UNIDIR

The Weaponization of Increasingly Autonomous Technologies:

Considering how Meaningful Human Control might move the discussion forward

No. 2 UNIDIR RESOURCES

Acknowledgements

Support from UNIDIR's core funders provides the foundation for all of the Institute's activities.

In addition, dedicated project funding was received from the governments of the Netherlands and Switzerland.

The Institute would also like to thank Tae Takahashi and Elena Finckh for their valuable assistance with this project.

About UNIDIR

The United Nations Institute for Disarmament Research (UNIDIR)—an autonomous institute within the United Nations—conducts research on disarmament and security. UNIDIR is based in Geneva, Switzerland, the centre for bilateral and multilateral disarmament and non-proliferation negotiations, and home of the Conference on Disarmament. The Institute explores current issues pertaining to the variety of existing and future armaments, as well as global diplomacy and local tensions and conflicts. Working with researchers, diplomats, government officials, NGOs and other institutions since 1980, UNIDIR acts as a bridge between the research community and governments. UNIDIR's activities are funded by contributions from governments and donor foundations.

Note

The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of the Secretariat of the United Nations concerning the legal status of any country, territory, city or area, or of its authorities, or concerning the delimitation of its frontiers or boundaries.

The views expressed in this publication are the sole responsibility of UNIDIR. They do not necessarily reflect the views or opinions of the United Nations or UNIDIR's sponsors.

www.unidir.org

The Weaponization of Increasingly Autonomous Technologies[1]: Considering how Meaningful Human Control might move the discussion forward

Recent discussions in a variety of intergovernmental, academic and policy forums have considered, among other issues, the objective of maintaining "meaningful human control" over increasingly autonomous weapons. This UNIDIR paper examines what may be understood by "meaningful human control", its strengths and weaknesses as a framing concept for discussions on autonomy and weapon systems, as well as other conceptual and policy-oriented approaches that address concerns about the weaponization of increasingly autonomous technologies. It is the second in a series[2] of UNIDIR papers on the weaponization of increasing autonomous technologies.[3]

Where did the concept of Meaningful Human Control come from?

The phrase "human control" in relation to weapon systems is not new—it has been used by both proponents and opponents of increasingly autonomous technologies. The 2012 US Department of Defence Directive 3000.09 on Autonomy in Weapon Systems uses the phase "human control" in its definition of semi-autonomous weapon systems.[4] The International Committee of the Red Cross has stressed the need for "human control" of certain "critical functions" of weapon systems.[5] Early statements of the International

1 UNIDIR has purposefully chosen to use the word "technologies" in order to encompass the broadest relevant categorization. In this paper, this categorization includes machines (inclusive of robots and weapons) and systems of machines (such as weapon systems), as well as the knowledge practices for designing, organizing and operating them.

2 UNIDIR's first paper is entitled "Framing Discussions on the Weaponization of Increasingly Autonomous Technologies" (May 2014). For more information about UNIDIR's project "The Weaponization of Increasingly Autonomous Technologies", sponsored by the Governments of Switzerland and the Netherlands, see www.unidir.org/programmes/security-and-society/the-weaponization-of-increasingly-autonomous-technologiesimplications-for-security-and-arms-control.

3 The views expressed in this paper are the sole responsibility of UNIDIR. UNIDIR would like to acknowledge the thoughtful contributions of the participants in a May 2014 meeting on meaningful human control convened by UNIDIR: John Borrie, Maya Brehm, Neil Davison, Kristian Hammond, Peter Herby, Patrick Lin, George Lucas, Noam Lubell, Richard Moyes, Erwan Roche, Lisa Rudnick, WSP Sidhu, Rob Sparrow, Alexandre Vautravers and Kerstin Vignard. UNIDIR would also like to acknowledge the contributions of those experts and interviewees who have requested to remain unnamed.

4 See United States, Department of Defence, "Directive 3000.09. Autonomy in Weapon Systems" of 21 November 2012, page 14, at www.dtic.mil/whs/directives/corres/pdf/300009p.pdf.

