ESIL Reflections

18 September 2019 Volume 8, Issue 5

Filling the Empty Box: A Principled Approach to Meaningful Human Control over Weapons Systems

Image by Geof Stearns (cc)

Daniele Amoroso (University of Cagliari)
and Guglielmo Tamburrini (University of Naples Federico II)

I. Introduction

In academic and diplomatic debates about ‘autonomous weapons systems’ (AWS), a watchword has rapidly gained ground across the opinion spectrum: all weapons systems, including autonomous ones, should be under meaningful human control (MHC). The UK-based NGO Article 36[1] can be credited for putting MHC at the centre of AWS debates by publishing a series of reports and policy-papers making the case for MHC over individual attacks as a requirement under international law.[2]

Unlike the calls for a pre-emptive ban on AWS, the ‘meaningful human control’ formula (and ones similar to this) was immediately met with interest by a number of States. MHC is in fact an easy-to-understand formula; it is characterised by constructive ambiguity, which may prove helpful to bridge the gap between various positions expressed at the international level; it enables one to sidestep intractable definitional problems regarding the distinction between autonomy and automation, by focusing on a more tractable normative problem concerning the types and levels of human control to be exerted on weapon systems in general.

[1] The NGO’s name is after Art. 36 of the 1977 Additional Protocol I to the Geneva Conventions, which requires States to review new weapons, means and methods of warfare in order to establish whether their employment would be contrary to international law. Needless to say, such provision applies also AWS.
[2] Article 36, *Killer Robots: UK Government Policy on Fully Autonomous Weapons*, April 2013.

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 AWS (GGE), which was established by the State Parties to the Convention on Conventional Weapons (CCW). This is reflected in the 'Possible Guiding Principles' adopted by the GGE at its August 2018 meeting. Most notably Principle 2 posits that 'Human responsibility for decisions on the use of lethal force must be retained […]'.[3] Here is, however, exactly where international consensus ends. As many commentators pointed out, it is far from settled – *even among those favouring a MHC requirement* – what its actual content should be or, to put it more sharply, what is normatively demanded to make human control over weapons systems actually 'meaningful'. It is, therefore, safe to say that in the current state of the AWS debate, this requirement is little more than an empty box.

This Reflection contributes to advancing the AWS debate by filling the MHC placeholder with more precise content. In particular, the ethical and legal reasons supporting the MHC requirement will be summarised with a view to pinpointing the functions that humans must perform to ensure MHC over weapon systems. It will then be argued that in order to secure and maintain these functions, two chief problems need to be addressed and solved, namely: (i) how to guarantee a proper quality of human involvement; (ii) how to properly establish exclusive control privileges for human operators. On this basis, some key aspects of a legal instrument enshrining the MHC requirement (such as a Protocol VI to the CCW) will be tentatively identified.

II. A Précis of the Ethical and Legal Case for Meaningful Human Control

Most legal and ethical debates about autonomy in weapons systems are based on a shared understanding of the critical functions of AWS, which are spelt out as a necessary condition of AWS, expressed – although with slightly different wording – by the US Department of Defense (DoD) and the International Committee for the Red Cross (ICRC): to count as autonomous, a weapon system must be able to select and engage targets without human intervention.[4]

A MHC requirement over weapon systems would aim at curbing the latter's autonomy over their critical target selection and engagement functions. This is the main reason why the necessary condition provides an adequate starting point for the ensuing discussion on the motives for and the content of a MHC requirement. By the same token, it is unsurprising that arguments supporting MHC largely coincide with arguments making the case for a ban on AWS. This pro-MHC/anti-AWS convergence notably emerges from the following three clusters of arguments.

[3] *Report of the 2018 session*, Geneva, 23 October 2018 (UN Doc. CCW/GGE.1/2018/3), para. 21(b). See also *Draft Report of the 2019 session*, Geneva, 21 August 2019 (UN Doc. CCW/GGE.1/2019/CRP.1/Rev.2), para. 17(e) ('Human judgement is essential in order to ensure […] compliance with international law').

