NEW WEAPONS, PROVEN PRECEDENT

Elements of and Models for a Treaty on Killer Robots

HUMAN RIGHTS WATCH

IHRC
INTERNATIONAL HUMAN RIGHTS CLINIC
HUMAN RIGHTS PROGRAM AT HARVARD LAW SCHOOL

HUMAN RIGHTS WATCH

IHRC

INTERNATIONAL HUMAN RIGHTS CLINIC

HUMAN RIGHTS PROGRAM AT HARVARD LAW SCHOOL

New Weapons, Proven Precedent

Elements of and Models for a Treaty on Killer Robots

Printed in the United States of America
ISBN: 978-1-62313-8745
Cover design by Rafael Jimenez

Human Rights Watch defends the rights of people worldwide. We scrupulously investigate abuses, expose the facts widely, and pressure those with power to respect rights and secure justice. Human Rights Watch is an independent, international organization that works as part of a vibrant movement to uphold human dignity and advance the cause of human rights for all.

Human Rights Watch is an international organization with staff in more than 40 countries, and offices in Amsterdam, Beirut, Berlin, Brussels, Chicago, Geneva, Goma, Johannesburg, London, Los Angeles, Moscow, Nairobi, New York, Paris, San Francisco, Sydney, Tokyo, Toronto, Tunis, Washington DC, and Zurich.

For more information, please visit our website: http://www.hrw.org

The International Human Rights Clinic (IHRC) at Harvard Law School seeks to protect and promote human rights and international humanitarian law through documentation; legal, factual, and strategic analysis; litigation before national, regional, and international bodies; treaty negotiations; and policy and advocacy initiatives. IHRC also engages in innovative clinical education to develop advanced practice techniques and approaches to human rights advocacy. IHRC's Armed Conflict and Civilian Protection Initiative (ACCPI) focuses on humanitarian disarmament and other measures to reduce the civilian suffering caused by armed conflict.

For more information, please visit IHRC's website: http://hrp.law.harvard.edu/clinic/

OCTOBER 2020

HUMAN RIGHTS WATCH

IHRC

INTERNATIONAL HUMAN RIGHTS CLINIC
HUMAN RIGHTS PROGRAM AT HARVARD LAW SCHOOL

ISBN: 978-1-62313-8745

New Weapons, Proven Precedent

Elements of and Models for a Treaty on Killer Robots

Summary

Fully autonomous weapons would usher in a new era of armed conflict, similar to the advent of air warfare or the proliferation of nuclear weapons. Also known as lethal autonomous weapons systems or "killer robots," these systems would select and engage targets without meaningful human control. The prospect of delegating life-and-death decisions to machines raises a host of moral, legal, accountability, and security concerns, and the systems' rapid development presents one of the most urgent challenges facing the world today. Since 2014, the Convention on Conventional Weapons (CCW) has held eight meetings, attended by more than 100 countries, to discuss these concerns, but the gravity of the problem warrants a much more urgent response.

A majority of CCW states parties and the Campaign to Stop Killer Robots, a global civil society coalition coordinated by Human Rights Watch, are calling for the negotiation of a legally binding instrument to prohibit or restrict lethal autonomous weapons systems. The Campaign advocates for a treaty to maintain meaningful human control over the use of force and prohibit weapons systems that operate without such control. While the exact language would be worked out during negotiations, the Campaign identified key elements of such a treaty in a November 2019 publication, prepared by Human Rights Watch and the Harvard Law School International Human Rights Clinic.[1]

While some states have suggested that the cutting-edge nature of fully autonomous weapons will significantly complicate the treaty process, drafters of an instrument on the topic can look to existing international law and principles for guidance. These weapons systems present distinctive challenges, and no single source constitutes a model response, but creating new law from scratch could unnecessarily slow the progress of negotiations. International law and non-legally binding principles of artificial intelligence (AI) provide ample precedent for the elements of a new treaty. Lessons from the past can and should be adapted to this emerging technology.

[1] Campaign to Stop Killer Robots, "Key Elements of a Treaty on Fully Autonomous Weapons," November 2019, https://www.stopkillerrobots.org/wp-content/uploads/2020/04/Key-Elements-of-a-Treaty-on-Fully-Autonomous-WeaponsvAccessible.pdf (accessed September 3, 2020). See also Bonnie Docherty, "The Need for and Elements of a New Treaty on Fully Autonomous Weapons," in *Rio Seminar on Autonomous Weapons* (Fundação Alexandre de Gusmão: Brasília, 2020), http://funag.gov.br/biblioteca/download/laws_digital.pdf (accessed October 3, 2020), pp. 223-234.