5 See ICRC statement of 13 May 2014 at www.icrc.org/eng/resources/documents/statement/2014/05-13-autonomous-weapons-statement.htm; see also *Report of the ICRC Expert Meeting on Autonomous Weapon Systems: Technical, Military, Legal and Humanitarian Aspects*, at www.icrc.org/eng/assets/files/2014/expert-meeting-autonomous-weapons-icrc-report-2014-05-09.pdf.

Committee for Robot Arms Control (ICRAC) spoke of the importance of human control, and the 2010 Berlin Statement of a group of experts convened by ICRAC warned of "the loss of human control over the maintenance of security, the use of lethal force and the conduct of war…".[6] However it was the UK NGO Article 36 that coined the term "Meaningful Human Control".[7]

The phrase Meaningful Human Control (MHC) has been taken up by civil society and some States as a useful framing concept for discussions on autonomy in weapon systems. ICRAC and the Campaign to Stop Killer Robots—as well as numerous individual experts—point to the challenges or inability of establishing MHC as one reason to call for a ban on autonomous weapon systems. Article 36 has continued to develop the intellectual underpinning of the concept, notably through its Briefing Papers[8] produced for delegates to the November 2013 meeting of the Convention on Certain Conventional Weapons (CCW) and the May 2014 CCW Meeting of Experts on Lethal Autonomous Weapon Systems. At the May meeting, many States welcomed the MHC concept as a useful basis for discussion, while others noted that the concept needed further development and study.[9] It is clear, however, that for a variety of stakeholders the idea of Meaningful Human Control is intuitively appealing even if the concept is not precisely defined.

As the international community grapples with how to address the weaponization of increasingly autonomous technologies, it is worthwhile to consider if the MHC concept could help frame[10] the nascent discussions and perhaps ultimately serve as the basis of an explicit international norm.

Meaningful Human Control over _____?

Article 36's concept of MHC is explicit—calling for Meaningful Human Control over individual attacks. However, the phrase is entering currency without this modifier, leading to various interpretations, e.g.:

- Meaningful Human Control over weapon systems?
- Over the critical functions[11] of autonomous weapons?
- Over individual attacks?

Each of these interpretations—and others—could be valid, but it will be crucial that those who use the term are explicit about the object of control in order to further their own thinking and to reduce potential misunderstandings.

---

6 See "Berlin Statement" at http://icrac.net/statements/.

7 See Article 36, "Killer Robots: UK Government Policy on Fully Autonomous Weapons", a commentary on the UK Ministry of Defense's 2011 Joint Doctrine Note on "The UK Approach to Unmanned Systems", at www.article36.org/wp-content/uploads/2013/04/Policy_Paper1.pdf.

8 Article 36, "Structuring Debate on Autonomous Weapon Systems" and "Key Areas for Debate on Autonomous Weapon Systems", available at www.article36.org/wp-content/uploads/2013/11/Autonomous-weapons-memo-for-CCW.pdf and www.article36.org/wp-content/uploads/2014/05/A36-CCW-May-2014.pdf.

9 See, for example, paragraph 20 of the Report of the 2014 Meeting of Experts on Lethal Autonomous Weapon Systems (Advance Copy), submitted by the Chairperson of the Meeting of Experts.

10 For the purposes of this paper, a framing refers to "conscious strategic efforts by groups of people to fashion shared understandings of the world and of themselves that legitimate and motivate collective action", Keck and Sikkink, as cited in J. Borrie, *International Affairs*, vol. 90, no. 3, 2014, p. 636.

11 The ICRC has described 'critical functions' as the acts of acquiring, tracking, selecting and attacking targets. See ICRC, *Report of the ICRC Expert Meeting on Autonomous Weapons Systems*, op. cit.

How might Meaningful Human Control advance the discussion conceptually?

