ICRAC

International Committee for Robot Arms Control

What makes human control over weapons systems "meaningful"?

August 2019

Report to the CCW GGE

WHAT MAKES HUMAN CONTROL OVER WEAPON SYSTEMS "MEANINGFUL"?

Daniele Amoroso (Università di Cagliari and ICRAC, daniele.amoroso@unica.it)

Guglielmo Tamburrini (Università di Napoli Federico II and ICRAC,
guglielmo.tamburrini@unina.it)

Contents

Executive Summary

"All weapon systems, including autonomous ones, should remain under meaningful human control". This "meaningful human control" (MHC) formula expresses a point of overwhelming consensus in the AWS debate, and was promptly met with interest by a substantial number of States participating in discussions at CCW meetings. MHC is in fact an easily understandable formula; it is characterized by a dialogue-facilitating constructive ambiguity; it enables one to sidestep recalcitrant definitional problems regarding the distinction between autonomy and automation. However, there are still many substantive issues to address in order to clarify what is normatively demanded to make human control over weapon systems truly "meaningful".

The present ICRAC Report contributes to move forward the AWS debate on MHC (i) by filling the MHC placeholder with more precise contents, and (ii) by identifying on this basis some key aspects of any legal instrument enshrining the MHC requirement (such as, e.g., a Protocol VI to the CCW).

Ethical and legal arguments advanced against autonomy in weapon systems go a long way towards shaping the content of MHC. Most relevant in this respect are the following three arguments against machines performing the critical functions of target selection and engagement without human intervention. *First*, AWS would be unable to comply with the principles of distinction, proportionality, and precaution embedded into International Humanitarian Law (IHL). *Second*, AWS are likely to result in an accountability gap. *Third*, AWS would run counter to the principle of human dignity. These claims provide guidelines to distinguish perfunctory from truly meaningful human control, by pinpointing functions that one must prescriptively assign to human control.

What are the ethically and legally motivated functions that humans must perform to ensure MHC over weapon systems? First, human control must afford a *fail-safe mechanism*, preventing IHL breaches from occurring. Second, it must serve as a *catalyst for accountability*. Third, it must ensure that it is a *moral agent*, and not an artificial one, that makes decisions affecting the life, physical integrity and property of people.

Two chief problems need to be addressed and solved in order to fill the MHC placeholder in a way that secures that these functions are maintained: *(i)* how to guarantee a proper *quality* of human involvement, through appropriate training and design; *(ii)* how to identify and properly establish *exclusive control privileges for human operators*.

While both issues are important for MHC, the most crucial one concerns (ii), that is, in the words of the Estonian and Finnish CCW delegations, "the final interaction between a human [and] a weapon system before that system delivers force". Indeed, the multiple attempts – made so far by scholars, states, and NGOs – to define control privileges for human operators under the MHC requirement are generally affected by a common weakness: they aspire to capture optimal partnership with one formula, which is supposed to apply uniformly to all kinds of weapon systems and to each one of their possible uses. In general, these attempts are either overly permissive or else overly restrictive.

On the side of overly permissive attempts, one may recall the Dutch (or "wider loop") approach, whereby MHC would be in fact exerted by human commanders at the planning stage of the targeting process. This approach is largely unhelpful with regard to *dynamic* targeting, which pursues targets of opportunity. To the extent that it warrants the weapon with humanly unrestrained autonomy after deployment, moreover, the Dutch approach appears to be deeply problematic, especially when the weapon is activated in a scenario populated by civilians.

On the opposite side of overly restrictive attempts, one finds endeavours to define the MHC requirement in rigorous terms and in an all-encompassing manner, by means of a uniform human control protocol over every and each kind of AWS and use thereof, which ranges, say, from AWS selecting and attacking targets of opportunity in civilian populated areas to defensive systems against incoming rockets and missiles. These attempts may prove inadequate in that milder forms of human control might be equally able to "purify" the autonomous action of certain defensive weapon systems and other weapons operating in some very limited operational environments of its ethically and legally troubling implications.

To avoid predicaments of overly restrictive or else overly permissive approaches to MHC, we suggest giving up the quest for a one-size-fits-all solution, in favour of a suitably differentiated approach to the issue of MHC. This differentiated approach to MHC is nonetheless based on the unifying grounds provided by the ethical and legal principles recalled above. The application of these overarching principles in concrete situations must be facilitated and given concrete operational content by the formulation of a set of rules. These rules take the form of "if-then" rules expressing the fail-safe, accountability, and moral agency conditions for exercising *in context* a genuinely MHC over weapon systems.

The "if-part" of these rules should include properties concerning *what* mission the weapon system is involved into, *where* it will be deployed and *how* it will perform its tasks. The "then-part" should establish what kind of human-machine shared control would be legally required on each single use of a weapon system. Following a taxonomy proposed by Noel Sharkey (and only slightly modified below), one may schematically consider five basic types of human-machine interaction for the "then-part" of these rules, ordered according to decreasing levels of human control and increasing levels of machine control:

> 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 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 attack on the basis of the mission goals as defined at the planning/activation stage, without further human involvement.

The gist of our differentiated approach to MHC is specified by means of (i) a general default policy and (ii) exceptions formulated as specific rules. The general default policy prescribes that the higher levels of human control (L1 and L2) should be exerted. *Lower levels of human control may become acceptable only as internationally agreed on exceptions, to be clearly formalized by means of specific "if-then" rules.*

Any deviation from the general default policy should take into account (at least) the following observations:

> 1. It may be permissible to employ AWS to conduct deliberate targeting (i.e. to engage military objectives that are known in advance to exist and can be mapped with reasonable certainty: *what-property*) at a lower level of human control (L3), since in this case the targeting decisions have actually been taken by humans at the planning stage. The same level should

be required in relation to AWS programmed to execute dynamic targeting in structured warfare scenarios, provided that a human confirms that civilians and civilian objects are not present there (*where-property*).

2. The (L4) human supervision and veto level might be deemed as an acceptable level of control in case of AWS entrusted with the task of defending human-occupied sites or vehicles (*what-property*). This is the case of, say, the US Phalanx, the Israeli Iron Dome and the German *Nächstbereichschutzsystem* (NBS) MANTIS.

3. The use of capabilities that may reduce the overall predictability of the AWS' behaviour, such as loitering, learned decision-making, swarming (*how-properties*), should always be treated as a compelling factor pushing towards the application of the higher levels of human control.

4. The (L5) full autonomy level should instead be considered incompatible with the MHC requirement.

The present proposal of relinquishing the quest for a one-size-fits-all solution to the MHC issue in favour of a suitably differentiated approach may contribute to sidestep present stumbling blocks at the CCW and other international policy venues. Indeed, diplomatic and political discontent about an MHC requirement which appears to be overly restrictive with respect to the limited autonomy of some weapon systems (such as Phalanx, Iron Dome and MANTIS) might be mitigated by recognizing the possibility of negotiating exceptions to L1-L2 human control levels, as long as one is able to identify weapon systems and contexts of use in which milder forms of human control will still ensure genuine MHC, that is, one reflecting the above fail-safe, accountability and human agency requirements.