[4] See, in almost identical terms, US DoD, Directive 3000.09, 'Autonomy in Weapons Systems', 21 November 2012, 13-14 and ICRC, *Views 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 2016, 1.

Firstly, AWS would be unable to comply with the basic tenets of International Humanitarian Law (IHL) (i.e. the principles of distinction, proportionality, and precaution). The development of AWS fulfilling the distinction and proportionality requirements, which match at least the performance of a competent and conscientious human soldier, presupposes a solution to many profound research problems in artificial intelligence (AI) and advanced robotics.[5] 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) harm to the civilian population, insofar as the regular behaviour of AI and robotic systems is disrupted by unpredicted dynamic changes occurring in warfare environments. In particular, adversarial testing have shown that systems developed with advanced machine learning technologies (e.g. deep learning) are prone to unexpected, counter-intuitive and potentially catastrophic mistakes, which a human operator would easily detect and avoid.[6]

Secondly, AWS are likely to raise an accountability gap. One cannot exclude that AWS will assume targeting decisions, which if taken by human agents, would trigger individual criminal responsibility. Who then will be held responsible for this conduct? The list of potentially responsible actors in the decision-making chain includes the operator and the military commander overseeing the AWS' mission as well as those involved in its development, manufacturing and procurement. Individuals on this list may raise a defence against responsibility charges and criminal prosecution in terms of their limited decision-making roles, or the complexities of AWS systems and their unpredictable behaviour in the battlefield. Cases may arise where it would be impossible to ascertain the existence of the mental element, which is required under International Criminal Law (ICL) in order to ascribe criminal responsibility. In this scenario, an individual would not be held criminally liable, notwithstanding the conduct in question objectively amounts to an international crime. This outcome is hardly reconcilable with the principle of individual criminal responsibility under ICL.

Thirdly, and finally, the principle of human dignity would dictate that decisions affecting the life, physical integrity and property of individuals involved in an armed conflict should be entirely reserved to humans and cannot be entrusted to an autonomous artificial agent. Otherwise, people subject to AWS' use of force would be put in a position where any appeal to the shared humanity of persons on the other side – and thus their inherent value as human beings – would be a priori and systematically denied.

III. Filling the Empty Box: How to Shape the Content of the MHC Requirement

What then should be the actual content of the MHC requirement? The ethical and legal reasons outlined already 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 control from truly meaningful human control. In particular, the above-mentioned arguments suggest a threefold role for human control on

[5] This is acknowledged also by roboticists who, in principle, are in favour of autonomy in weapons systems. See Ronald Arkin, *Governing Lethal Behavior in Autonomous Robots* (CRC Press, 2019), 211-212.
[6] Christian Szegedy et al. 'Intriguing properties of neural networks', *arxiv.org*, 19 February 2014.

weapons systems in order to be truly meaningful. Firstly, human control must afford a *fail-safe mechanism*, which is designed to prevent a weapons malfunction from resulting in a direct attack against protected persons and objects, or excessive collateral damages.[7] Secondly, it must serve as a *catalyst for accountability*, insofar as it stipulates the legal conditions for the attribution of responsibility when a weapon follows a course of action in breach of international law.[8] Thirdly, it must ensure that it is a *moral agent*, and not an artificial one, who takes decisions affecting the life, physical integrity and property of people (including combatants) involved in an armed conflict.[9]

The preservation by human agents of these various properties in relation to increasingly autonomous weapons 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*.

A. The Quality of Human Involvement: Training and Design

In the first place, military personnel training should foster awareness of both established and possible limits in the proper functioning of weapons systems, and related human predicaments in the capability to predict and control their behaviour. Awareness-raising efforts concerning limitations in AWS functioning should be part of more comprehensive training exercises, whereby the military personnel learn to use advanced technologies without forfeiting human judgement and critical thinking, and without succumbing to so-called automation biases.