This report provides precedent for each of the treaty elements and shows that constructing a legally binding instrument does not require an entirely new approach. Earlier law and principles, often driven by similar concerns and objectives, can inform the structure of a treaty on fully autonomous weapons, and when negotiations start, facilitate crafting of language. The existence of relevant models should make it legally, politically, and practically easier to develop a new treaty.

Elements of a New Treaty

The proposed treaty elements apply to all weapons systems that select and engage targets based on sensor processing, rather than human inputs. They include three types of obligations. First, a general obligation requires maintaining meaningful human control over the use of force. Second, prohibitions ban the development, production, and use of weapons systems that autonomously select and engage targets and by their nature pose fundamental moral or legal problems. These prohibitions cover weapons that always operate without meaningful human control and those that rely on data, like weight, heat, or sound, to select human targets. Third, specific positive obligations aim to 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, another key element of the treaty, cuts across all three types of obligations. It can be distilled into decision-making, technological, and operational components.

Methodology

This report elaborates on the content and rationale for the treaty elements listed above. It then presents parallels with existing law and principles. It draws on two main sources.

The report examines international legal instruments, especially international humanitarian law and disarmament treaties. The instruments include Additional Protocol I to the Geneva Conventions, a cornerstone of civilian protection, the Arms Trade Treaty, and numerous conventions banning specific weapons. The report also considers relevant precedent from international human rights law and international environmental law. These legal sources provide especially useful support for the obligations and terminology proposed for the treaty.

Given that emerging weapons systems raise some novel issues for international law, the report also considers principles from the technology sector regarding the development, deployment, and governance of artificial intelligence. The report draws in particular from "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI," a white paper produced by the Berkman Klein Center at Harvard Law School, which analyzes 36 "AI principles" documents issued by governments, the private sector, and civil society from around the world.[2] That paper identifies common themes and principles that cut across the diverse documents. The conclusion of "Principled Artificial Intelligence" and the sources it cites support the objectives of the proposed prohibitions and the understanding of the concept of meaningful human control.

[2] Jessica Fjeld et al., "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI," *Berkman Klein Center Research Publication 2020-1* (2020), accessed September 3, 2020, doi: 10.2139/3518482. The documents were curated for variety in the sectors, geographies, and approaches represented, with a focus on documents that have been especially visible or influential.

Recommendations

To increase momentum for adoption of a timely and effective legally binding instrument on fully autonomous weapons, states should:

- Agree to launch negotiations by the end of 2021, with the aim of swiftly adopting a new international treaty to retain meaningful human control over the use of force and prohibit weapons systems that lack such control.
- Consider and build on the precedent provided by earlier treaties and normative frameworks to address the concerns posed by fully autonomous weapons and expedite their work towards a new treaty.
- Articulate their national positions on the structure and content of a new treaty.

I. General Obligation

At the heart of the proposed elements for a treaty on fully autonomous weapons rests a general obligation for states parties to "maintain meaningful human control over the use of force."[3] It establishes an overarching principle that can close unexpected loopholes in the treaty's other provisions and guide interpretation of the prohibitions and positive obligations. The general obligation's focus on control over conduct ("use of force") rather than control over a specific system helps future-proof the treaty by obviating the need to foresee all possible technologies in a rapidly developing field. In addition, because the term "use of force" is used in both international humanitarian law (the laws of war) and international human rights law, the general obligation ensures that the treaty applies to situations of armed conflict and law enforcement operations.[4] Finally, regulating conduct allows the obligation to cover algorithmic decision-making throughout the targeting process, and thus reflects modern targeting practices, which are characterized by distributed decision-making across actors and technologies.

The general obligation also references "meaningful human control," a concept that is present in all of the proposed treaty elements. While states have used different terms to describe the human role in the use of force, virtually all agree a role is necessary. The content of meaningful human control and the advantages of the specific term are discussed in the last section of this report.