In the light of the foregoing, an "MHC Convention/Protocol" should have, as a minimum, the following contents:

1. The Preamble should recall the essential elements of the ethical and legal concerns stirred by the technological possibility of weapons autonomy in the critical target selection and engagement functions
2. The requirement of MHC over all weapon systems must be stated in a provision of general purport, followed by three Sections on "Training", "Control by design", and "Control in use"
3. The Section on "Training" must spell out state obligations to foster awareness among AWS decision-makers and users about the limits affecting autonomous targeting by weapon systems
4. The Section on "Control by design" must crucially include provisions prescribing that the design and development of weapon systems always comply with interpretability and explainability requirements: to achieve situational awareness for MHC purposes, human operators should be put in a position to get a sufficient amount of humanly understandable information about machine data processing (*interpretability* requirement), and to obtain an account of the reasons why the machine is suggesting or going to take a certain course of action (*explainability* requirement)
5. The "Control in use" part will establish higher levels of human control (L1-L2) as a default policy and the obligation to regulate exceptions thereto by way of suitable rules (as schematized above)
6. Transparency obligations, verification and confidence-building measures will be explicitly included, given their crucial role for the actual implementation and respect of the MHC Convention/Protocol.

What makes human control over weapon systems "meaningful"?

Daniele Amoroso (Università di Cagliari and ICRAC)

and Guglielmo Tamburrini (Università di Napoli Federico II and ICRAC)[*]

1. Introduction

In academic and diplomatic debates about so-called "autonomous weapon systems" (AWS), a watchword has rapidly gained ground across the opinion spectrum: all weapon systems, including autonomous ones, should remain under human control. While references to the human element were already present in early documents on AWS,[1] the UK-based NGO Article 36 must be credited for putting it at the centre of 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 call for a pre-emptive ban on AWS, the MHC formula (and the like) was promptly met with interest by a substantial number of States. This response is explainable by a variety of converging reasons. To begin with, human control is an easily *understandable* concept, which "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 this formula is its *constructive ambiguity*,[4] which may prove helpful to bridge the gap between various positions expressed at the international level on the AWS issue. Third, and finally, the notion at hand allows one to shift the focus of the AWS debate from a recalcitrant definitional problem, i.e. the precise drawing of boundaries between automation and autonomy in weapon systems, to a normative problem, i.e. what kinds and levels of human control ought to be exerted on weapon systems.[5] Unlike the former definitional problem, the latter appears to be more tractable and likely to be successfully addressed through negotiations.[6] In this perspective, it was correctly underlined that one should not look at MHC necessarily as a "solution", as it rather indicates the right "approach" to cope with ethical and legal implications of autonomy in weapon systems.[7]

And indeed, growing attention to the issue of human control emerges from diplomatic talks that have been taking place in Geneva within the Group of Governmental Experts on lethal AWS (GGE) established by the State Parties to the Convention on Conventional Weapons (CCW).[8] In

[*] **Acknowledgments.** We are grateful to Jürgen Altmann for detailed and stimulating comments on a previous version of this report. We wish to thank also Amanda Sharkey, Noel Sharkey, and Steve Wright for their stimulating comments. Parts of this Report also appeared as an ESIL Reflection (volume 8, issue 5 (2019)), published by the European Society of International Law (https://esil-sedi.eu/category/reflections/).

[1] See US Department of Defense (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] Article 36 (2013). For an earlier use of this expression in the context of robotic warfare, see Adams (2001/2002), p. 67.

[3] UNIDIR (2014), p. 3.

[4] Crootof (2016), pp. 58-60; Rosert (2017), pp. 2-3.

[5] Brehm (2015), p. 5. See also UNIDIR (2014), p. 4, as well as, more recently, UK (2018), para. 8.

[6] Brehm (2015), p. 5.

[7] UNIDIR (2014), p. 4.

[8] See the powerful statement by Germany (2019), which deems "the definition of the human role as the single most relevant deliverable of this group and a central element of any outcome document."

addition to the fact that a number of participating State Parties have explicitly endorsed the call for an MHC requirement,[9] the importance of this issue was underscored by most delegations taking part in the CCW proceedings, in both their official speeches and working papers.[10] Such a convergence of views is reflected in the "Possible Guiding Principles" adopted by the GGE at its August 2018 meeting, and notably in Principle 2, which posits that "Human responsibility for decisions on the use of lethal force must be retained […]".[11] This is, however, exactly where international consensus stops. As many commentators pointed out, it is far from settled – *even among those favouring 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.

The present ICRAC Report intends to contribute to move forward the AWS debate (i) by filling the MHC placeholder with more precise contents, and (ii) by identifying on this basis some key aspects of any legal instrument enshrining the MHC requirement (such as, e.g., a Protocol VI to the CCW).

2. The Ethical and Legal Case in Favour of the MHC requirement: an Overview

Most legal and ethical debates about autonomy in weapon systems are based on a shared understanding of AWS critical functions, which are spelled out in a necessary condition on AWS, propounded – with slightly different wordings – by the US Department of Defense (DoD)[12], the International Committee for the Red Cross (ICRC)[13], as well as by the NGOs campaigning for banning AWS[14]: *to count as autonomous, a weapon system must be able to select and engage targets without human intervention.*[15]

[9] See, e.g., Austria, Brazil, and Chile (2018).

[10] Indeed, at both the 2018 and 2019 GGE meetings an agenda item has been specifically devoted to the "consideration of the human element in the use of lethal force". See Singh Gill (2018), 6(b) and Gjorgjinski (2019), 5(b). It is remarkable, in this respect, that the five permanent members of the UN Security Council, while not necessarily converging on the MHC formula, agreed on the need to keep humans involved to a certain extent. See, in particular, the views expressed by China (2018), para. 3 ("Discussions on Human-Machine Interaction should […] define the mode and degree of human involvement and intervention. Concepts such as meaningful human control and human judgment are rather general and should be further elaborated and clarified"; France (2018) para. 13 ("The use of force remains an inherent responsibility of human command, particularly in cases of violations of international humanitarian law"); Russian Federation (2018), para. 11 ("We do not doubt the necessity of maintaining human control over the machine"); United Kingdom (2018), para. 6 ("Operation of our weapons will always be under human control as an absolute guarantee of human oversight and authority, and of accountability for weapons usage"). As to the US official position, see note 1 above.

[11] GGE (2018), para. 21(b).

[12] US Department of Defense (2012), pp. 13-14.

[13] ICRC (2016), p. 1.

[14] Campaign to Stop Killer Robots (2013).

[15] An alternative definition is the one propounded by the UK Ministry of Defence in its Joint Doctrines on unmanned aircraft systems. See, most recently, UK Ministry of Defence (2017), p. 13, where autonomous systems are defined as follows: "An autonomous system is capable of understanding higher-level intent and direction. From this understanding and its perception of its environment, such a system is able to take appropriate action to bring about a desired state. It is capable of deciding a course of action, from a number of alternatives, without depending on human oversight and control, although these may still be present. Although the overall activity of an autonomous unmanned aircraft will be predictable, individual actions may not". This definition, which projects AWS in some undetermined technological future has been stigmatized by the House of Lords Select Committee on Artificial Intelligence as "out of step" with that generally agreed upon at the international level. See UK House of Lords Select Committee (2018), para. 345. See also, for further discussion, Amoroso and Tamburrini (2017), pp. 3-4.