MHC may be seen as a useful approach to discussing what is problematic about increasingly autonomous technologies that can identify and attack targets without any human intervention. The MHC concept could be considered *a priori* to exclude the use of such systems. This is how it is often understood intuitively. However, whether this is in fact the case depends on how each of the words involved is understood. "Meaningful" is an inherently subjective concept as individuals give different meanings to the same sets of facts. Meaningful might refer to whether there is sufficient time for a human to intervene, exercise judgment, override or terminate an attack. "Human control" may likewise be understood in a variety of ways. To some, control will mean that a human operator is monitoring the system and making all critical decisions including, in particular, the decision to attack a given target. Others will argue that human control can be sufficiently exercised through the design of a system and by ensuring that it functions reliably and predictably without having a human "in the loop" for each targeting and attack decision.[12]

The concept of MHC as described by Article 36 is distinct from the traditional characterization of a human "in or on the loop" as it offers more precision (control versus the somewhat ambiguous conceptual "loop" or the more passive "judgment"), it explicitly emphasizes the quality of control ("meaningful") and it implicitly accords responsibility to human agents for decisions concerning each individual attack.

Using the MHC concept as a frame to discuss the weaponization of increasing autonomous technologies has a range of positive aspects:

- It provides a common language for discussion that is **accessible to a broad range of governments and publics** regardless of their degree of technical knowledge.
- It focuses on a potentially shared objective of maintaining some form of control over all weapon systems and is **more comprehensive** than approaches that attempt to predict or regulate the evolution of the rapidly moving fields of technology, robotics and artificial intelligence.
- It is **consistent with international humanitarian law** regulating the use of weapons in armed conflict, which implicitly entails a certain level of human judgment and explicitly assigns responsibility for decisions made.
- It is a **concept broad enough to integrate consideration of ethics, human-machine interaction and the "dictates of the public conscience"** which are often side-lined in approaches that narrowly consider technology, law or functions.

If States wish to move from using MHC simply to structure policy discussion to using it as a basis for an international norm, further work will be needed to develop a shared understanding of how such control is operationalized. Questions such as these will need to be addressed:

- Consideration of the different interpretations of the words "meaningful", "human" and "control" and how these are similar to or distinct from other concepts, such as

12 Variations in this approach can be seen in US Department of Defence Directive 3000.09 (op. cit.) which requires that 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" and the United Kingdom policy that that the "... operation of weapon systems will always be under human control".

"human control", "appropriate levels of human judgment" and the "man in, on or out the loop".

- "What" must be the subject of MHC? For example, is it the weapon system itself, each individual attack or something else?
- If MHC becomes widely accepted as a general principle, what sorts of parameters could be identified that would provide assurance that it is indeed being exercised?
- Given that "control", in military terms, includes a variety of processes (such as intelligence collection, context analysis, target identification, a determination that the attack will be discriminate, a proportionality calculation and the decision to attack) is MHC equally necessary at each stage?
- Would a norm based on MHC raise questions about certain existing technologies?[13]
- Given that the characteristic of autonomy might be highly attractive in militarily terms when communication with a weapon or weapon system is not feasible or is lost or jammed, how could MHC be exercised in such contexts?
- How could MHC be maintained in conflicts between opponents with roughly equal levels of technological sophistication—where the speed of machine decision-making outpaces humans' ability to follow or intervene?

Deeper discussion of such issues is useful as it helps to better articulate what some find unsettling about the weaponization of increasingly autonomous technologies. Such reflections will help to identify and focus on the precise nature of the challenges increasingly autonomous technologies pose, including those related to human-machine interaction, decision-making, responsibility and accountability.

It is perhaps not necessary that the MHC concept be precisely defined. Many widely shared concepts ultimately contained in international humanitarian law instruments are not defined in themselves—including, for example, "unnecessary suffering" and "indiscriminate effects". The fact that at least two countries[14] have explicitly stated policy acknowledging the necessity of either human control or human judgment in relation to autonomous weapon systems indicates that some the notions at the heart of the concept of Meaningful Human Control are already accepted principles by some States.

In summary, at this early stage of discussions on autonomy, the concept of Meaningful Human Control **provides an approach** to discussing the weaponization of increasingly autonomous technologies, **not a solution** to the technical, legal, moral and regulatory questions they pose. MHC turns our attention away from speculation about technological development and future capabilities and toward articulating the expectation that the development and use of emerging technologies will conform to established norms of responsibility, accountability, legality, and other principles of international humanitarian and human rights law.

