KEY ELEMENTS OF A TREATY

ON FULLY AUTONOMOUS WEAPONS

CAMPAIGN TO STOP
KILLER ROBOTS

KEY ELEMENTS OF A TREATY ON FULLY AUTONOMOUS WEAPONS

The increasing technological capacity for autonomy in weapons systems raises a host of moral, legal, accountability, technological, and security concerns. Weapons systems that select and engage targets without meaningful human control—known as fully autonomous weapons, lethal autonomous weapons systems, or killer robots—would cross the threshold of acceptability and should be prevented and prohibited through new international law.

The Campaign to Stop Killer Robots is calling for a legally binding instrument to address such emerging technology by preserving meaningful human control over the use of force. The instrument should apply to the range of weapons systems that select and engage targets on the basis of sensor inputs, that is, systems where the object to be attacked is determined by sensor processing, not by humans.[1] This broad scope is designed to ensure problematic technology does not escape regulation.

The treaty's restrictions, however, would focus on those systems that contravene the requirement of meaningful human control. It would use a combination of prohibitions and positive obligations effectively to ban systems that amount to, or are used as fully autonomous weapons. While specific language and content would have to be worked out during multilateral discussions and treaty negotiations, the final instrument should incorporate the key elements identified in this paper.

[1] For more on this categorization, see Richard Moyes, Article 36, "Target Profiles," August 2019, http://www.article36.org/wp-content/uploads/2019/08/Target-profiles.pdf, p. 3.

This paper examines the concept of meaningful human control, which would be central to the new treaty or protocol. It then proposes three types of core obligations:

- A general obligation to maintain meaningful human control over the use of force;

- Prohibitions (i.e., negative obligations) on weapons systems that select and engage targets and by their nature pose fundamental moral or legal problems; and

- Specific positive obligations to help ensure that meaningful human control is maintained in the use of all other systems that select and engage targets.

THE CONCEPT OF MEANINGFUL HUMAN CONTROL

The proposed legally binding instrument should focus on meaningful human control because many of the concerns raised by fully autonomous weapons are attributable to the absence of such control. This absence would undermine human dignity by delegating life-and-death determinations to inanimate machines that reduce humans to datapoints yet could not comprehend the value of human life. Such weapons systems would also lack the capacity for human judgment necessary, for example, to weigh the proportionality of an attack, as required under international law. Furthermore, it would be legally difficult and arguably unjust to hold a human liable for the actions of a system operating beyond his or her control.[2]

For these and other reasons, states as well as international and non-governmental organizations have expressed widespread agreement about the need for some form of human control over the use of force. Their choice of terminology and specific views of the human role may differ, but they have identified many of the same factors. Drawing on international discussions and numerous publications, this paper distills the concept of meaningful human control into decision-making, technological, and operational components.[3]

[2] For more information on the problems of fully autonomous weapons, see Human Rights Watch and the Harvard Law School International Human Rights Clinic, *Making the Case: The Dangers of Killer Robots and the Need for a Preemptive Ban* (2016), https://www.hrw.org/report/2016/12/09/making-case/dangers-killer-robots-and-need-preemptive-ban.

[3] While there are different ways to frame this concept, the phrase “meaningful human control” has many advantages. “Control” is a term widely used in international law and is stronger and broader than the alternatives proposed by a few states, such as intervention and judgment. The qualifier “meaningful” works to ensure that control is substantive rather than superficial and is less context specific or outcome driven than alternatives like appropriate and effective.

DECISION-MAKING 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. In particular, the human operator of a weapon system should have: an understanding of the operational environment; an understanding of how the system functions, including what it might identify as a target; and sufficient time for deliberation.

TECHNOLOGICAL COMPONENTS

Technological components are embedded features of a weapon system that can enhance meaningful human control. They include: predictability and reliability;[4] 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.

OPERATIONAL COMPONENTS

Operational components make human control more meaningful by limiting when and where a weapon system can operate and what it can target. Factors that could be constrained include: the time between a human's legal assessment and the system's application of force; the duration of the system's operation; the nature and size of the geographic area of operation; and the permissible types of targets (e.g., personnel or material).