An MHC requirement over weapon systems would be aimed at curbing the latter's autonomy exactly in the way of their critical target selection and engagement functions.[16] This is the main reason why the above necessary condition provides an adequate starting point for the ensuing discussion on the motives for and the contents of an MHC requirement. By the same token, it is no surprise that the arguments supporting MHC largely coincide with those substantiating the case for a ban on AWS. This pro-MHC and pro-ban convergence notably emerges from the following three arguments.[17]

First, AWS would be unable to comply with the principles of distinction, proportionality, and precaution embedded into International Humanitarian Law (IHL). The development of AWS fulfilling distinction and proportionality requirements at least as well as a competent and conscientious human soldier presupposes the solution of many profound research problems in artificial intelligence (AI) and advanced robotics.[18] Furthermore, it is 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, insofar as the regular behaviour of AI and robotic systems is perturbed by unpredicted dynamic changes occurring in warfare environments. Notably, systems developed by means of advanced machine learning technologies (e. g. deep learning) have been extensively demonstrated by adversarial testing to be prone to unexpected, counter-intuitive and potentially catastrophic mistakes, which a human operator would easily detect and avoid.[19]

Second, AWS are likely to determine an accountability gap.[20] One cannot exclude that AWS will make targeting decisions that, were they taken by human agents, would trigger individual criminal responsibility. Who will be held responsible for this conduct? The list of potentially responsible persons in the decision-making chain includes the military commander in charge and those overseeing the AWS operation, in addition to manufacturers, robotics engineers, software programmers, and those who conducted the AWS weapons review. People in this list may cast their defence against responsibility charges and criminal prosecution in terms of their limited decision-making roles, as well as of the complexities of AWS systems and their unpredictable behaviour in the battlefield.[21] Cases may occur where it is impossible to ascertain the existence of the mental element (intent, knowledge or recklessness), which is required under International Criminal Law (ICL) to ascribe criminal responsibilities. Consequently, no one would be held criminally liable, notwithstanding the conduct at stake materially amounts to an international crime. This outcome is hardly reconcilable with the moral duty of military commanders and operators to be accountable for their own actions, as well as with the related principle of individual criminal responsibility under ICL.

Third, AWS would run counter to the principle of human dignity, which would dictate that decisions affecting the life, physical integrity and property of human beings involved in an armed conflict should be entirely reserved to human operators and cannot be entrusted to an autonomous artificial

[16] Bhuta, Beck and Geiss (2016), p. 381.

[17] For a first, comprehensive exposition of the ethical and legal problems at stake, see Heyns (2013), paras. 63-81, 89-97. See also, more recently, Amoroso et al. (2018).

[18] This is acknowledged also by roboticists who, in principle, are in favour of autonomy in weapons systems. See Arkin (2009), pp. 211-212 (listing the "daunting problems" to be addressed in order to develop an IHL-compliant AWS).

[19] Szegedy et al. (2014). See also, with specific reference to AWS, UNIDIR (2018), p. 9; ICRC (2019a), p. 11.

[20] This problem was already flagged by Sparrow (2007). For an updated discussion, as well as for further references, see Amoroso and Giordano (2019).

[21] It is worth noting that AWS' unpredictability has many sources. In particular, it does not depend solely upon the complexity of the system's architecture and the AI built into it, but also on the features of the AWS operational environments. See Tamburrini (2016), pp. 127-128.

agent. Otherwise, people subject to AWS' use of force would be placed in a position where any appeal to the shared humanity of persons placed on the other side – and thus their inherent value as human beings – would be a priori and systematically denied.[22]

Each one of these arguments against machine autonomy in the critical functions of selecting and engaging targets is *deontological* in character: an appeal to moral duties (IHL-embedded duty to protect the non-combatants, duty to preserve human responsibilities, duty to respect human dignity) is made there to guide the use of weapon systems and to assess the moral worth of deploying AWS.

In addition to these deontological arguments, there are *consequentialist* arguments questioning the moral worth of deploying AWS in the light of the expected negative repercussions on global peace and stability. Consequentialist arguments have played a central role in the debate concerning the ethical and legal acceptability of AWS.[23] Consequentialist appraisals of AWS, however, are comparatively less relevant than the above deontological arguments in connection with the main problem that we are concerned with, that is, *the problem of shaping the contents of MHC*. An analysis of their relationship with the MHC requirement, therefore, is postponed to the concluding section.

3. Shaping the Content of the (Meaningful) Human Control Requirement

The foregoing ethical and legal reasons go a long way towards shaping the content of MHC, by pinpointing functions that are prescriptively assigned to human control and by providing criteria that enable one to distinguish perfunctory from truly meaningful human control. In particular, the arguments above suggest a threefold role for human control on weapon systems to be "meaningful". First, human control must afford a *fail-safe mechanism*, which is meant to prevent a malfunctioning of the weapon from resulting in a direct attack against the civilian population and objects, or in excessive collateral damages.[24] Second, it must serve as a *catalyst for accountability*, insofar as it secures the legal conditions for responsibility ascription in case a weapon follows a course of action that is in breach of international law.[25] Third, it must ensure that it is a *moral agent*, and not an artificial one, that takes decisions affecting the life, physical integrity and property of people (including combatants) that are involved in an armed conflict.[26] To be ethically and legally sound, therefore, rules aimed at determining the obligations flowing from the MHC requirement should guarantee that all these functions are properly fulfilled.

The preservation by means of human agency of these various properties in relation to increasingly autonomous weapon systems requires one to address and solve two crucial problems: *i)* how to ensure a proper *quality* of human involvement; *ii)* how to properly establish *exclusive control privileges for human operators*. In this respect, it is important to note that neither problem is properly addressed by considering only actual battlefield deployment and use, as humans are involved in different capacities throughout "the entire life cycle of the weapon system",[27] which includes training of

[22] See, among others, Asaro (2012), p. 689; Heyns (2016); and Sharkey (A.) (2019)

[23] On this point, see extensively Altmann and Sauer (2017). See also Tamburrini (2016), and Amoroso and Tamburrini (2017), pp. 9-12.

[24] Scharre (2017), 154.

[25] Chengeta (2017).

[26] ICRC (2018), paras. 23-26.

[27] GGE (2018), para. 21(b).

commanders and operators, research and development (R&D), as well as the weapon testing, evaluation and certification (T&E).

A. The Quality of Human Involvement: Training and Design

As it will be seen in Section 3.B, it is possible to identify a set of hypotheses where autonomy in weapon systems is not irreconcilable with the MHC requirement. It will be shown, however, that also in those hypotheses human operators will retain some exclusive control privileges, and notably the power to approve or veto the machines' targeting decisions. Should that be the case, it is all the more important to ensure the quality of human involvement, so that residual control privileges are exercised in a meaningful way.