Other international humanitarian law treaties have used general obligations to lay out the core purpose and foundational principles of an instrument and to inform interpretation of more specific provisions. The 1977 Additional Protocol I to the Geneva Conventions (Protocol I) provides apt precedent because its origins resemble those of the proposed fully autonomous weapons treaty. Protocol I was drafted in part to respond to changes in the nature of warfare and developments in weapons technology. Aerial bombing, for example, did not exist at the time of the 1907 Hague Regulations, the previous

[3] Campaign to Stop Killer Robots, "Key Elements of a Treaty on Fully Autonomous Weapons."

[4] Although international humanitarian law and international human rights law govern the use of force in somewhat different ways, the new treaty can take such differences into account.

international effort to address the methods and means of warfare.[5] The development of autonomy in weapons systems likewise creates an impetus to clarify and strengthen the law.

The structure of Protocol I's section on protecting civilians from the effects of hostilities parallels that of the proposed elements for a fully autonomous weapons treaty. Article 48 constitutes a general obligation akin to the one requiring meaningful human control over the use of force. The article seeks to further the protocol's objective of strengthening civilian protection by stating the principle of distinction, which is fundamental to international humanitarian law: "[T]he Parties to the conflict shall at all times distinguish between the civilian population and combatants." Protocol I then unpacks this general obligation through a number of more specific provisions, akin to the prohibitions and positive obligations on fully autonomous weapons discussed below. For example, Article 51(4-5) prohibits indiscriminate and disproportionate attacks, which violate the principle of distinction, and Article 57 obliges parties to a conflict to take affirmative precautions to spare civilians and civilian objects. The principle of distinction is recognized as customary international law, binding on all warring parties.

The Arms Trade Treaty also contains a general provision that informs interpretation of the instrument. Article 1, along with the treaty's preamble and enumerated principles, articulates the treaty's two-part objective: to "establish the highest possible common international standards for regulating [the arms trade]," and to "prevent and eradicate the illicit trade in conventional arms and prevent their diversion." Strictly speaking, the article does not create an obligation, but by communicating the treaty's overarching goals, it serves as a touchstone for interpretation of other provisions, which identify specific actions that states must undertake or refrain from taking.[6]

Several weapons ban treaties, including the Chemical Weapons Convention, Mine Ban Treaty, and Convention on Cluster Munitions, contain articles entitled "General Obligations." Rather than presenting overarching principles, these provisions lay out

[5] Philip Spoerri, Director of International Law, International Committee of the Red Cross, "The Geneva Conventions of 1949: Origins and Current Significance," address at a ceremony to celebrate the 60th anniversary of the Geneva Conventions, August 12, 2009, https://www.icrc.org/en/doc/resources/documents/statement/geneva-conventions-statement-120809.htm (accessed October 3, 2020).

[6] See Andrew Clapham et al., *The Arms Trade Treaty: A Commentary* (Oxford: Oxford University Press, 2016), paras. 1.01, 1.03.

specific prohibitions and highlight certain positive obligations.[7] Such prohibitions and positive obligations will be explored in greater detail in the next two sections.

[7] Article 1 of the Chemical Weapons Convention lists as general obligations states' commitments not to use, develop, produce, stockpile, or transfer chemical weapons as well as commitments to destroy chemical weapons and chemical weapons production facilities. Article 1 of the Mine Ban Treaty enumerates prohibitions and a positive obligation to "destroy or ensure the destruction of all anti-personnel mines." Article 1 of the Convention on Cluster Munitions is entitled "General Obligations and Scope of Application," yet outlines the prohibitions established by the convention. Prohibitions will be explored in greater detail in the following section.

II. Prohibitions

The prohibitions proposed for a fully autonomous weapons treaty cover "the development, production, and use of weapons systems that select and engage targets and are inherently unacceptable for ethical or legal reasons." They prohibit systems that "pose fundamental moral or legal problems" by their nature rather than due to their manner of use. Clear prohibitions make implementation, monitoring, and enforcement easier. They also create a strong stigma against the banned weapons systems, which can influence even states not party and non-state actors.

