KEY ELEMENTS OF A TREATY ON FULLY AUTONOMOUS WEAPONS

FREQUENTLY ASKED QUESTIONS

The Campaign to Stop Killer Robots and others have built a strong case against fully autonomous weapons, also referred to as lethal autonomous weapons systems or "killer robots." These weapons systems, which would select and engage targets without meaningful human control, raise a host of moral, legal, accountability, security, and technological concerns. Having examined the topic since 2013, states should now move from discussing the challenges to crafting a solution. They can do so by launching negotiations of a treaty to prohibit fully autonomous weapons and retain meaningful human control over the use of force. The Campaign regards such a new international treaty as a humanitarian priority, a legal necessity, and an ethical obligation.

To inform this process, the Campaign to Stop Killer Robots has identified key elements of a legally binding instrument on fully autonomous weapons. Its proposal, presented in a paper originally distributed in November 2019, distills meaningful human control into three categories of components and lays out a combination of treaty prohibitions and positive obligations. This complementary "frequently asked questions" paper expands on the Campaign's position and responds to some of the challenging questions raised by the proposal.

While the specific language and content of a final treaty will depend on the results of negotiations, the Campaign's proposal demonstrates the feasibility of developing a new instrument and provides a starting point for further discussion.

CAMPAIGN TO STOP KILLER ROBOTS

This Campaign to Stop Killer Robots briefing paper was prepared by Bonnie Docherty of Human Rights Watch and the Harvard Law School International Human Rights Clinic, with the support of her law students in the Clinic.

1. WHY IS A NEW TREATY NECESSARY?

A new treaty is necessary to clarify and strengthen existing international law. Many states argue that international humanitarian law is sufficient, but its rules were written for humans not machines. Drafters could not envision weapons systems with full autonomy and did not intend the law to govern them. A new treaty would establish clear international rules designed to address the specific problem of autonomy in weapons systems. In so doing, it would promote consistency of interpretation and implementation and facilitate enforcement.

The treaty proposed by the Campaign to Stop Killer Robots extends the traditional scope of existing international humanitarian law. It addresses not only use but also production and development. In addition, it covers the use of technology in law enforcement operations as well as situations of armed conflict. While international human rights law applies to law enforcement operations, that body of law would also be strengthened by a treaty dedicated to fully autonomous weapons.

A new legally binding instrument would go beyond the "normative and operational framework" proposed by the states parties to the Convention on Conventional Weapons (CCW). A treaty would set international standards for dealing with the complexities of fully autonomous weapons. It would bind states parties and influence states not party and non-state actors. Working toward a "normative and operational framework," an intentionally ambiguous goal, distracts states from the priority of developing an effective response to the challenges posed by fully autonomous weapons.

2. WHAT IS THE SCOPE OF THE WEAPONS SYSTEMS COVERED BY THE PROPOSED TREATY?

The proposed treaty has a broad scope of application encompassing all weapons systems that select and engage targets on the basis of sensor inputs. In other words, it covers systems that rely on sensor processing, not humans, to identify and apply force to objects that match a preprogrammed profile. By necessitating a thorough assessment of all systems that operate in this way, the treaty seeks to ensure that any subset of systems posing legal and ethical concerns does not escape regulation.

Though the proposed treaty is broad in scope, its restrictions are narrower. It imposes limitations on only two categories of weapons systems: (1) those that inherently—i.e., by their design rather than by their manner of use—raise fundamental moral or legal problems, and (2) weapons that may not be inherently unacceptable, but have the potential to be used without meaningful human control. By prohibiting systems in the first category and regulating those in the second, the treaty would help preserve meaningful human control over the use of force.

3. DOES THE SCOPE OF THE PROPOSED TREATY EXTEND TO EXISTING WEAPONS SYSTEMS?

The proposed treaty applies to existing weapons systems if they rely on sensor processing to select and engage targets. Its future-looking restrictions, however, focus on systems that would operate without meaningful human control and other systems that would raise fundamental legal and moral concerns. For example, when sensors in Israel's Iron Dome or the US Phalanx Close-In Weapon System detect incoming missiles or rockets, they respond quickly to shoot down the threat. While these automatic missile-defense systems rely on sensor processing, they operate within tight parameters in relatively controlled environments and target munitions rather than people. In addition, there is an opportunity for a human override. These systems thus seem to function within the bounds of meaningful human control and the final treaty would be unlikely to restrict their use.