To begin with, military personnel training should foster awareness of both ascertained and likely limits in the proper autonomous functioning of weapon systems, and related human predicaments in the capability to predict and control their behaviour.[28] Well-known performance degradation factors originate in task environment changes that are difficult to model, notably including unpredicted competitive interactions with an adversary's autonomous artificial agents, be them other AWS endowed with kinetic capabilities or software agents performing cyberattacks.[29] Awareness building efforts concerning limitations in AWS proper functioning should be part of more encompassing training efforts, whereby the military personnel are trained to use advanced technologies without forfeiting human judgement and critical sense, and without succumbing to so-called automation biases.[30]

If humans are expected not to blindly trust the machine, moreover, they should be put in a position to get a sufficient amount of humanly understandable information about machine data processing (*interpretability* requirement), so as to provide operators with this crucial sort of situation data, and to achieve adequate situation awareness on this account too; and to additionally obtain an account of the reasons why the machine is suggesting or going to take a certain course of action (*explainability* requirement). Both interpretability and explainability requirements concern the *design* of weapon systems and must be addressed by R&D and T&E teams.[31]

To fulfil the interpretability requirement, one should map machine data and information processing into domains that humans can make sense of.[32] Accordingly, AWS should be designed so as to provide commanders and operators with "access to the sources of information" handled by the system,[33] in a way that allows humans to take in and process data "at the level of meaning", rather than "in a purely syntactic manner".[34] The latter would occur if users were offered only a representation of the battlefield they could not reliably take in, without giving them any element to understand the situational features that are crucial to play the fail-safe, accountability and moral

[28] Margulies (2017), p. 441.

[29] ICRAC (2015), point 10.

[30] On which see Cummings (2006); Sharkey (N.) (2016), pp. 32-33; ICRAC (2018), pp. 2-3. A tragical example of the backlashes of automation-biases on the use of weapons systems is offered by the sadly known Patriot fratricides, on which see Scharre (2018), pp. 139-140.

[31] This point is well captured by the formula "control by design", coined by the International Panel on Regulation of Autonomous Weapons (iPRaW) (2018), p. 14.

[32] Montavon, Samek, and Müller (2018).

[33] Breton and Bossé (2003), pp. 10-11.

[34] Hew (2016), p. 230.

agency roles.[35] One may recall, in this regard, the well-known example made by Article 36 of a human "simply pressing a 'fire' button in response to indications from a computer".[36] In other words, military technological advances should empower human combatants, by enhancing their situational awareness, rather than substituting artificial agents for human understanding and judgement.[37]

To fulfil the explainability condition, AWS should be equipped with the capability to provide explanations of courses of actions that are being suggested or undertaken.[38] On account of the interpretability requirement, these explanations must be cast in terms that are cognitively accessible to human users.[39] Meeting the explainability requirement might prove particularly demanding in relation to weapon systems endowed with machine-learning capabilities. Indeed, currently used learning technologies are often based on sub-symbolic data representations and other information processing methods that are not transparent to human users. Notably, deep neural networks are currently achieving previously unmatched algorithmic classification and decision-making results, but are mostly unable to fulfil interpretability and explainability requirements.[40] The development of AI systems that are capable of providing humanly understandable explanations for their decisions and actions is the focus of the rapidly expanding XAI (eXplainable AI) research area. Scientifically challenging issues in XAI are, by no coincidence, central themes of research programs supported by the US Defense Advance Research Project Agency (DARPA).[41] Pending significant breakthroughs in XAI, one cannot but acknowledge the present technological difficulty of ensuring sufficient levels of system interpretability and explainability that are necessary to establish MHC on AI-based weapon systems. Weapon systems not meeting these conditions are incompatible with the proper exercise of MHC as spelled out here.[42]

B. Exclusive Control Privileges for Human Operators

The aforementioned training and design requirements should not be deemed as a "panacea", able to address adequately all the normative concerns regarding human control over weapon systems.

[35] On a technical level, a significant move towards empowering human judgment would be to mount on weapons systems not only sensors that "apply algorithms to the measurements to infer the existence and properties of objects" (object-inferencing sensors), but also devices that "extend the range of a human's organic sensors" (extended human sensors). Hew (2016), p. 229.

[36] Article 36 (2016), p. 2.

[37] US Air Force Chief Scientist (2015), p. 8.

[38] It should be underscored that the explainability requirement must be read in conjunction with the need to retain exclusive control privileges for human operators (see below Section 3.B). The explanations provided by the machine, indeed, are aimed at enhancing the human operator's control when she chooses among the alternatives suggested by the weapon, as well as – in the hypotheses where lower levels of human control are allowed – when she approves the targeting decision suggested by the weapon or supervises its targeting activities. They should in no way be understood as a means to legitimize, in general terms, autonomous targeting by weapons systems.

[39] See US Department of Defense (2012), requiring the interface between people and machines to "be readily understandable to trained operators" (para. 4.3.a).

[40] Holzinger et al. (2017).

[41] https://www.darpa.mil/program/explainable-artificial-intelligence accessed 20 July 2019.

[42] Admittedly, there is one hypothesis where the fulfilment of the interpretability and explainability requirements could prove particularly demanding. Reference is made to the case of autonomous defensive systems tasked to protect inhabited vehicles and buildings from sudden and/or saturation attacks, in relation to which a split-second response may be required, i.e. one that is incompatible with human reaction times. In this case, it could be expedient to think of a specific regime that, while requiring special design efforts (e.g. graphic representation of incoming objects, with different colours representing the degree of certainty/similarity to known projectiles to alert operator about possible changes), takes into due account that certain defensive functions may be beyond the limits of human-following.

As well underlined by Estonia and Finland, “the most critical point”, in this respect, remains “the final interaction between a human [and] a weapon system before that system delivers force”,[43] which leads us to the core of the MHC conundrum, viz. the identification of exclusive control privileges for human operators, whose assignment to human beings only is ethically and legally justified.

Several attempts have so far been made – by scholars, states, and NGOs – to define the human-weapon shared control policies dictated by the MHC requirement. While significantly different from each other, these various proposals are generally affected by a common weakness: they aspire to capture optimal partnership with one formula, which is supposed to apply uniformly to all kinds of weapon systems and to each one of their possible uses.[44] This flaw is particularly evident as regards the so-called “wider loop” approach, advocated by the Dutch government, whereby MHC would be in fact exerted by human commanders at the planning stage of the targeting process.[45]

The Dutch approach, indeed, may have limited applicability and relevance with regard to *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, a largely unhelpful approach with regard to *dynamic* targeting, which pursue targets of opportunity. To the extent that it warrants the weapon with humanly unrestrained autonomy after deployment, moreover, the Dutch approach appears to be deeply problematic when the weapon is activated in a scenario populated by civilians, in that it drives a wedge between the State owing a duty of care towards the civilian population and the actual possibility to comply with that duty by influencing the course of events through its agents.[46] This is especially true for AWS endowed with the capability of “loitering” for sustained periods of time in search of enemy targets,[47] for the conditions licensing the activation of a loitering AWS by human operators may rapidly change in many warfare scenarios characterized by erratic dynamics and surprise-seeking behaviors.