If humans are expected to not blindly trust the machine, they should be in a position to obtain sufficient humanly understandable information about machine data processing, so as to allow operators to achieve 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). Both interpretability and explainability requirements concern the *design* of weapons systems and must be addressed by R&D and T&E teams.

To fulfil the interpretability requirement, it is necessary to map machine data and information processing into domains that humans can make sense of. Accordingly, AWS should be designed so as to provide commanders and operators with ‘access to the sources of information’ handled by the system[10] in a way that allows humans to absorb and process data ‘at the level of meaning’, rather than ‘in a purely syntactic manner’.[11] In general,

[7] Paul Scharre, ‘Centaur Warfighting: The False Choice of Humans vs. Automation’ (2016), 30 *Temple International and Comparative Law Journal* 151, 154.

[8] Thompson Chengeta, ‘Defining the Emerging Notion of “Meaningful Human Control” in Autonomous Weapon Systems’ (2017), 49 *New York Journal of International Law & Politics* 833.

[9] ICRC, *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 2018 (UN Doc. CCW/GGE.1/2018/WP.5), paras. 23-26.

[10] Richard Breton and Eloi Bossé, **‘**The Cognitive Costs and Benefits of Automation’ (2013), in NATO (ed.), *The Role of Humans in Intelligent and Automated Systems*, RTO Meeting Proceedings MP-088, 1, 10-11.

[11] Patrick Chisan Hew, ‘Preserving a combat commander’s moral agency: The Vincennes Incident as a Chinese Room’ (2016), 18 Ethics and Information Technology 227, 230.

military technological advances should empower human combatants by enhancing their situational awareness, rather than substituting artificial agents for human understanding and judgment.[12]

To fulfil the explainability requirement, AWS should come with the capability to provide, in terms that are cognitively accessible to human users, explanations of why courses of actions are being suggested or undertaken. Meeting the explainability requirement might prove particularly demanding in relation to AWS equipped with machine-learning capabilities. Learning technologies – most notably, deep neural networks – have enabled one to achieve remarkable results in algorithmic classification and decision-making. However, they involve sub-symbolic data representations and information processing that are not transparent to human users. A new and rapidly growing area of research, eXplainable AI (XAI), is addressing interpretability and explainability issues for learning AI systems. However, pending significant breakthroughs in XAI, one cannot but acknowledge the present technological difficulties in ensuring sufficient levels of system interpretability and explainability for exercising MHC on the more advanced AI-based weapons systems.

B. Exclusive Control Privileges for Human Operators

Several attempts have been made to define control privileges of human operators under the MHC requirement. While differing in significant ways from each other, these various proposals are generally affected by a common weakness. All of these proposals look for one formula, uniformly capturing optimal human-machine partnership for all kinds of weapons systems and for each of their possible uses. In general, these attempts are either overly permissive or overly restrictive.

On the side of overly permissive attempts, one may recall the Dutch (or 'wider loop') approach, whereby MHC would be exerted by human commanders at the planning stage of the targeting process.[13] This approach 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 pursues targets of opportunity. To the extent that it warrants the weapon with humanly unrestrained autonomy after deployment, the Dutch approach appears to be deeply problematic, in that it drives a wedge between the State owing a duty of care towards the civilian population (and other protected persons) and the actual possibility to comply with that duty by influencing the course of events through its agents. This wedge emerges already in the case of loitering weapons systems, which explore a given area for sustained periods of time in search of enemy targets to attack:[14] the conditions authorising the activation of a loitering AWS by human operators may rapidly change in warfare scenarios characterised by erratic or surprise-seeking behaviours.

[12] US Air Force Chief Scientist, *Autonomous Horizons. System Autonomy in the Air Force – A Path to the Future. Vol. I Human Autonomy Teaming*, AF/ST TR 15 01, June 2015, 8.