Earlier disarmament treaties provide precedent for banning weapons that are morally and/or legally problematic. The Mine Ban Treaty seeks to "put an end to the suffering and casualties caused by anti-personnel mines."[8] The Oslo Declaration, which launched the process to negotiate the Convention on Cluster Munitions, committed states to developing a treaty that prohibited cluster munitions that caused "unacceptable harm to civilians."[9] The convention's preamble reaffirms the Oslo Declaration and indicates that states found unacceptable "the suffering … caused by cluster munitions," including death and injury, interference with economic and social development, and obstacles to the return of displaced persons.[10] The Treaty on the Prohibition of Nuclear Weapons (TPNW) highlights the "unacceptable suffering" experienced by victims of use and testing and acknowledges the "ethical imperatives for nuclear disarmament."[11] All three treaties also place their obligations in a legal context, underscoring the importance of fundamental international humanitarian law principles, such as distinction. The TPNW in particular states that any use of the weapons would violate international humanitarian law.[12]

Specific references to the Martens Clause reinforce that the weapons prohibited by these treaties are morally and legally problematic. The Martens Clause is a provision of

[8] Convention on the Prohibition of the Use, Stockpiling, Production and Transfer of Anti-Personnel Mines and on Their Destruction (Mine Ban Treaty), adopted September 18, 1997, entered into force March 1, 1999, pmbl., para (1).

[9] "Oslo Declaration," Oslo Conference on Cluster Munitions, February 22-23, 2007, reprinted in Gro Nystuen and Stuart Casey-Maslen (eds), *The Convention on Cluster Munitions: A Commentary* (Oxford University Press, 2010) annex 2, p. 631.

[10] Convention on Cluster Munitions, adopted May 30, 2008, entered into force August 1, 2010, pmbl., paras. 18, 2.

[11] Treaty on the Prohibition of Nuclear Weapons (TPNW), adopted July 7, 2017, opened for signature September 20, 2017, pmbl., paras. 6, 5.

[12] Ibid., pmbl., para. 10.

international humanitarian law that requires states to consider ethical standards, i.e., the dictates of public conscience and principles of humanity, in the protection of civilians.[13] In its preamble, the Mine Ban Treaty suggests that antipersonnel landmines contravene these standards by noting that the call for a ban evinced the public conscience and furthered the principles of humanity.[14] The TPNW "reaffirm[s] that any use of nuclear weapons would … be abhorrent to the principles of humanity and the dictates of public conscience."[15]

The proposed prohibitions on fully autonomous weapons are motivated by similar factors. Given that these systems would have the power to kill without meaningful human control, they would cross a moral redline for many people, face significant challenges in complying with international law, and raise concerns under the Martens Clause.

Scope of the Prohibitions

The prohibitions in disarmament treaties cover a range of activities that parallel the proposed ban on the development, production, and use of fully autonomous weapons. The Chemical Weapons Convention, Mine Ban Treaty, and Convention on Cluster Munitions ban use, development, production, acquisition, stockpiling, retention, and transfer as well as assistance with those prohibited activities.[16] The TPNW contains similar prohibitions and adds testing and threatening to use nuclear weapons. These treaties thus recognize that banning use alone is insufficient to address the problems of an unacceptable weapon system.

All four treaties prohibit these activities "under any circumstances." As a result, they apply in times of peace and war. This broad scope is important in the fully autonomous weapons context given that the systems could be used in law enforcement operations as well as in situations of armed conflict.

Precedent shows that such comprehensive prohibitions work even for weapons systems incorporating dual-use technology. Drafters of earlier disarmament treaties created a clear

[13] The Martens Clause was introduced in the preamble of the 1899 Hague Convention II with Respect to the Laws of War, and it appeared in a slightly modified form in the 1907 Hague Regulations. It has since appeared in Article 1(2) of Protocol I and the preamble to Protocol II to the Geneva Convention as well as in other international humanitarian law treaties.

[14] Mine Ban Treaty, pmbl., para. 8.

[15] TPNW, pmbl., para. 11.

[16] The Biological Weapons Convention bans development, production, stockpiling or otherwise acquiring or retention. Convention on the Prohibition of the Development, Production and Stockpiling of Bacteriological (Biological) and Toxin Weapons and on their Destruction, adopted December 17, 1972, entered into force March 26, 1975, art. I.