The purpose of the broad scope is not to restrict the use of existing systems but to ensure that emerging technologies do not cross a threshold of acceptability, especially given the rapid pace of development. It seeks to limit systems that have more sweeping capacities to select and engage a range of targets, including humans, in unpredictable and dynamic environments. During treaty negotiations, states can determine the parameters of meaningful human control and decide how existing systems fit into its framework. An assessment of existing systems may help inform how meaningful human control is ultimately understood and operationalized.

4. WHAT KEY ELEMENTS SHOULD THE TREATY INCLUDE?

The Campaign to Stop Killer Robots proposes that the treaty include three key elements: (1) a general obligation to maintain meaningful human control over the use of force, (2) prohibitions on specific weapons systems that select and engage targets independently and by their nature pose fundamental moral or legal problems, and (3) specific positive obligations to ensure that meaningful human control is maintained in the use of all other systems that select and engage targets.

The general obligation articulates the central principle of the treaty and provides guidance for interpreting the rest of the instrument. It would be operationalized through the prohibitions and positive obligations, which would work together not only to ban the development, production, and use of the most concerning weapons systems but also to impose regulations on the use of all weapons systems that may operate without meaningful human control. Although the exact language of the new treaty will be finalized by states in the course of negotiations, this framework constitutes a comprehensive approach to addressing the dangers posed by fully autonomous weapons.

5. WHY IS THE CONCEPT OF MEANINGFUL HUMAN CONTROL AT THE HEART OF THE TREATY?

Meaningful human control is fundamental to all three elements of the proposed treaty because most of the concerns arising from the use of fully autonomous weapons are attributable to the lack of such human control.[1] For instance, the use of fully autonomous weapons would undermine human dignity by delegating life-and-death determinations to inanimate machines; machines cannot comprehend the value of human life and would reduce people to data points when executing their attacks. Such weapons systems would also be unable to replicate the human judgment necessary, for example, to weigh the proportionality of an attack as required under international law. Even if the systems could replicate human judgment, the law is designed to be implemented by humans. Finally, it would be legally difficult and arguably unjust to hold a human liable for the actions of a system operating beyond the human's control. All of these concerns demonstrate the need to maintain meaningful human control over the use of force.

6. WHEN IS CONTROL MEANINGFUL?

Human control is a spectrum, ranging from no control to absolute control. The qualifier "meaningful" ensures that human control over selecting and engaging targets is substantive. To satisfy that standard, it would not, for example, be sufficient for a human operator merely to flip a switch to turn on a weapon system. Instead, the contours of meaningful human control can be determined by a combination of three components, which have been distilled from international discussions and expert publications: (1) decision-making components, (2) technological components, and (3) operational components.

The decision-making components of meaningful human control give humans the information and ability to make decisions about whether the use of force complies with legal rules and ethical principles. Human operators should have an understanding of the operational environment, such as who and what is in a battlespace; an understanding of the weapons technology, such as what it could select and engage; and sufficient time for deliberation to allow the operators to make decisions that satisfy critical legal requirements, like that of distinction and proportionality. The decision-making components provide human operators context that is essential when making decisions on the use of force in complex and dynamic environments.

The technological components are embedded features of a weapon system that can enhance meaningful human control. Such components include features to ensure a system's predictability and reliability, thus enabling a human to use the system with confidence that it will act as directed and will perform consistently. Other technological components include the ability of the system to relay relevant information to the human operator and the ability for a human to intervene after the activation of the system, which allow operators to react to changes in the environment, re-evaluate the decision to apply force after a system's activation, and redirect or abort an attack if civilians have entered an area or combatants have surrendered.

Operational components impose constraints on autonomy that increase human control. They include limits on when and where a weapon system can operate and what it can target. By restricting the system's ability to act independently, these constraints reduce the likelihood that information considered at the time of a weapon system's activation would become outdated and help ensure that the weapon system would operate as intended in a dynamic environment.

