislav Yevgrafovich Petrov, who in fact prevented a malfunctioning in the Soviet satellite warning system from unleashing a large-scale nuclear war by correctly qualifying in the system's report that a number of missiles had been launched by the United States as a false alarm. For this reason, one could name this requirement of human control over weapons systems the "Petrov condition." On the relevance of the Petrov episode in the AWS debate, see Paul Scharre, *Army of None: Autonomous Weapons and the Future of War* (New York: W. W. Norton, 2018), pp. 1–2, 6, 144–45, 207, and 305.

[33] The "accountability-responsibility" gap issue was first raised in 2004 by Andreas Matthias in relation to autonomous learning machines in general, and then by Robert Sparrow with specific regard to AWS. See Andreas Matthias, "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata," *Ethics and Information Technology* 6 (2004), pp. 175–83; and Robert Sparrow, "Killer Robots," *Journal of Applied Philosophy* 24, no. 1 (February 2007), pp. 62–77.

[34] Susanne Burri, "What Is the Moral Problem with Killer Robots?," in Bradley Jay Strawser, Ryan Jenkins, and Michael Robillard, eds., *Who Should Die?: The Ethics of Killing in War* (New York: Oxford University Press, 2018), p. 170.

[35] Johannes Himmelreich, "Responsibility for Killer Robots," *Ethical Theory and Moral Practice* 22, no. 3 (June 2019), pp. 731–47, at p. 735.

[36] Ibid., p. 736. A similar emphasis on the "tracking condition," but in these articles in explicit reference to the notion of MHC, may be found in Filippo Santoni de Sio and Jeroen van den Hoven, "Meaningful Human Control over Autonomous Systems: A Philosophical Account," *Frontiers in Robotics and AI* 5, no. 15 (February 2018), pp. 6–8; and Giulio Mecacci and Filippo Santoni de Sio, "Meaningful Human Control as Reason-Responsiveness: The Case of Dual-Mode Vehicles," *Ethics and Information Technology* 22 (June 2020), pp. 103–15.

[37] We are not concerned here, therefore, with an (alleged) *moral* "responsibility gap," whose existence is denied by—among others—Robillard. See Robillard, "No Such Thing as Killer Robots."

[38] Roff, "Killing in War," pp. 357–58.

[39] See, for example, Rebecca Crootof, "War Torts: Accountability for Autonomous Weapons," *University of Pennsylvania Law Review* 146, no. 6 (2016), pp. 1347–1402.

[40] Andrea Bianchi, "State Responsibility and Criminal Liability of Individuals," in Antonio Cassese, ed., *The Oxford Companion to International Criminal Justice* (Oxford: Oxford University Press, 2009), p. 24. See also, with specific reference to the AWS debate, Thompson Chengeta, "Accountability Gap, Autonomous Weapon Systems and Modes of Responsibility in International Law," *Denver Journal of International Law & Policy* 45, no. 1 (April 2016), pp. 1–50, at pp. 49–50.

[41] Miriam Gur-Arye and Alon Harel, "Taking Internationalism Seriously: Why International Criminal Law Matters," in Kevin Jon Heller, Frédéric Mégret, Sarah M. H. Nouwen, Jens David Ohlin, Darryl Robinson, eds., *The Oxford Handbook of International Criminal Law* (New York: Oxford University Press, 2020), p. 234. See also, with specific regard to AWS, Darren M. Stewart, "New Technology and the Law of Armed Conflict: Technological Meteorites and Legal Dinosaurs?," *International Law Studies* 87 (2011), p. 292; and John Danaher, "Robots, Law and the Retribution Gap," *Ethics and Information Technology* 18, no. 4 (2016), pp. 299–309.