divide between weaponized and peaceful technology, protecting the development of the latter. The Chemical Weapons Convention addresses concerns about stigmatizing constructive uses of chemicals by applying its prohibitions only to those toxic chemicals and associated equipment that are designed to be used as weapons. In its preamble, it expresses states parties' desire to "promote free trade in chemicals as well as international cooperation and exchange of scientific and technical information in the field of chemical activities for purposes not prohibited under this Convention in order to enhance the economic and technological development of all States Parties."[17] While prohibiting numerous activities related to biological weapons, the Biological Weapons Convention allows states parties to divert their stockpiles to "peaceful purposes."[18] In addition, Article X of this instrument requires states parties to "facilitate" the "fullest possible exchange of equipment, materials and scientific and technological information" for peaceful purposes, such as preventing disease, and declares that states parties have the right to participate in these exchanges. The TPNW's preamble refers to the right to "research, production and use of nuclear energy for peaceful purposes without discrimination."[19] Drafters of a treaty on fully autonomous weapons can look to these models to ensure a ban on the systems does not chill the development of autonomous technology for other purposes.

Protocol IV to the Convention on Conventional Weapons (CCW) exemplifies a treaty governing technology that is emerging as well as dual use.[20] During the 1990s, states, international organizations, and civil society groups expressed outrage at the prospect of new laser weapons that could cause permanent blindness. Like opponents to fully autonomous weapons, they argued that blinding lasers were "abhorrent" weapons that would violate the Martens Clause, endanger civilians, and proliferate to irresponsible actors. In 1995, CCW states parties adopted a preemptive ban on the use and transfer of blinding laser weapons, and "for the first time in human history, an inhumane weapon had

[17] Convention on the Prohibition of the Development, Production, Stockpiling and Use of Chemical Weapons and on Their Destruction (Chemical Weapons Convention), adopted January 13, 1993, entered into force April 29, 1997, pmbl., para. 10.

[18] Convention on the Prohibition of the Development, Production and Stockpiling of Bacteriological (Biological) and Toxin Weapons and on their Destruction (Biological Weapons Convention), signed April 10, 1972, entered into force March 26, 1975, art. II.

[19] TPNW, pmbl., para. 22.

[20] Convention on Certain Conventional Weapons Protocol IV on Blinding Laser Weapons, adopted October 13, 1995, entered into force October 7, 1998, art. 1. For more details on the specific humanitarian concerns posed by these weapons systems, see Human Rights Watch and Harvard Law School International Human Rights Clinic (IHRC), "Precedent for Preemption: The Ban on Blinding Lasers as a Model for a Killer Robots Prohibition," November 2015, https://www.hrw.org/sites/default/files/supporting_resources/robots_and_lasers_final.pdf.

been declared illegal and prohibited before it had actually been used."[21] Despite fears that a prohibition would interfere with the use of non-weaponized laser technology, lasers continue to be used widely for peaceful purposes, such as corrective eye surgery and other medical procedures. Lasers are also used by militaries against anti-materiel targets. While blinding lasers are a narrower class of weapons than fully autonomous weapons, the parallels show that drawing the line on problematic emerging technologies through prohibitions is feasible and effective.

Control-Based Prohibitions

The proposed prohibitions in the fully autonomous weapons treaty encompass two major categories of systems. First, the treaty would ban weapons systems that "by their nature select and engage targets without meaningful human control." The prohibition should cover, for example, complex systems that, due to their machine-learning algorithms, would produce unpredictable or inexplicable effects.

The lack of human control in weapons systems has motivated several prior disarmament treaties.[22] Fully autonomous weapons raise similar concerns as victim-activated landmines, biological weapons, and chemical weapons because the systems have the ability to take life without meaningful human control.

Maintaining control is an underlying principle of the Mine Ban Treaty. The treaty prohibits antipersonnel mines, which are victim-activated explosive devices "designed to be exploded by the presence, proximity or contact of a person."[23] Such landmines endanger civilians and violate international humanitarian law's principle of distinction because humans cannot control when they explode. The Mine Ban Treaty does not apply to command-detonated mines. Human operators detonate these mines by remote control and thus can determine whom the mines target and when they explode. As one state party explained to the Landmine Monitor, such command-detonated mines are "designed to be placed on the ground, aimed and controlled by a soldier who assesses the situation and

[21] Summary Record (Partial) of the 13th Meeting, Review Conference of the States Parties to the Convention on the Prohibitions or Restrictions on the Use of Certain Conventional Weapons which may be Deemed to be Excessively Injurious or to have Indiscriminate Effects, CCW/CONF.I/SR.13, May 3, 1996, para. 69.

[22] Human Rights Watch and IHRC, "Killer Robots and the Concept of Meaningful Human Control," April 2016, https://www.hrw.org/sites/default/files/supporting_resources/robots_meaningful_human_control_final.pdf, p. 10.