7. WHAT ARE THE ADVANTAGES OF USING THE SPECIFIC TERM "MEANINGFUL HUMAN CONTROL"?

Since 2014, almost all states parties to the CCW have agreed that humans have an essential role to play in the use of force. In their discussions of lethal autonomous weapons systems, however, they have used different terms to describe this concept, including "meaningful human control," "appropriate levels of human judgment," and "human intervention."[2] The specific language of "meaningful human control," which is employed by a large number of states, international organizations, nongovernmental organizations, and other experts, offers several advantages. "Control" is a strong word, meaning as a noun, the "power or authority to guide or manage," and as a verb, "to exercise restraining or directing influence over; to have power over."[3] Control is also a concept familiar in international law. The term has been used as a prerequisite for accountability,[4] and although they do not use the actual term, treaties banning landmines, chemical weapons, and biological weapons prohibit weapons that are beyond human control after their emplacement or release.[5]

Other terms, like judgment and intervention, imply a weaker role for humans than control, and they would be insufficient to address the problems posed by fully autonomous weapons. While human judgment facilitates compliance with the proportionality test and other rules of international law, it is a narrower concept than human control. Defined as "belief or decision"[6] or as the exercise of "discernment,"[7] "judgment" focuses on thought processes rather than action. Humans who exercise control, by contrast, can both apply their legal and moral reasoning and act to ensure a machine follows it. "Intervention," which can be defined as "the act of interfering with the outcome or course especially of a condition or process,"[8] implies that humans can interfere in the direction of events, but not necessarily dictate them. As a result, limited human oversight over the use of force might be sufficient to meet this standard. Under international law, intervention is understood to require a lower level of involvement than control.[9]

Using the qualifier "meaningful" ensures that the degree of control is substantive. Although various other adjectives could be used to qualify human control, e.g., appropriate, effective, sufficient, or necessary, the word "meaningful" has distinct advantages. According to Article 36, meaningful is "general rather than context specific (e.g. appropriate), derives from an overarching principle rather being outcome driven (e.g. effective, sufficient), and it implies human meaning rather than something administrative, technical or bureaucratic."[10]

8. IS THE CONCEPT OF MEANINGFUL HUMAN CONTROL TOO ABSTRACT?

Concerns that the term "meaningful human control" is too abstract a legal standard on which to base a treaty are misguided. The law often relies on similarly subjective standards. For example, under the international criminal law rule of command responsibility, a commander can be held criminally liable for the actions of subordinates over whom the commander has "effective command and control." Similarly, under both *jus ad bellum* (law on the use of force) and *jus in bello* (law on armed conflict), legal accountability often requires "effective control" or "overall control."[11] The specific term "meaningful" has been used in setting a standard for adequate engagement or consultation with affected groups.[12] As is the case with most areas of law, clarity in the meaning of terms develops over time, including through judicial decisions, authoritative commentaries, other legal analyses, and evolving government positions.

Furthermore, an understanding of meaningful human control has already begun to take shape. States, international organizations, civil society, and other experts have all discussed the contents of the meaningful human control standard. States can draw upon and refine the components outlined above in the process of determining the exact contours of the meaningful human control standard during negotiations of the treaty.

9. WHY SHOULD THE TREATY INCLUDE A GENERAL OBLIGATION TO MAINTAIN MEANINGFUL HUMAN CONTROL OVER THE USE OF FORCE?

The general obligation sets the stage for the rest of the treaty. It establishes a principle to guide interpretation of the other provisions, and its generality will close unexpected loopholes in the treaty's prohibitions and positive obligations. These factors are particularly important given that novel issues could arise as technology evolves.

The general obligation focuses on the regulation of conduct (i.e., use of force) rather than a specific system in order to capture future, potentially unforeseeable technologies. The language "use of force" also has the benefit of making the obligation applicable to situations of armed conflict and law enforcement operations. Although international humanitarian law and international human rights law govern use of force in somewhat different ways, the new treaty can take such differences into account.

10. WHAT DOES THE PROPOSED TREATY PROHIBIT?

The proposed treaty prohibits the development, production, and use of weapons systems that select and engage targets and are inherently unacceptable for ethical or legal reasons. In other words, it prohibits systems that pose fundamental problems due to their design rather than manner of use. Clear prohibitions make monitoring and enforcement easier, and they create a strong stigma against the banned weapons systems.