[42] Trial of German Major War Criminals, Judgment, pt. 22, *Proceedings of the International Military Tribunal Sitting at Nuremberg, Germany* (October 1, 1946), p. 447.
[43] For a compelling version of this argument, see Christof Heyns, "Autonomous Weapons Systems: Living a Dignified Life and Dying a Dignified Death," in Nehal Bhuta, Susanne Beck, Robin Geiß, Hin-Yan Liu, Claus Kreß, eds., *Autonomous Weapons Systems: Law, Ethics, Policy* (Cambridge, Mass.: Cambridge University Press, 2016), pp. 3–20.
[44] Burri, "What Is the Moral Problem with Killer Robots?," p. 173.
[45] Thomas Nagel, "War and Massacre," *Philosophy & Public Affairs* 1, no. 2 (1972), pp. 123–44.
[46] Sparrow, "Robots and Respect," pp. 106–107. After all, if the principles of humanity and human dignity have any minimal function in warfare, it is precisely that of reminding belligerents of the fact that the killing of another human being, even if in full compliance with the law, remains "a truly existential choice that each soldier needs to justify before his own conscience." Alex Leveringhaus, *Ethics and Autonomous Weapons* (London: Palgrave, 2016), p. 92.
[47] For these arguments, see Jürgen Altmann and Frank Sauer, "Autonomous Weapon Systems and Strategic Stability," *Survival* 59, no. 5 (September 2017), pp. 117–42; and Guglielmo Tamburrini, "On Banning Autonomous Weapons Systems: From Deontological to Wide Consequentialist Arguments," in Bhuta et al., *Autonomous Weapons Systems*, pp. 122–42. As Sparrow correctly pointed out, however, these wide consequentialist concerns are not specific to AWS (Sparrow, "Robots and Respect," p. 106).
[48] Paul Scharre, "Centaur Warfighting: The False Choice of Humans vs. Automation," *Temple International and Comparative Law Journal* 30, no. 1 (Spring 2016), pp. 151–165, at p. 154.
[49] Thompson Chengeta, "Defining the Emerging Notion of 'Meaningful Human Control' in Autonomous Weapon Systems," *New York Journal of International Law & Politics* 49 (2017), pp. 833–90.
[50] International Committee of the Red Cross, "Ethics and Autonomous Weapon Systems: An Ethical Basis for Human Control?" (working paper CCW/GGE.1/2018/WP.5, submitted to the Group of Governmental Experts on Lethal Autonomous Weapons Systems conference, Convention on Certain Conventional Weapons, Geneva, March 29, 2018), paras. 23–26.
[51] For a similar approach, see the joint SIPRI-International Committee of the Red Cross report *Limits on Autonomy in Weapon Systems*: Netta Goussac, Vincent Boulanin, Moa Peldán Carlsson, and Neil Davison, *Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control* (Solna, Sweden: Stockholm International Peace Research Institute, 2020), pp. 25–26.
[52] Advisory Council on International Affairs (AIV) and Advisory Committee on Issues of Public International Law (CAVV), *Autonomous Weapon Systems: The Need for Meaningful Human Control*, AIV no. 97/CAVV no. 26 (Hague: Advisory Council on International Affairs, October 2015).
[53] International Committee of the Red Cross, "Statement on Agenda Item 5(a)" (transcript of the statement delivered at the March 2019 meeting of the Group of Governmental Experts on Lethal Autonomous Weapons Systems, Convention on Certain Conventional Weapons, Geneva, March 26, 2019), p. 3, available at: reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2019/gge/statements/26March_ICRC5a.pdf; and David Akerson, "The Illegality of Offensive Lethal Autonomy," in Dan Saxon, ed., *International Humanitarian Law and the Changing Technology of War* (Leiden, Netherlands: Brill, 2013), p. 87.
[54] Chengeta, "Defining the Emerging Notion of 'Meaningful Human Control,'" pp. 888–89: "[MHC] entails that: (a) The decision to kill and the legal judgment pertaining to individual attacks must be made by a human in real time, in other words the actual time during which a target is to be killed. (b) The weapon system depends on the authorization of the operator to execute his or her decision to kill without which, it cannot proceed. (c) The weapon system has an abort mechanism that allows the operator to abort an attack in the event that it is no longer lawful to kill a target due to changed circumstances or other reasons prescribed in international law. (d) Operators have an inherent obligation to monitor weapon systems they activate while the weapon systems execute operators' decisions to kill."
[55] U.S. Department of Defense, "Autonomy in Weapons Systems," p. 13.
[56] Ibid., p. 3, para. 4(c)(2).
[57] Noel E. Sharkey, "Staying in the Loop: Human Supervisory Control of Weapons," in Bhuta et al., *Autonomous Weapons Systems*, pp. 23–38, at pp. 32–33. In light of the above, one may question the viability of the guidelines on AWS (Gs) recently adopted by the French Defence Ethics Committee, which combine the "boxed autonomy" and "supervised autonomy" approaches (Defence Ethics Committee, *Opinion on the Integration of Autonomy into Lethal Weapons Systems* [Paris, France: Ministère des Armées, April 29, 2021]). In brief, the Committee makes the case for the ethical acceptability of so-called "Partially Autonomous Lethal Weapons Systems," whose defining features would lie in the fact that (1) human commanders set in advance "the target to be reached, space and time limits, constraints, engagement rules, for each mission performed" (G6) and that (2) "emergency deactivation systems" are designed into the weapons' systems (G17).