13 Paul Scharre notes, for example, that large numbers of existing, uncontroversial weapons systems fail to meet one set of minimum standards for meaningful control proposed at the May 2014 CCW meeting. See P. Scharre, "Autonomy, 'Killer Robots', and Human Control in the Use of Force" (part II), published on *Just Security*, 9 July 2014, at http://justsecurity.org/12712/autonomy-killer-robots-human-control-force-part-ii/

14 The United Kingdom and the United States.

What sorts of parameters shape human control over weapons systems and attacks?

In considering MHC as a frame for discussions on autonomy, it is useful to consider parameters that facilitate control over weapons generally as well as those where increasing autonomy raises specific issues or concerns.

1) **Function of the weapon**—Control is first and foremost based on knowledge of the weapon system. A thorough understanding by the operator and the commander of the selected weapon's functions and effects, coupled with contextual information (such as awareness of the situation on the ground, what other objects are in the target area, etc), contribute to the assessment of whether a weapon is appropriate for a particular attack.

   Increasing autonomy in the critical functions of weapons means that contextual assessment about targeting and attack would be delegated to the weapons themselves. This raises issues about accountability and responsibility if the effects of the attack deviate from those intended by a human commander. A greater understanding of the degree of human oversight or control of existing systems that have highly automatic characteristics, such as Phalanx and Iron Dome, would be a useful area of inquiry not because these systems necessarily pose problems but for the opposite reason—in order to consider why haven't the use of these systems raised widespread concern before.

2) **Spatial limitations**—Restricting the use of certain weapon systems to operations in specific environments seems like a reasonable mechanism for ensuring control over increasingly autonomous technologies—particularly in relation to concerns of ensuring predictability (see below) in diverse environments. Yet, would such a geographic or spatial restriction be realistic or effective? Throughout history, weapon systems designed for one type of use have been employed for other uses and in new contexts based on need and innovation.

3) **Time limitations**—Limiting the time in which a weapon system can operate is an additional way of exercising control. Precedents include time limitations on active life of unanchored sea mines contained in the 1907 Hague Convention (VIII) and those on active life of remotely delivered anti-personnel mines in Amended Protocol 2 of the Convention on Certain Conventional Weapons.

   A limitation on the time frame in which a system could function at higher levels of autonomy might be considered. Beyond the expected time frame for mission accomplishment the system would be required to return to a mode directly controlled by humans or to be redeployed. Such a parameter would also be appropriate for extended missions (such as those foreseen in the marine environment) where communications are limited, difficult or impossible. Yet would such limitations be adequate? Would systems simply be automatically reactivated at the end of each required time frame? How would one verify that a system outside of communication range had actually ended its autonomous functioning?

4) **Predictability**—For a weapon system to be under control it needs to behave in predictable ways in the environment in which it is deployed and the effect that is intended. This is in large part a product of the function of the weapon, and the space and time in which it is used.

A system that will be predictable in one environment (for example in the sea and air environments) may not behave in the same manner in another (such as a city with a wide range of objects and stimuli). In addition, a system that behaves predictably for a weapon designer or programmer may in the real world environment produce results that others, including a commander, may not expect or be able to foresee. It should also be noted that predictability of actions of such a system is not the same as predictability of outcomes/impacts. A given action may be predictable, but in interaction with a particular environment the results of that action may be surprising.

5) **Distinction between systems intended for anti-material targets and those for use against combatants or with foreseeable consequences for civilians**—It is possible that the degree of autonomy permitted for systems that target anti-materiel systems could be different from those designed for targeting persons. Yet, as with geographic restrictions mentioned above, is it realistic to expect that weapon systems will only be used for the purpose for which they are designed? And what of secondary effects? In urban areas, for example, "anti-materiel" weapons can have severe impacts on civilians. In light of the speculative nature of discussions on future autonomous weapons, it is worthwhile to consider what observations can be made about past practice: how have weapons technologies been used in the past for purposes or in situations other than that for which they were designed and what have been the consequences of their use.

6) **Stationary versus mobile roles**—It could be argued that it is more acceptable for stationary systems (e.g. those defending a given location or installation against specific types of threats) to have higher degrees of autonomy,[15] whereas those that are mobile in order to search out targets require a higher level of human control.[16] Yet what standards would be needed to ensure that even stationary systems would achieve the degree of distinction and ability to judge proportionality that is required by international humanitarian law?