Two main categories of systems fall under the prohibition. First, the treaty bans weapons systems that by their nature select and engage targets without meaningful human control. For example, the prohibition should cover systems that become too complex for human users to understand, like those that apply force based on machine learning, and thus produce unpredictable or inexplicable effects. The proposed treaty also prohibits weapons systems that select and engage humans as targets, regardless of whether they operate under meaningful human control. Such systems would rely on target profiles, i.e., certain types of data, such as weight, heat, or sound, to represent people. In killing or injuring people based on such data, these systems would violate human dignity and dehumanize violence. Systems that deliberately or unintentionally target people based on discriminatory indicators, such as age, gender, or other social identities, are particularly problematic.

11. WOULD THE PROHIBITIONS ON THE DEVELOPMENT AND PRODUCTION OF SUCH WEAPONS SYSTEMS STUNT RESEARCH AND INNOVATION IN AUTONOMOUS TECHNOLOGY?

The prohibitions on development and production are designed to further stigmatize and to prevent the existence of fundamentally flawed weapons systems that can in turn proliferate. These prohibitions would not hinder development and production of civilian or non-weaponized military autonomous technology. Research and development activities would be banned if they were directed at technology that could be used exclusively for fully autonomous weapons or that was explicitly intended for use in such weapons. Furthermore, the treaty would prohibit the production of weapons systems from dual use technology if the resulting systems would lack meaningful human control.

12. WHAT ARE THE PROPOSED TREATY'S POSITIVE OBLIGATIONS?

The proposed treaty also includes positive obligations, i.e., requirements for states to take affirmative action, to ensure that meaningful human control is maintained in the use of systems that select and engage targets. These obligations cover systems that are not inherently unacceptable by design but that have the potential to select and engage targets without meaningful human control. Like the prohibitions, the positive obligations serve as a means to implement the general obligation by requiring that the weapons systems covered by the treaty are used only with meaningful human control.

The content of the positive obligations should draw on the components of meaningful human control discussed above. For example, the treaty could require that operators understand how a weapon system functions before activating it. It could set minimum standards for predictability and reliability. In addition, or alternatively, the treaty could limit permissible systems to those operating within certain temporal or geographic parameters. In so doing, the positive obligations would help preserve meaningful human control over the use of force and establish requirements that in effect render the use of systems operating as fully autonomous weapons unlawful.

13. WOULD THE INCLUSION OF THESE POSITIVE OBLIGATIONS WEAKEN THE PROHIBITION?

The positive obligations would complement rather than compete with the proposed treaty's prohibitions. Rather than creating exceptions to the category of prohibited weapons systems, they fill a gap by regulating weapons systems that may not be inherently problematic but may still be used in ways that raise significant moral and legal problems. The positive obligations also ensure that the treaty is not limited to technology that drafters can envision today. It sets parameters of acceptability for both current and future technologies. At the same time, the positive obligations promote technological development by allowing for new technology as long as it does not cross the redline of being used without meaningful human control.

14. WHAT OTHER ELEMENTS SHOULD THE TREATY INCLUDE?

This treaty, like all legally binding instruments, should complement the core obligations discussed above with other elements. The treaty should include a preamble that articulates the purpose of the instrument, highlights the risks of fully autonomous weapons that motivated its creation, and places the issue in the context of relevant international law. Its operative part should include additional provisions that advance implementation and compliance. The treaty should include reporting requirements to promote transparency and facilitate independent monitoring. Cooperative compliance mechanisms and rigorous verification measures would help prevent violations of the treaty. Regular meetings of states parties are needed to review the status and operation of the treaty, identify implementation gaps, and set goals for the future. An obligation to adopt national implementation measures, including domestic legislation that imposes penal sanctions for violations, would further promote implementation and enforcement. There must also be a reasonable threshold for entry into force that allows the treaty to take effect in a timely manner.

15. IN WHAT FORUM COULD THIS TREATY BE NEGOTIATED?

The proposed legally binding instrument could be negotiated in a number of forums, including an independent process launched and led by like-minded states. The issue was first debated at the Human Rights Council in 2013, and CCW states parties have held informal and formal discussions on lethal autonomous weapon systems since 2014. CCW states parties agreed to a set of guiding principles in 2018 and set a plan in 2019 to "consider the development of aspects of the normative and operational framework" for these weapons systems ahead of the 2021 Review Conference. Progress towards a credible CCW outcome, particularly a mandate to negotiate a new legally binding protocol, however, has been blocked by a small number of military powers acting under the CCW's tradition of consensus decision-making. Therefore, it is doubtful that states will produce a new protocol under the auspices of the CCW, let alone one that sets a strong international standard.