[58] Sparrow, "Robots and Respect," pp. 94–97; and Daniele Amoroso, *Autonomous Weapons Systems and International Law: A Study on Human-Machine Interactions in Ethically and Legally Sensitive Domains* (Baden-Baden, Germany: Nomos, 2020), pp. 222–24.

[59] United States, "Human-Machine Interaction in the Development, Deployment and Use of Emerging Technologies in the Area of Lethal Autonomous Weapons Systems" (working paper CCW/GGE.2/2018/WP.4, submitted to the Group of Governmental Experts on Lethal Autonomous Weapons Systems conference, Convention on Certain Conventional Weapons, Geneva, August 28, 2018), para. 9: "There is not a fixed, one-size-fits-all level of human judgment that should be applied to every context."

[60] International Committee of the Red Cross, "Statement on Agenda Item 5(b)" (transcript of the statement delivered at the March 2019 meeting of the Group of Governmental Experts Lethal Autonomous Weapons Systems conference, Convention on Certain Conventional Weapons, Geneva, March 25, 2019), p. 3, available at: reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2019/gge/statements/25March_ICRC.pdf.

[61] Sharkey, "Staying in the Loop," pp. 34–37. Deviations concern, notably, levels 4 and 5.

[62] Heather M. Roff and Richard Moyes, Article 36, "Meaningful Human Control, Artificial Intelligence and Autonomous Weapons" (briefing paper prepared for the Informal Meeting of Experts on Lethal Autonomous Weapons Systems, Convention on Certain Conventional Weapons, Geneva, April 11–15, 2016), p. 6.

[63] In a sense, this policy would be an application of the precautionary principle (or approach), which mainly emerged in the field of international environmental law, to the domain of human-weapons interaction. For a similar view, although with regard to the wider strategic implications of AWS, see Denise Garcia, "Future Arms, Technologies, and International Law: Preventive Security Governance," *European Journal of International Security* 1, no. 1 (February 2016), pp. 94–111.

[64] Jean Kumagai, "A Robotic Sentry for Korea's Demilitarized Zone," *IEEE Spectrum* 44, no. 3 (April 2007), pp. 16–17.

[65] International Panel on Regulation of Autonomous Weapons, *Focus on Technology and Application of Autonomous Weapons*, "Focus On" report no. 1 (Berlin: iPRAW, August 2017), pp. 15–16.

[66] Scharre, *Army of None*, p. 107.

[67] MBDA, "MBDA Presents a More Versatile Brimstone," Defense-Aerospace.com, February 10, 2009, www.defense-aerospace.com/articles-view/release/3/102330/mbda-details-more-versatile,-dual_mode-brimstone.html.

[68] Goussac et al., *Limits on Autonomy in Weapon Systems*, pp. 33–35.