The overly permissive guideline sketched 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 overly restrictive endeavours to define the MHC requirement in rigorous terms and in an all-encompassing manner, so as to prescribe a uniformly applicable form of human control over every and each kind of weapon system and use thereof. While undoubtedly praiseworthy for their attention to humanitarian concerns, these attempts may prove inadequate in that they run the risk of banning some weapons (i) whose lawfulness has been so far gone undisputed,[48] and (ii) for which milder forms of human control might be equally able to “purify” the autonomy of weapon systems of its ethically and legally troubling implications, however only in certain limited operational environments. Significant cases in point are the already deployed US Phalanx,[49] Israeli Iron Dome[50]

[43] Estonia and Finland (2018), para. 17. See also Brazil (2019): “The concepts of control by design and control in use can be useful […]. Notwithstanding, the brunt of IHL obligations relates to the use phase, particularly to the decision to engage, and so it must remain as the primary reference for further work”.

[44] US (2018a), para. 9 (“there is not a fixed, one-size-fits-all level of human judgment that should be applied to every context”).

[45] See AIV/CAVV (2015). See also Roorda (2015) and Ekelhof (2018).

[46] Akerson (2013), p. 87. See also the explicit criticism by the ICRC, according to which “concepts of ‘human control in the wider loop’ and the use of autonomy to ‘effectuate the intention of commanders and the operators of weapons systems’ do not adequately capture the requirement for human control under IHL” (ICRC (2019b), p. 3).

[47] A well-known sample of loitering munition is the Harpy NG system, produced by the Israel Aerospace Industries (https://www.iai.co.il/p/harpy accessed on 20 July 2019).

[48] Horowitz and Scharre (2015), pp. 9-10.

[49] https://www.army-technology.com/news/raytheon-us-phalanx-land-based-weapon-system/ accessed 20 July 2019.

[50] https://www.army-technology.com/projects/irondomeairdefencemi/ accessed 20 July 2019.

and the German *Nächstbereichschutzsystem* (NBS) MANTIS,[51] when both are used as intended, that is, as protective shields from incoming shells and missiles. As we shall point out below, a reflection on these systems suggests that, in some very limited circumstances, a proper exercise of MHC is not incompatible with the autonomy of weapon systems.

In brief, along with humanitarian concerns and motivated restrictions, also military motivations for granting limited forms of critical autonomy to weapon systems should be taken into consideration, as long as the latter do not jeopardize the fail-safe, accountability, and moral agency properties that we have identified above as core components of the MHC requirement. To this end, we suggest giving up the quest for a one-size-fits-all solution to the issue of MHC in favour of a suitably differentiated approach, which is nonetheless based on the unifying ground provided by the converging ethical and legal principles outlined above. In our view, notably, the application of these overarching principles in concrete situations must be facilitated and given concrete operational content by the formulation of a set of rules bridging the gap between ethical and legal principles on the one hand, and specific sorts of weapon systems and their concrete uses on the other hand. These "if-then" bridge rules should be able to express the fail-safe, accountability, and moral agency conditions for exercising *in context* a genuinely MHC over weapon systems.

The "if-part" of these rules should include properties concerning *what* mission the weapon system is involved into, *where* it will be deployed and *how* it will perform its tasks.[52] The "what-properties", in particular, must relate to the 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 inhabited military objects vs. uninhabited vehicles and military objects). The "where-properties" must concern the dynamical features of the operational environment, including interactions with an adversary's autonomous artificial agents, and having special regard to the presence/absence of civilians, civilian objects and friendly forces. The "how-properties", finally, must regard the information-processing and sensory-motor capabilities that the system puts at work to carry out its mission and that may affect its overall controllability and predictability. Learned decision-making and "swarm intelligence"[53] abilities, which may be increasingly implemented on future AWS, jointly with loitering capabilities of existing or developing weapon systems, are significant cases in point of how-properties that raise serious concern from an MHC perspective.[54]

The "then-part" of bridge rules should establish what kind of human-machine shared control would be legally required on each single use of a weapon system. Following a taxonomy proposed by Noel Sharkey (only slightly modified below),[55] one may sensibly consider five basic types of human-machine interaction for the "then-part" of bridge rules, ordered according to decreasing levels of human control and increasing levels of machine control in connection with the critical target selection and engaging tasks:

L1. A human engages with and selects targets, and initiates any attack;

[51] https://www.army-technology.com/projects/mantis/ accessed 20 July 2019. See also

[52] For an analysis of these properties in the different context of autonomous robotic surgery, see Ficuciello, Tamburrini, Arezzo, Villani and Siciliano (2019).

[53] See, in this respect, the Pentagon's Perdix Project (https://dod.defense.gov/News/News-Releases/News-Release-View/Article/1044811/department-of-defense-announces-successful-micro-drone-demonstration/ accessed 20 July 2019).

[54] For a brief account of the legal problems raised by machine-learning and loitering technologies see above the text accompanying footnotes 40-41 and 46-47, respectively. In relation to swarm technology, see ICRC (2019c), p. 3.

[55] Sharkey (N.) (2016), pp. 34-37; ICRAC (2018), pp. 3-4. Deviations concern, notably, levels 4 and 5.

L2. A program suggests alternative targets and a human chooses which to attack;

L3. A program selects targets and a human must approve 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 attack on the basis of the mission goals as defined at the planning/activation stage, without further human involvement.

Against the background of L1-L5, the gist of our differentiated approach to MHC is specified by means of (i) a general default policy and (ii) exceptions formulated as specific bridge rules. In the light of the ethical and legal arguments for MHC examined above, we suggest as a general default policy that the higher levels of human control (L1 and L2) be exerted. Under this proviso, *lower levels of human control may become acceptable only as internationally agreed on exceptions, clearly formalized as specific bridge rules*. These bridge rules should establish what level is required to grant the fulfilment of a genuinely *meaningful* human control, as well as the values of the what/where/how properties (or combinations thereof) that justify the identification of some specific level in the above list.

Any deviation from the general default policy should be crafted for specific types and models of weapon systems (and use thereof), and should take into account at least the following observations:

1. Deliberate targeting (*what-property*) by AWS may be pursued at a one-step lower level of human control (L3), since targeting decisions have actually been taken by humans at the planning stage: the human operator, therefore, has only to confirm that there have not been changes in the battlespace that may affect the lawfulness of the operation. The same level should be required, *as a minimum*, in relation to AWS programmed to engage human or humanly inhabited targets in structured scenarios, e.g. high seas or deserts, where civilians and civilian objects are not present (*where-property*),[56] so that it is granted that there is a human on the attacking end who can verify, in order to deny or grant approval, whether there are persons *hors de combat* and take appropriate measures accordingly. An example of a (potential) AWS used in a structured environment is the legacy South Korean robotic sentry SGR-A1, deployed on the South side of the Korean demilitarized zone, which is reportedly able to function in full autonomous mode, although it seems that it has not been (yet) activated in this mode.[57]

2. The (L4) human supervision and veto level might be deemed as an acceptable level of control only in case of AWS with exclusively defensive functions (*what-property*). This is the case of the US Phalanx, the Israeli Iron Dome and the German *Nächstbereichschutzsystem* (NBS) MANTIS, according to their customary uses as protective shields from incoming shells and rockets for human-inhabited sites or vehicles.

3. The use of capabilities that may reduce the overall predictability of the weapon systems' behaviour, such as loitering, learned decision-making, swarming (*how-properties*), should always be treated as a compelling factor pushing towards the application of higher levels of human control (L1 and L2).