[23] Mine Ban Treaty, art. 2(1).

makes a deliberate decision as to detonation."[24] In differentiating between the two types of landmines, the Mine Ban Treaty highlights the significance of the element of control in defining a category of problematic weapons.

The Biological Weapons Convention and Chemical Weapons Convention were also driven in part by concerns about the controllability of the weapons. A 1969 report from the UN secretary-general commented that "controllability ... is a most important consideration in their [biological and chemical agents'] use as weapons."[25] The report observes that once in the atmosphere, biological and chemical agents can be dispersed by elements of nature, such as the wind, and "control is thus possible only to the extent that the meteorological situation can be predicted." The United Nations General Assembly subsequently adopted a resolution declaring the use of biological and chemical weapons to be counter to general principles of international law, explicitly noting that the weapons' "effects are often uncontrollable and unpredictable and may be injurious without distinction to combatants and non-combatants."[26] States adopted the treaty banning biological weapons in 1971 and the treaty banning chemical weapons in 1993.

Human control is an underlying theme of many existing sets of artificial intelligence principles. In an analysis of 36 such "principles documents," the Berkman Klein Center white paper "Principled Artificial Intelligence" explains: "our society, governments, and companies alike are grappling with a potential shift in the locus of control from humans to AI systems."[27] The concept of control is closely linked to other cross-cutting themes the paper identifies because "human involvement is often presented as a mechanism to accomplish those ends." These documents provide additional support for the content of meaningful human control, which is discussed in more detail below.

[24] International Campaign to Ban Landmines, "Country Profile: Canada," in *Landmine Monitor 2002*, eds. Stephen Goose et al. (New York: Human Rights Watch, August 2002), http://archives.the-monitor.org/index.php/publications/display?url=lm/2002/canada.html#fnB858 (accessed October 5, 2020) (quoting "ILX0149: Response to Query," email to MAC from Shannon Smith, DFAIT/ILX, May 2, 2002, and also citing "The Canadian Forces and Anti-Personnel Landmine," DND document BG-02.007, February 13, 2002).

[25] United Nations Secretary-General, "Chemical and Bacteriological (Biological) Weapons and the Effects of their Possible Use: Report of the Secretary-General" Doc.A/7575/Rev.1, (1969), http://repository.un.org/bitstream/handle/11176/75955/A_7575_Rev.1%3bS_9292_Rev.1-EN.pdf?sequence=10&isAllowed=y (accessed April 23, 2020), para. 28.

[26] United Nations General Assembly, Resolution 2603 (1969), A/7890. The resolution was adopted by a vote of 80 in favor to 3 against; 36 states abstained from voting and 7 states were non-voting.

[27] Fjeld et al., "Principled Artificial Intelligence," p. 53.

Human Dignity-Based Prohibitions

The proposed elements of a fully autonomous weapons treaty also prohibit a second category of weapons systems that select and engage humans as targets in morally or legally unacceptable ways. Such systems would rely on target profiles, i.e., "certain types of data, such as weight, heat, or sound, to represent people or categories of people."[28] In killing or injuring people based on such data, these systems would violate human dignity and dehumanize violence. They also have the potential to be discriminatory because they could be programmed to target people based on discriminatory indicators, such as age, gender, race, or other social identities, or algorithmic bias could lead to discriminatory results.[29]

International human rights law provides long-standing precedent for protecting human dignity. The opening words of the Universal Declaration of Human Rights assert that "recognition of the inherent dignity and of the equal and inalienable rights of all members of the human family is the foundation of freedom, justice and peace in the world."[30] In ascribing inherent dignity to all human beings, the Universal Declaration implies that everyone has worth that deserves respect.[31] The International Covenant on Civil and Political Rights (ICCPR) establishes the inextricable link between dignity and human rights, stating in its preamble that the rights it enumerates "derive from the inherent dignity of the human person."[32]

International human rights law also emphasizes the importance of non-discrimination. The Universal Declaration of Human Rights states, "All are equal before the law and are entitled without any discrimination to equal protection of the law."[33] The ICCPR elaborates on prohibited criteria for distinguishing people, including "race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other

[28] Richard Moyes, Article 36, "Target Profiles," August 2019, http://www.article36.org/wp-content/uploads/2019/08/Target-profiles.pdf (accessed September 8, 2020), p. 3.