7) **Standards of compliance**—Soldiers are trained and expected to fully comply with international humanitarian law in every instance. When individuals fail to comply, there are degrees of consequences for the individual, and perhaps further up in the command hierarchy as well. Given that no system may be considered 100% predictable,[17] would a lower standard of compliance be agreed upon? Would it be acceptable or legal to deploy an autonomous weapon system with the knowledge that its actions would be compliant less than 100% of the time? When humans behave in ways that fall short of compliance, there is the possibility they will face consequences. Is there an equivalent for an object?

8) **Verification**—How can measures of transparency, reporting and accountability be developed to ensure that whatever norm may be adopted at the national or

15 In fact, this is already the case. Iron Dome and the Aegis system have been deployed for some time as defensive weapons with little resistance to or concern for the degree of their autonomy.

16 Weapons such as the Harpy (a wide-area loitering munition) and encapsulated torpedo mines (a type of sea mine) are two currently deployed systems that are mobile and operate without human approval of specific target selection. P. Scharre notes that it is worth considering why so few countries have chosen to deploy these types of weapons despite the technology being fairly accessible. See P. Scharre, "Autonomy, 'Killer Robots', and Human Control in the Use of Force" (part I), published on *Just Security*, 9 July 2014, at http://justsecurity.org/12708/autonomy-killer-robots-human-control-force-part/.

17 If for no other reason that no technological system operates without fault or error 100% of the time.

international level is fully implemented? A good starting point for reflection is what sorts of transparency, reporting and accountability frameworks are in place for existing weapon technologies and whether States are likely to apply these to increasingly autonomous technologies. An additional test of this aspect is whether more countries are willing to share relevant national policies and practices, such as the United States DOD Directive. A culture of transparency offers both opportunities for exchange and learning, as well as serves as the basis to identify common underlying principles or approaches of States.

What other approaches have been suggested to help frame discussions of increasing autonomy?

Putting aside the initial traction that the concept of Meaningful Human Control has enjoyed, it is worth considering if other frameworks have the potential to provide an overarching approach that could unify discussions and form a basis for normative development. While less clearly articulated as a framing *per se*, other possibilities were put forth at the May 2014 CCW informal Meeting of Experts as useful for discussions on the weaponization of increasingly autonomous technologies. These include discussions based on technical definitions, predictability, and intent. As the disadvantages of a technology-centric frame have already been noted in UNIDIR's first observation paper,[18] here we address predictability and intent.

Predictability of compliance with international humanitarian law

An approach based solely on predictability of compliance with IHL (notably rules on distinction, proportionality and precautions in attack) takes as a starting point that international humanitarian law has already integrated ethical as well as military considerations into universally accepted norms on the use of weapons in armed conflict. Rather than re-opening such issues in relation to increasingly autonomous weapon systems, it is suggested that it would be more effective to simply determine the capacity of any future system to comply with existing core rules. Compliance with known and accepted norms would therefore be the measure of both acceptability and legality.

If this approach were to be pursued, one would have to determine what level of compliance with existing rules is deemed acceptable. As discussed above in relation to MHC, a judgment would need to be made as to what level of compliance would be necessary or acceptable; whether technology will be evaluated using the same standard as human beings; or whether standards would need to be higher. Questions also arise about responsibility for the deployment of systems with a known error or fail rate (i.e. level of non-compliance).

Positive attributes of an approach based on predictability include:

- Compliance with known and already accepted norms becomes the measure of both acceptability and legality.
- The level of predictability can, at least in theory, be tested and quantified prior to deployment.

18 See UNIDIR, 2014, "Framing Discussions on the Weaponization of Increasingly Autonomous Technologies", op. cit., pp. 7–8.

- The predictability of compliance by increasingly autonomous systems can be compared to that of human combatants/operators.