[56] An example of a CCW provision based on "where-properties" is Art. 2(2) of the Protocol III on Prohibitions or Restrictions on the Use of Incendiary Weapons (10 October 1980), which prohibits "in all circumstances to make any military objective located within a concentration of civilians the object of attack by air-delivered incendiary weapons".
[57] Kumagai (2007).

The (L5) full autonomy level should be considered incompatible with the MHC requirement. While it is true that operational constraints set at the planning and/or activation stages may play an important role in limiting weapons' autonomy,[58] such a "boxed autonomy" alone is not enough to ensure MHC,[59] unless the operational space and time frames are so strictly circumscribed to make targeting decisions entirely and reliably traceable to human operators.

This point can be better illustrated by referring to homing fire-and-forget munitions, i.e. a kind of precision-guided munition which use passive sensors and/or active seekers to track onto moving targets. Most fire-and-forget munitions simply lock on targets pre-selected by human operators,[60] so falling under the highest level (L1) of human control. The most sophisticated among them can be fired in a salvo to simultaneously attack multiple targets moving close to each other, which were previously identified by a human operator, by coordinating among them in order to avoid hitting the same target.[61] In this case, each munition could be said *prima facie* to select and engage targets on its own, based on the mission goals defined by the human operator, viz. a (L5) level of human control. At closer scrutiny, however, the mission is defined at such level of detail (e.g. "destroy *that specific* line of enemy tanks") that nobody could question the conclusion that those targeting decisions are collectively attributable to the human launching the attack and thus are under the (L1) level of human control.

The same conclusion does not hold for fire-and-forget munitions – like the UK Brimstone – when these are endowed with the capability of searching for and hitting targets within a "kill box" designated by the 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),[62] 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 and not by the human user. While this is admittedly a borderline case, to be treated with pondered 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 UK Brimstone is particularly instructive. Indeed, a previous version of the Brimstone, operating solely in the "autonomous" mode, was deemed incompatible with the rules of engagement of the Afghanistan campaign, which prompted the UK Royal Air Force to issue in 2007 an Urgent Operation Requirement aimed at modifying the existing missiles in order to enable the human selection of targets through laser guidance (in the words of the manufacturer, "a man-in-the-loop capability").[63] Contrary to what is sometimes maintained, therefore, the ethical and legal acceptability of offensive autonomy – even in such a limited form – is far from being undisputed, even among the primary users of these weapon systems.

4. Conclusions

At the August 2018 meeting of the GGE on lethal AWS, the delegations of Austria, Brazil and Chile jointly submitted a proposal for a mandate to "negotiate a legally binding instrument to ensure

[58] ICRC (2019c), p. 4.

[59] On the notion of "boxed autonomy", see iPRaW (2017), pp. 15-16.

[60] Scharre (2018), p. 106.

[61] Boulanin and Verbruggen (2017), p. 50.

[62] Scharre (2018), p. 107.

[63] MBDA (2009). This led to the current dual-mode version of the Brimstone missile.

meaningful human control over critical functions in lethal autonomous weapon systems".[64] One may doubt that this proposal will be promptly followed up within the institutional framework of the CCW, in light of the fact that some major military powers, including the US, have been opposing a solution of this kind. At the same time, however, the present proposal of relinquishing the quest for a one-size-fits-all solution to the MHC issue in favour of a suitably differentiated approach may contribute to sidestep present stumbling blocks at the CCW and other international policy venues. Indeed, diplomatic and political discontent about an MHC requirement which appears to be overly restrictive with respect to the limited autonomy of some weapon systems (such as Phalanx, Iron Dome and MANTIS) might be mitigated by recognizing the possibility of negotiating exceptions to L1-L2 human control as long as one is able to identify weapon systems and contexts of use in which milder forms of human control will do. In any case, it cannot be ruled out that, at some point in the near future, concerned states will explore alternative venues to negotiate an international agreement establishing the MHC requirement, as it already occurred with the Anti-Personnel Mine Ban Convention.[65]

It seems therefore appropriate to begin thinking about the content of such – for now wholly hypothetical – treaty.[66] In the light of the foregoing, we suggest that an "MHC Convention/Protocol" should have, as a minimum, the following contents:

1. The essential elements of the ethical and legal concerns outlined in Section 2 must be included in the Preamble, so as to provide the "context" for the interpretation of the treaty under Article 31(2) of the 1969 Vienna Convention on the Law of Treaties.
2. The requirement of MHC over all weapon systems must be stated in a provision of general purport. The content of this requirement should then be clarified in three ensuing parts, concerning "Training", "Control by design" and "Control in use" respectively.[67]
3. Provision(s) on "Training" must spell out State obligations to foster awareness among AWS decision-makers and users about the limits affecting autonomous targeting by weapon systems, and to train them to preserve critical sense and countervail risks of so-called automation bias.[68]
4. The "Control by design" part must include provisions prescribing design and development compliance of weapon systems with interpretability and explainability requirements, as set out in Section 3.A. Technical specifications for these requirements might be detailed in some specific Annex.
5. The "Control in use" part will be undoubtedly the more important and challenging to agree upon. Our suggestion is to establish higher levels of human control (L1-L2) as a default policy and regulate exceptions thereto by way of bridge rules like those suggested in Section 3.B. In this way, one relinquishes the quest for a one-size-fits-all solution to the MHC issue in favour of a suitably differentiated approach, which is nonetheless based on the unifying ground provided by the above converging ethical and legal principles.
6. Crucial to the actual implementation of the MHC requirement will be the introduction of transparency obligations, as well as of verification and confidence-building measures. This aspect cannot be addressed here in detail.[69] However, the analysis carried out in this Report provides some indications as to *what* information State Parties should share with the others.

[64] Austria, Brazil and Chile (2018)
[65] Campaign to Stop Killer Robots (2018).
[66] For a first, interesting attempt in this sense, see Homayounnejad (2018).
[67] With regard to the latter two terms, see iPRaW (2018).
[68] The provision of a training obligation would not be a novelty within the CCW system. See, for instance, Art. 2 of the Protocol IV on Blinding Laser Weapons (13 October 1995).
[69] In this regard, see the valuable insights by Knuckey (2016).

> For instance, States might be required to communicate to the other Parties the weapon systems in relation to which they wish to adopt a control policy different from the default policy and on what basis they consider applicable one of the exceptions set forth in the bridge rules. Furthermore, as suggested by Gubrud and Altmann, State Parties might have to be obliged to secure "records of each engagement", by making them "available to a Treaty Implementing Organization, on request, when sufficient evidence exists to support suspicions of illegal autonomous operation".[70]

Before concluding, a final remark is in order, which concerns the relationship between the outlined contents of MHC and the substantive stability and peace concerns raised by AWS. While ultimately based on deontological reasons, the MHC requirement – at least as set out in this contribution – would prove effective also from the consequentialist perspective in normative ethics that was mentioned at the end of section 2. 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 which is incompatible with the limitations of the human cognitive and sensory-motor coordination system.[71]

[70] Gubrud and Altmann (2013), p. 2.

[71] Altmann and Sauer (2017).