States should identify the most efficient and effective path to a strong treaty, which will likely require leaving the CCW. They could turn to the UN General Assembly, where the 2013 Arms Trade Treaty and the 2017 Treaty on the Prohibition of Nuclear Weapons were negotiated and adopted. Alternatively, they could pursue an independent process, like the Ottawa Process that produced the 1997 Mine Ban Treaty and Oslo Process that led to the 2008 Convention on Cluster Munitions. A negotiating process that is not bound by consensus would be able to move faster and aim higher. The process should also include all states as well as the Campaign to Stop Killer Robots, International Committee of the Red Cross, and other international organizations.

ENDNOTES

[1] For a more detailed description of the concerns arising from the use of weapons systems without meaningful human control, see Human Rights Watch and Harvard Law School's International Human Rights Clinic (IHRC), "Killer Robots and the Concept of Meaningful Human Control: Memorandum to Convention on Conventional Weapons (CCW) Delegates," April 2016, https://www.hrw.org/sites/default/files/supporting_resources/robots_meaningful_human_control_final.pdf (accessed February 17, 2020).

[2] States that used this term in 2019 include Argentina, Austria, Germany, Ireland, Japan, Mexico, Norway, Sweden, and Switzerland. Others used the term in previous years. See, for example, Human Rights Watch and IHRC , "Killer Robots and the Concept of Meaningful Human Control," pp. 7-16.

[3] "Control," Merriam-Webster, https://www.merriam-webster.com/dictionary/control (accessed February 17, 2020).

[4] See, for example, Rome Statute of the International Criminal Court (Rome Statute), A/CONF.183/9, July 17, 1998, entered into force July 1, 2002, art. 28; see also Prosecutor v. Ignace Bagilishema, Judgment (Trial Chamber), June 7, 2001, para. 45 ("[T]he essential element is not whether a superior had authority over a certain geographical area, but whether he or she had effective control over the individuals who committed the crimes."). Additionally, in *jus ad bellum* and *jus in bello*, legal accountability often requires "effective control" or "overall control." See Military and Paramilitary Activities in and against Nicaragua (Nicaragua v. United States of America), International Court of Justice, Judgment, June 24, 1986.

[5] Human Rights Watch and IHRC, "Killer Robots and the Concept of Meaningful Human Control," pp. 10-12.

[6] "Judgement," Oxford Reference, http://www.oxfordreference.com/abstract/10.1093/acref/9780199264797.001.0001/acref-9780199264797-e-1298?rskey=VCoo4X&result=2 (accessed February 17, 2020).

[7] "Judgment," Merriam-Webster, https://www.merriam-webster.com/dictionary/judgment (accessed February 17, 2020).

[8] "Intervention," Merriam Webster, https://www.merriam-webster.com/dictionary/intervention (accessed February 17, 2020).

[9] For example, the International Court of Justice's judgment in U.S. v. Nicaragua found that sending financial aid to *contra* guerrillas qualified as impermissible intervention in Nicaragua's internal affairs. This same aid did not rise to the level of "effective control" over the *contras*' actions, however, so the United States could not be held responsible for alleged violations of international humanitarian and human rights laws. Military and Paramilitary Activities in and against Nicaragua (Nicaragua v. United States of America), International Court of Justice, Judgment, June 24, 1986, ¶ ¶ 110, 115, 228.

[10] Article 36, "Key Elements of Meaningful Human Control," April 2016, http://www.article36.org/weapons/autonomous-weapons/mhc-2016-papers/ (accessed February 17, 2020), p. 2.

[11] See supra note 4.

[12] See, for example, Department of Justice, Government of Canada, "Principles Respecting the Government of Canada's Relationship with Indigenous Peoples," 2018, https://www.justice.gc.ca/eng/csj-sjc/principles.pdf (accessed February 17, 2020), p. 12; UN Economic Commission for Africa, Guiding Principles on Large Scale Land Based Investments in Africa (2014), https://www.uneca.org/sites/default/files/PublicationFiles/guiding_principles_eng_rev_era_size.pdf (accessed February 17, 2020) p. 22.