While none of these components are independently sufficient to amount to meaningful human control, all have the potential to enhance control in some way. In addition, the components often work in tandem. Further analysis of existing and emerging technology could help determine which these or other components should be codified in a legal instrument as prerequisites for meaningful human control.

[4] In general, predictability refers to the degree to which a weapon system operates as humans expect, and reliability refers to the degree to which the system will perform consistently. International Committee of the Red Cross statement under Agenda Item 5(b), Convention on Conventional Weapons Group of Governmental Experts on Lethal Autonomous Weapons Systems, Geneva, March 2019.

CORE OBLIGATIONS OF THE TREATY

The heart of the legally binding instrument should consist of three core obligations: a general obligation along with prohibitions and positive obligations to implement it.

A GENERAL OBLIGATION TO MAINTAIN MEANINGFUL HUMAN CONTROL OVER THE USE OF FORCE

This overarching provision would facilitate compliance with applicable legal and ethical norms by obliging states parties to maintain meaningful human control over the use of force. The generality of the obligation would help avoid loopholes, and the principle it embodies could inform interpretation of the treaty's other provisions. As noted above, most states have already expressed support for a requirement of human control.

The general obligation should focus on control over conduct ("use of force") rather than specific technology. This approach would help future-proof the treaty by obviating the need to predict how technology will develop. The term "use of force" also makes the general obligation applicable to situations of armed conflict and law enforcement.[5]

[5] While the term "use of force" frequently appears in discussions and documents of international humanitarian law and international human rights law, the two bodies of law govern it is somewhat different ways. The new treaty may need to take such differences into account.

PROHIBITIONS ON SPECIFIC WEAPONS SYSTEMS THAT SELECT AND ENGAGE TARGETS AND BY THEIR NATURE POSE FUNDAMENTAL MORAL OR LEGAL PROBLEMS

The treaty should prohibit the development, production, and use of weapons systems that select and engage targets and are inherently unacceptable for ethical or legal reasons. The clarity of the prohibitions would facilitate monitoring and enforcement, and their absoluteness would create a strong stigma against the banned systems.

The new instrument should prohibit weapons systems that by their nature select and engage targets without meaningful human control. The prohibition should cover, for example, systems that become too complex for human users to understand and thus produce unpredictable and inexplicable effects. These complex systems might apply force based on prior machine learning or allow critical system parameters to change without human authorization. Such weapons systems would run afoul of the new instrument's general obligation discussed above.

The prohibitions could also extend to specific other weapons systems that select and engage targets and are by their nature, rather than their manner of use, problematic. In particular, the treaty could prohibit weapons systems that select and engage humans as targets, regardless of whether they operate under meaningful human control.[6] Such systems would rely on certain types of data, such as weight, heat, or sound, to represent people or categories of people. In killing or injuring people based on such data, these systems would contravene the principle of human dignity and dehumanize violence. A prohibition on this category of systems would also encompass systems that, deliberately or unintentionally, target groups of people based on discriminatory indicators related to age, gender, or other social identities.

[6] For more information on such systems and the proposal to prohibit them, see generally Moyes, "Target Profiles."

SPECIFIC POSITIVE OBLIGATIONS TO ENSURE THAT MEANINGFUL HUMAN CONTROL IS MAINTAINED IN THE USE OF ALL OTHER SYSTEMS THAT SELECT AND ENGAGE TARGETS

The new instrument's positive obligations should cover weapons systems that are not inherently unacceptable but that might still have the potential to select and engage targets without meaningful human control. The obligations would require states parties to ensure that weapons systems that select and engage targets 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 system operating as fully autonomous weapons unlawful.

OTHER ELEMENTS

While the key elements outlined above are critical to achieving the objectives of the new instrument, other elements should complement them. For example, a preamble should articulate the purpose of the treaty and place it in the context of relevant law. Reporting requirements would promote transparency and facilitate independent monitoring. Detailed verification measures or cooperative compliance mechanisms would help prevent violations of the treaty. Regular meetings of states parties would provide an opportunity to review the status and operation of the treaty, identify implementation gaps, and set goals for the future. Other important elements would include a requirement to adopt national implementation measures and a threshold for entry into force.

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.

Retain meaningful human control over the use of force.
Prohibit fully autonomous weapons.
WWW.STOPKILLERROBOTS.ORG