References

Adams, T.K. (2001/2002), 'Future warfare and the decline of human decisionmaking', *Parameters*, **31**(4), pp. 57-71

Advisory Council on International Affairs (AIV) and Advisory Committee on Issues of Public International Law (CAVV) (2015), *Autonomous weapon systems: the need for meaningful human control*, No. 97 AIV / No. 26 CAVV

Akerson, D. (2013), 'The Illegality of Offensive Lethal Autonomy', in D. Saxon (ed), *International Humanitarian Law and the Changing Technology of War*, Leiden, the Netherlands and Boston, US: Martinus Nijhoff, pp. 65-98

Altmann, J. and F. Sauer (2017), 'Autonomous Weapon Systems and Strategic Stability', *Survival*, **59**(5), pp. 117-142

Amoroso, D., F. Sauer, N. Sharkey, L. Suchman and G. Tamburrini (2018), *Autonomy in Weapon Systems. The Military Application of Artificial Intelligence as a Litmus Test for Germany's New Foreign and Security Policy*, Berlin: Heinrich Böll Foundation

Amoroso, D. and B. Giordano (2019), 'Who Is to Blame for Autonomous Weapons Systems' Misdoings?', in N. Lazzerini and E. Carpanelli (eds), *Use and Misuse of New Technologies. Contemporary Challenges in International and European Law*, The Hague, the Netherlands: Springer, pp. 211-232

Amoroso, D. and G. Tamburrini (2017), 'The Ethical and Legal Case Against Autonomy in Weapons Systems', *Global Jurist*, **17**(3), pp. 1-20

Arkin, R. (2009), *Governing Lethal Behavior in Autonomous Robots*, Boca Raton, FL, US: CRC Press

Article 36 (2013), *Killer Robots: UK Government Policy on Fully Autonomous Weapons*, policy paper

Article 36 (2016), *Key elements of meaningful human control*, Background paper working paper submitted to the Group of Governmental Experts on lethal autonomous weapons of the CCW, Geneva, April

Asaro, P. (2012), 'On banning autonomous weapon systems: human rights, automation, and the dehumanization of lethal decision-making', *International Review of the Red Cross*, **94**, 687-709

Austria, Brazil, and Chile (2018), '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)', submitted to the Group of Governmental Experts on lethal autonomous weapons of the CCW, Geneva, 30 August (UN Doc. CCW/GGE.2/2018/WP.7)

Bhuta, N, S. Beck and R. Geiss (2016), 'Present futures: concluding reflections and open questions on autonomous weapons systems', in N. Bhuta et al. (eds), *Autonomous Weapons Systems: Law, Ethics, Policy*, Cambridge, UK: Cambridge University Press, pp. 347-383

Boulanin, V. and M. Verbruggen (2017), 'Mapping the Development of Autonomy in Weapon Systems', SIPRI Report, Solna, November

Brazil (2019), 'Statement on Agenda Item 5(d)', delivered at the Group of Governmental Experts on lethal autonomous weapons of the CCW, Geneva, March

Brehem, M. (2015), 'Meaningful human control', paper presented to the Informal meeting of experts on lethal autonomous weapons systems of the Convention on Certain Conventional Weapons (CCW), Geneva, 14 April

Breton, R. and E. Bossé (2013), 'The Cognitive Costs and Benefits of Automation', in NATO (ed), *The Role of Humans in Intelligent and Automated Systems*, RTO Meeting Proceedings MP-088, pp. 1-11

Campaign to Stop Killer Robots (2013), 'Urgent Action Needed to Ban Fully Autonomous Weapons. Non-governmental organizations convene to launch Campaign to Stop Killer Robot', press release, London, 23 April

Campaign to Stop Killer Robots (2018), 'Fragile diplomatic talks limp forward', press release, 23 November

Chengeta, T. (2017), 'Defining the Emerging Notion of "Meaningful Human Control" in Autonomous Weapon Systems', *New York Journal of International Law & Politics*, **49**, pp. 833-890

China (2018), 'Position paper', submitted to the Group of Governmental Experts on lethal autonomous weapons of the CCW, Geneva, 11 April (UN Doc. CCW/GGE.1/2018/WP.7)

Crootof, R. (2016), 'A Meaningful Floor for "Meaningful Human Control"', *Temple Journal of International & Comparative Law*, **30**, pp. 53-62

Cummings, M.L. (2006), 'Automation and Accountability in Decision Support System Interface Design', *Journal of Technology Studies*, **32**(1), pp. 23-31

Ekelhof, M.A.C. (2018), 'Lifting the Fog of Targeting: "Autonomous Weapons" and Human Control through the Lens of Military Targeting', *Naval War College Review*, **71**(3), Article 6

Estonia and Finland, 'Categorizing lethal autonomous weapons systems – A technical and legal perspective to understanding LAWS', working paper submitted to the Group of Governmental Experts on lethal autonomous weapons of the CCW, Geneva, 24 August (UN Doc. CCW/GGE.2/2018/WP.2)

Ficuciello, F., G. Tamburrini, A. Arezzo, L. Villani and B. Siciliano (2019), 'Autonomy in surgical robots and its meaningful human control', *Paladyn Journal of Behavioral Robotics*, **10**, pp. 30-43

France (2018), 'Human-Machine Interaction in the Development, Deployment and Use of Emerging Technologies in the Area of Lethal Autonomous Weapons Systems', working paper submitted to the Group of Governmental Experts on lethal autonomous weapons of the CCW, Geneva, 28 August (UN Doc. CCW/GGE.2/2018/WP.3)

Germany (2019), 'Statement on Agenda Item 5(b)', delivered at the Group of Governmental Experts on lethal autonomous weapons of the CCW, Geneva, 26 March

Gjorgjinski, L.J. (2019), 'Provisional agenda', submitted to the Group of Governmental Experts on lethal autonomous weapons of the CCW, Geneva, 8 February (UN Doc. CCW/GGE.1/2019/1)

Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems (GGE) (2018), 'Report of the 2018 session', Geneva, 23 October (UN Doc. CCW/GGE.1/2018/3)

Gubrud, M. and J. Altmann (2013), 'Compliance Measures for an Autonomous Weapons Convention', ICRAC Working Paper #2, May

Heyns, C. (2013), 'Report of the Special Rapporteur on extrajudicial, summary or arbitrary executions', 9 April (UN Doc. A/HRC/23/47)

Heyns, C. (2016), 'Autonomous weapons systems: living a dignified life and dying a dignified death', in N. Bhuta et al. (eds), *Autonomous Weapons Systems: Law, Ethics, Policy*, Cambridge, UK: Cambridge University Press, pp. 3-19

Hew, P.C. (2016), 'Preserving a combat commander's moral agency: The Vincennes Incident as a Chinese Room', *Ethics and Information Technology*, 18, pp. 227-235

Holzinger, A., M. Plass, K. Holzinger, G. Cerasela Crisan, C.-M. Pintea and V. Palade (2017), 'A glass-box interactive machine learning approach for solving NP-hard problems with the human-in-the-loop', *arxiv.org*, 3 August

Homayounnejad, M. (2018), 'Some Thoughts on Negotiating a Treaty on Autonomous Weapon Systems', OpinioJuris, 3 January