[69] For a fuller account of these aspects of the MHC requirement, see Roff and Moyes, "Meaningful Human Control," pp. 2–3. Special attention to training requirements is paid in the 2021 guidelines on AWS by the French Defence Ethics Committee. See Defence Ethics Committee, *Opinion on the Integration of Autonomy into Lethal Weapons Systems*, G21–G25.

[70] Austria, Brazil, and Chile, "Proposal for a Mandate to Negotiate a Legally-Binding Instrument," p. 1.

[71] For more on the expressions "control in use" and "control by design," see International Panel on Regulation of Autonomous Weapons, *Focus on the Human-Machine Relation in LAWS*, "Focus on" report no. 3 (Berlin: iPRAW, March 2018).

[72] International Committee of the Red Cross, *ICRC Position on Autonomous Weapons Systems*, p. 2.

[73] Ibid. Further restraints include limits on the types of targets, limits on the duration, geographical scope and scale of use, and limits on situations of use.

[74] Switzerland, "A 'Compliance-Based' Approach to Autonomous Weapon Systems" (working paper CCW/GGE.1/2017/WP.9, submitted to the Group of Governmental Experts on Lethal Autonomous Weapons Systems conference, Convention on Certain Conventional Weapons, Geneva, November 10, 2017).

[75] In this respect, it has been correctly noted that "without expressly endorsing the term, the DOD seems to be currently following a policy requiring meaningful human control in all but name." Adam Cook, *Taming Killer Robots: Giving Meaning to the "Meaningful Human Control" Standard for Lethal Autonomous Weapon Systems*, JAG School paper no. 1 (Maxwell Airforce Base, Ala.: Air University Press, 2019), p. 17.

[76] U.S. Department of Defense, "Autonomy in Weapons Systems," p. 3, para. 4(c)(3).

[77] Ibid., p. 3, para. 4(c)(2).

[78] Ibid., p. 3, para. 4(d).

[79] Altmann and Sauer, "Autonomous Weapon Systems and Strategic Stability."

Abstract: The notion of meaningful human control (MHC) has gathered overwhelming consensus and interest in the autonomous weapons systems (AWS) debate. By shifting the focus of this debate to MHC, one sidesteps recalcitrant definitional issues about the autonomy of weapons systems and profitably moves the normative discussion forward. Some delegations participating in discussions at the Group of Governmental Experts on Lethal Autonomous Weapons Systems meetings endorsed the notion of MHC with the proviso that one size of human control does not fit all weapons systems and uses thereof. Building on this broad suggestion, we propose a "differentiated"—but also "principled" and "prudential"—framework for MHC over weapons systems. The need for a differentiated approach—namely, an approach acknowledging that the extent of normatively required human control depends on the kind of weapons systems used and contexts of their use—is supported by highlighting major drawbacks of proposed uniform solutions. Within the wide space of differentiated MHC profiles, distinctive ethical and legal reasons are offered for principled solutions that invariably assign to humans the following control roles: (1) "fail-safe actor," contributing to preventing the weapon's action from resulting in indiscriminate attacks in breach of international humanitarian law; (2) "accountability attractor," securing legal conditions for international criminal law (ICL) responsibility ascriptions; and (3) "moral agency enactor," ensuring that decisions affecting the life, physical integrity, and property of people involved in armed conflicts be exclusively taken by moral agents, thereby alleviating the human dignity concerns associated with the autonomous performance of targeting decisions. And the prudential character of our framework is expressed by means of a rule, imposing by default the more stringent levels of human control on weapons targeting. The default rule is motivated by epistemic uncertainties about the behaviors of AWS. Designated exceptions to this rule are admitted only in the framework of an international agreement among states, which expresses the shared conviction that lower levels of human control suffice to preserve the fail-safe actor, accountability attractor, and moral agency enactor requirements on those explicitly listed exceptions. Finally, we maintain that this framework affords an appropriate normative basis for both national arms review policies and binding international regulations on human control of weapons systems.

Keywords: autonomous weapons systems, meaningful human control, human dignity, just war theory, accountability