However, this approach is not without challenges:

- Levels of predictability in a range of environments will be difficult to reliably forecast as testing conditions might not correspond to those present at time, place and context of use.
- Predictability levels will likely be based upon "state of the art" systems. As similar systems are produced by an increasing number of actors, or modified by commanders in battle or other actors, the levels of compliance with international humanitarian law are likely to be lower.
- Is it technically possible—and is it desirable—to replace human judgment with a set of data-based "proxies" (i.e. data that is readable by computerized systems) for assessments of distinction, proportionality and precaution. For example, how nuanced would a targeting identification algorithm need to be? Would a targeting algorithm that classifies all males between the ages of 15 and 45 as combatants be acceptable? One might consider that this would be technologically codifying the already controversial practice of signature strikes.
- Predictability focuses on legal compliance with the existing IHL regime, which represents a negotiated outcome among States. This approach assumes that all ethical concerns have already been codified within IHL. As a result, this frame could make it difficult to address the possibility that new ethical considerations might arise from the weaponization of increasingly autonomous technologies.
- Such an approach excludes scope for ethical and social approaches in the debate on autonomous systems. It side-lines the concern about control and responsibility that many stakeholders see as central to the debate. It also neglects further consideration of human rights and the "dictates of the public conscience" that international humanitarian law recognizes as relevant to the determination of the legality of weapons.

Human intent

This approach attempts to address what is seen as the fundamental issue presented by the weaponization of highly autonomous technologies: that a machine could take life and death decisions, without any human intent involved. A norm or conceptual approach that requires that human intent must always be part of decisions to attack a target with potentially lethal consequences could be seen to articulate a generalized feeling that the automated taking of human life by machines is somehow wrong.

Some might argue that if a machine acts in a way that is consistent with the intent of the human responsible, does it matter whether the weapon is actually controlled/operated/supervised by a human? Others would say that "automated" killing that is not based on a human's intent challenges core principles of both human rights law and IHL that protect human dignity and prohibit the arbitrary taking of life.

Positive attributes of an approach based on intent include:

- That it addresses a fundamental issue of human rights and dignity that human life should not be ended or damaged with no person ever intending this result.

- That it would capture the belief that the tools of violence are used by humans to achieve human-determined objectives.

Challenges of an approach based solely on intent would include:

- Clarity about how human intent is expressed and what evidence would be required to prove intent. Also, whose intent? The designer(s) of a weapon, or its programmer(s)? Manufacturer? Commander? Operator? Could, for example design and programming alone could be taken as an expression of human intent?
- Focusing on intent might be too narrow. The approach may not adequately capture the range of social/ethical concerns of which intent is only one aspect.

It is interesting to note that both predictability and human intent are captured within the broad frame of MHC and thus these frames are not mutually exclusive approaches. Proponents of all three frames already have common ground for discussion.

Final observation

An attempt to develop national or international norms to address the challenges associated by the weaponization of increasingly autonomous technologies will need organizing principles that can be understood and broadly shared by a variety of States, civil society organizations, the public and the media.

To date the conceptual approach that has attracted the most interest has been that of Meaningful Human Control. Although this concept already serves as a welcome entry point for discussing the relevant issues, its strength as a frame for future discussions depends on how willing States, experts, organizations and other stakeholders are to engage in substantive exploration of fundamental and challenging issues, including those noted in this paper.

Any attempt to address the weaponization of increasingly autonomous technologies, at the global or national level, will need to be primarily a social and political process and not merely—or even primarily—a technical or legal exercise. The fact that the international community has indicated a willingness to engage on the topic of autonomy in weapon systems, and as different approaches to the topic already indicate some common ground, there is a solid foundation to build upon.

UNIDIR

The Weaponization of Increasingly Autonomous Technologies:

Considering how Meaningful Human Control might move the discussion forward

Recent discussions in a variety of intergovernmental, academic and policy forums have considered, among other issues, the objective of maintaining "meaningful human control" over increasingly autonomous weapons. This UNIDIR paper examines what may be understood by "meaningful human control", its strengths and weaknesses as a framing concept for discussions on autonomy and weapon systems, as well as other conceptual and policy-oriented approaches that address concerns about the weaponization of increasingly autonomous technologies. It is the second in a series of UNIDIR papers on the weaponization of increasing autonomous technologies.