Horowitz, M. and P. Scharre (2015), 'Meaningful Human Control In Weapon Systems: A Primer', working paper of the Center for a New American Security (CNAS), March

International Committee for Robot Arms Control (ICRAC) (2015), 'LAWS: Ten Problems For Global Security', Memorandum for delegates at the Convention on Certain Conventional Weapons (CCW) Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS), Geneva, April

ICRAC (2018), 'Guidelines for the human control of weapons systems', working paper submitted to the Group of Governmental Experts on lethal autonomous weapons of the CCW, Geneva, April

International Committee of the Red Cross (ICRC) (2016), 'Views of the International Committee of the Red Cross (ICRC) on autonomous weapon system', paper submitted to the Informal meeting of experts on lethal autonomous weapons systems of the Convention on Certain Conventional Weapons (CCW), Geneva, 11 April

ICRC (2018), 'Ethics and autonomous weapon systems: An ethical basis for human control?', working paper submitted to the Group of Governmental Experts on lethal autonomous weapons of the CCW, Geneva, 29 March (UN Doc. CCW/GGE.1/2018/WP.5)

ICRC (2019a), 'Artificial intelligence and machine learning in armed conflict: A human-centred approach'. Report, Geneva, 6 June

ICRC (2019b), 'Statement on Agenda Item 5(a)', delivered at the Group of Governmental Experts on lethal autonomous weapons of the CCW, Geneva, March

ICRC (2019c), 'Statement on Agenda Item 5(b), delivered at the Group of Governmental Experts on lethal autonomous weapons of the CCW, Geneva, March

International Panel on Regulation of Autonomous Weapons (iPRaW) (2017), 'Technology and Application of Autonomous Weapons', Report No. 1, August

iPRaW (2018), 'Focus on the Human-Machine Relation in LAWS', Report No. 3, March

Knuckey, S. (2016), 'Autonomous weapons systems and transparency: towards an international dialogue', in N. Bhuta et al. (eds), *Autonomous Weapons Systems: Law, Ethics, Policy*, Cambridge, UK: Cambridge University Press, pp. 164-184

Kumagai, J. (2007), 'A Robotic Sentry For Korea's Demilitarized Zone', *IEEE Spectrum*, 1 March

Margulies, P. (2017), 'Making autonomous weapons accountable: command responsibility for computer-guided lethal force in armed conflicts', in J.D. Ohlin (ed), *Research Handbook on Remote Warfare*, Cheltenham UK and Northampton, MA, USA: Edward Elgar, pp. 405-442

MBDA (2009), 'MBDA Presents A More Versatile Brimstone', 10 February, available at: http://www.defense-aerospace.com/articles-view/release/3/102330/mbda-details-more-versatile,-dual_mode-brimstone.html accessed 20 July 2019

Montavon, G., W. Samek, and K.-R. Müller (2018), 'Methods for interpreting and understanding deep neural networks', *Digital Signal Processing*, 73, pp. 1-15

Roorda, M. (2015), 'NATO's Targeting Process: Ensuring Human Control Over and Lawful Use of 'Autonomous' Weapons', in A. Williams and P. Scharre (eds), *Autonomous Systems: Issues for Defence Policymakers*, The Hague, The Netherlands: NATO Communication and Information Agency, pp. 152-168

Rosert, E. (2017), 'How to Regulate Autonomous Weapons. Steps to Codify Meaningful Human Control as a Principle of International Humanitarian Law', *PRIF Spotlight*, **6**

Russian Federation (2018), 'Russia's Approaches to the Elaboration of a Working Definition and Basic Functions of Lethal Autonomous Weapons Systems in the Context of the Purposes and Objectives of the Convention', working paper submitted to the Group of Governmental Experts on lethal autonomous weapons of the CCW, Geneva, 4 April (UN Doc. CCW/GGE.1/2018/WP.6)

Scharre, P. (2017), 'Centaur Warfighting: The False Choice of Humans vs. Automation', *Temple International and Comparative Law Journal*, **30**(1), pp. 151-165

Scharre, P. (2018), *Army of None. Autonomous Weapons and the Future of War*, New York US and London UK: W.W. Norton & Company

Sharkey, A. (2019), 'Autonomous weapons systems, killer robots and human dignity', *Ethics and Information Technology*, **21**(2), pp. 75-87

Sharkey, N. (2016), 'Staying the Loop: human supervisory control of weapons', N. Bhuta et al. (eds), *Autonomous Weapons Systems: Law, Ethics, Policy*, Cambridge, UK: Cambridge University Press, pp. 23-38

Singh Gill, A. (2018), 'Provisional agenda', submitted to the Group of Governmental Experts on lethal autonomous weapons of the CCW, Geneva, 16 March (UN Doc. CCW/GGE.1/2018/1)

Sparrow, R. (2007), 'Killer Robots', *Journal of Applied Philosophy*, **24**, pp. 62-77

Szegedy, C., W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow and R. Fergus (2014), 'Intriguing properties of neural networks', *arxiv.org*, 19 February

Tamburrini, G. (2016), 'On Banning Autonomous Weapon Systems: From Deontological to Wide Consequentialist Reasons', in N. Bhuta et al. (eds), *Autonomous Weapons Systems: Law, Ethics, Policy*, Cambridge, UK: Cambridge University Press, pp. 122-141

United Nations Institute for Disarmament Research (UNIDIR) (2014), 'The Weaponization of Increasingly Autonomous Technologies: Considering how Meaningful Human Control might move the discussion forward', UNIDIR Resource No. 2

UNIDIR (2018), 'The Weaponization of Increasingly Autonomous Technologies: Artificial Intelligence', UNIDIR Resource No. 8

United Kingdom (UK) (2018), 'Human Machine Touchpoints: The United Kingdom's perspective on human control over weapon development and targeting cycles', working paper submitted to the Group of Governmental Experts on lethal autonomous weapons of the CCW, Geneva, 8 August (UN Doc. CCW/GGE.2/2018/WP.1)

UK House of Lords Select Committee (2018), 'AI in the UK: ready, willing and able?', Report of Session 2017–19, 16 April

UK Ministry of Defence (2017), 'Joint Doctrine Publication 0-30.2. Unmanned Aircraft Systems', August

United States (US) (2018a), 'Human-Machine Interaction in the Development, Deployment and Use of Emerging Technologies in the Area of Lethal Autonomous Weapons Systems', working paper submitted to the Group of Governmental Experts on lethal autonomous weapons of the CCW, Geneva, 28 August (UN Doc. CCW/GGE.2/2018/WP.4)

US (2018b), 'Humanitarian benefits of emerging technologies in the area of lethal autonomous weapon systems', working paper submitted to the Group of Governmental Experts on lethal autonomous weapons of the CCW, Geneva, 28 March (UN Doc. CCW/GGE.1/2018/WP.4)

US Air Force Chief Scientist (2015), 'Autonomous Horizons. System Autonomy in the Air Force – A Path to the Future. Vol. I Human Autonomy Teaming', AF/ST TR 15 01, June

US Department of Defense (2012), Directive 3000.09, "Autonomy in Weapons Systems", 21 November