[13] The Dutch position is based on: 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*, No. 97 AIV / No. 26 CAVV (2015).

[14] A well-known sample of loitering munition is the Israeli Harpy NG system (https://www.iai.co.il/p/harpy accessed 20 July 2019).

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 each and every type of AWS and use thereof. This uniform protocol should equally apply to AWS selecting and attacking targets of opportunity in civilian populated areas and to defensive systems autonomously firing against incoming rockets and missiles.[15] These attempts may prove inadequate, in that milder forms of human control might be equally able to 'purify' of its ethically and legally troubling implications the autonomous action of certain defensive weapon systems and other weapons operating in some very limited operational environments.

To avoid predicaments of both overly restrictive and 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 is nevertheless based on the unifying grounds provided by the ethical and legal principles outlined in Section II. 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 a genuine MHC over weapon systems *in context*.

The 'if-part' of these rules should include properties concerning *what* mission the weapons system is involved in, *where* it will be deployed, and *how* it will perform its tasks. The 'what-properties', in particular, must address operational goals (defensive vs. offensive), targeting modes (deliberate vs. dynamic), and the nature of targets to be engaged (human combatants, manned military vehicles, inhabited military buildings vs. uninhabited military vehicles and buildings). The 'where-properties' must address dynamical features of the operational environment, including interactions with the adversary's autonomous artificial agents, having special regard to the presence or absence of civilians, civilian objects and friendly forces. Finally the 'how-properties' must address the information-processing and sensory-motor capabilities that the system puts to work for carrying out its mission and that may affect its overall controllability and predictability. Learned decision-making and 'swarm intelligence'[16] abilities, which may be increasingly implemented in future AWS, together with loitering capabilities of existing weapons, are significant examples of 'how-properties' that raise serious concern from a MHC perspective.[17]

The 'then-part' of bridge rules should establish what kind of human-machine shared control would be legally required for each single use of a weapons system. Following a taxonomy

[15] See, e.g., the Israeli Iron Dome (https://www.army-technology.com/projects/irondomeairdefencemi/ accessed 20 July 2019) and the German Nächstbereichschutzsystem (NBS) MANTIS (https://www.army-technology.com/projects/mantis/ accessed 20 July 2019

[16] 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).

[17] See above the text accompanying n 6 (machine-learning) and n 14 (loitering technology). In relation to swarm technology, see ICRC, 'Statement on Agenda Item 5(b)', delivered at the Group of Governmental Experts on lethal autonomous weapons of the CCW, Geneva, March 2019.

proposed by Noel Sharkey (and only slightly modified below),[18] one may schematically consider five basic levels 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 programme suggests alternative targets and a human chooses which one to attack;
L3. A programme selects targets and a human must approve before the attack;
L4. A programme selects and engages targets, but is supervised by a human who retains the power to override the programme's choices and abort the attack;
L5: A programme selects targets and initiates attack on the basis of the mission defined at the 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 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 applied. Under this proviso, lower levels of human control may only become acceptable as internationally agreed upon exceptions, which are clearly formalised as specific bridge rules. These bridge rules should establish what level is required to ensure 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 list above.

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

1. Deliberate targeting (*what-property*) by AWS may be pursued at a 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 any 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, where civilians and civilian objects are not present (*where-property*), so as to ensure that there is a human on the attacking end, who can verify whether there are persons *hors de combat* and take appropriate measures accordingly. A case in point is the Samsung SGR-A1, a South Korean military robot sentry, deployed on the south-end side of the Korean demilitarised zone.[19]
2. The (L4) human supervision and veto level might be deemed as an acceptable level of control in the case of AWS with exclusively defensive functions (*what-property*). This is the case of the Israeli Iron Dome and the German *Nächstbereichschutzsystem*

[18] Noel Sharkey, 'Staying the Loop: human supervisory control of weapons', in Nehal Bhuta et al. (eds), *Autonomous Weapons Systems: Law, Ethics, Policy* (CUP, 2016) 23, 34-37. Deviations concern, notably, levels L4 and L5.
[19] Jean Kumagai, 'A Robotic Sentry For Korea's Demilitarized Zone', IEEE Spectrum, 1 March 2007.

(NBS) MANTIS, according to their customary uses as protective shields from incoming shells and rockets.
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 higher levels of human control (L1 and L2).

IV. Conclusions

At the August 2018 meeting of the GGE on lethal AWS, the Austrian, Brazilian and Chilean delegations jointly submitted a proposal for a mandate to 'negotiate a legally binding instrument to ensure meaningful human control over critical functions in lethal autonomous weapon systems'.[20] It is doubtful that this proposal will be followed up within the institutional framework of the CCW, at least in the short term, given that some major military powers, including the US, oppose such a solution. At the same time, however, the proposal of relinquishing the quest for a one-size-fits-all solution to the MHC issue in favour of a suitably differentiated approach may help sidestep current stumbling blocks.[21] Diplomatic and political discontent about a MHC requirement which appears to be overly restrictive with respect to the limited autonomy of some weapons systems might be mitigated by recognising the possibility of negotiating exceptions to L1-L2 human control if one is able to identify weapons systems and contexts of use where milder forms of human control will suffice. It therefore seems appropriate to start thinking about the content of such – for now hypothetical – treaty. In the light of the foregoing, we suggest that a 'MHC Convention/Protocol' should address, as a minimum, the following points:

1. The essential elements of the ethical and legal concerns outlined in Section II 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 weapons systems must be stated in a provision of general purpose. The content of this requirement should then be clarified in three ensuing parts, concerning 'Training', 'Control by design' and 'Control in use' respectively.[22]
3. Provisions on 'Training' must spell out State obligations to foster awareness among AWS decision-makers and users about the limits affecting autonomous targeting by weapons systems, and to train them to preserve a critical approach and countervail risks of so-called automation bias.
4. The 'Control by design' part must include provisions prescribing the satisfaction of interpretability and explainability requirements, as set out in Section III.A. Technical specifications for these requirements can be detailed in a specific Annex.

[20] UN Doc. CCW/GGE.2/2018/WP.7 (30 August 2018).

[21] This perspective seems to underpin the additional guiding principle agreed on by the High Contracting at the last GGE meeting, whereby '[…] [i]n determining the quality and extent of human-machine interaction, a range of factors should be considered including the operational context, and the characteristics and capabilities of the weapons system as a whole.' (*Draft Report of the 2019 session* (n 3), para. 16(a)).

[22] With regard to the latter two terms, see International Panel on Regulation of Autonomous Weapons (iPRaW), *Focus on the Human-Machine Relation in LAWS,* Report No. 3, March 2018.

5. The ‘Control in use’ part will undoubtedly be most important and challenging to agree upon. Our suggestion is to establish higher levels of human control (L1-L2) as a default policy and to regulate exceptions thereto by way of bridge rules like those suggested in Section III.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 grounds of the converging ethical and legal principles described above.
6. Crucial to the actual implementation of the MHC requirement will be the introduction of transparency obligations, verification procedures, and confidence-building measures. This aspect cannot be addressed here in detail.[23] However, the analysis carried out in this Reflection provides some indication as to *what* information State Parties should share with others. For instance, States might be required to communicate to other Parties the weapons systems where they wish to adopt a control policy different from the default policy and on what basis they consider applicable one of the exceptions set down in the bridge rules.

Cite as: Daniele Amoroso and Guglielmo Tamburrini, 'Filling the Empty Box: A Principled Approach to Meaningful Human Control over Weapons Systems', ESIL Reflections 8:5 (2019).

[23] In this regard, see Sarah Knuckey, ‘Autonomous weapons systems and transparency: towards an international dialogue‘, in Bhuta (n 18), 164.