[29] For more information on algorithmic biases, see generally UN Institute for Disarmament Research (UNIDIR), *Algorithmic Bias and the Weaponization of Increasingly Autonomous Technologies* (2018), https://unidir.org/sites/default/files/publication/pdfs/algorithmic-bias-and-the-weaponization-of-increasingly-autonomous-technologies-en-720.pdf (accessed September 18, 2020).

[30] Universal Declaration of Human Rights, adopted December 10, 1948, G.A. Res. 217A(III), U.N. Doc. A/810, pmbl., para. 1.

[31] Jack Donnelly, "Human Dignity and Human Rights," in *Protecting Dignity: Agenda for Human Rights*, Government of Switzerland, June 2009, https://www.legal-tools.org/doc/e80bda/pdf/ (accessed February 16, 2014), p. 10.

[32] International Covenant on Civil and Political Rights (ICCPR), adopted December 16, 1966, G.A. Res. 2200A (XXI), 21 U.N. GAOR Supp. (No. 16) at 52, U.N. Doc. A/6316 (1966), 999 U.N.T.S. 171, entered into force March 23, 1976, pmbl., para. 2.

[33] Universal Declaration of Human Rights, art. 7.

status."[34] Later instruments reiterate these criteria and enumerate other ones, such as disability and sexual orientation.[35]

Dignity and non-discrimination play a role in humanitarian disarmament law, which seeks to reduce arms-inflicted human suffering. The preamble of the Convention on Cluster Munitions, for example, expresses states parties' determination to "ensure the full realisation of the rights of all cluster munition victims and recognis[e] their inherent dignity." The belief in human dignity was the foundation of the "rights-based" approach to victim assistance, which is reaffirmed and articulated in Article 5.[36] The Convention on Cluster Munitions and the Treaty on the Prohibition of Nuclear Weapons further specify that victim assistance should be provided without discrimination. While the principles of dignity and non-discrimination must be upheld during decision-making about the use of force as well as in remedial measures after use, the reference to the principles in these disarmament treaties bolsters the case for them playing a role in a fully autonomous weapons instrument.

The frequent references to these principles in existing artificial intelligence documents reinforce their relevance to a treaty on fully autonomous weapons. Of the principles documents surveyed in "Principled Artificial Intelligence," close to half stress the importance of prioritizing human values, including human dignity, which is often tied to human rights.[37] In addition, almost 90 percent of the documents studied by the Berkman Klein Center include the principle of "non-discrimination and the prevention of bias."

[34] Under Article 2 of the ICCPR, each state "undertakes to respect and to ensure to all individuals within its territory and subject to its jurisdiction the rights recognized in the present Covenant, without distinction of any kind, such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status." ICCPR, art. 2. Article 26 reads that, "All persons are equal before the law and are entitled without any discrimination to the equal protection of the law. In this respect, the law shall prohibit any discrimination and guarantee to all persons equal and effective protection against discrimination on any ground such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status." ICCPR, art. 26.

[35] Convention on the Rights of Persons with Disabilities (CRPD), adopted May 3, 2008, 2515 U.N.T.S. 3, entered into force May 3, 2008, art. 5(2); UN Human Rights Council, "Human Rights, Sexual Orientation and Gender Identity," Resolution 17/19, A/HRC/Res/17/19, pmbl.

[36] Gro Nystuen and Stuart Casey-Maslen eds., *The Convention on Cluster Munitions: A Commentary* (Oxford: Oxford University Press, 2010), p. 55. Drawing on the foundation of international human rights law, the rights-based approach recognizes that victims, as all other human beings or groups of human beings, are subjects or bearers of rights.

[37] Fjeld et al., "Principled Artificial Intelligence," pp. 60-61. For example, the European High-Level Expert Group guidelines uphold respect for human dignity as part of a family of fundamental rights set out in international human rights law particularly apt to cover AI systems. European Commission's High-Level Expert Group on Artificial Intelligence, "Ethics Guidelines for Trustworthy AI" (2019), https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (accessed September 9, 2020), p. 10.

According to its report, some of the documents warn that “AI is not only replicating patterns of bias, but also has the potential to significantly scale discrimination and to discriminate in unforeseen ways.”[38]

[38] Fjeld et al., “Principled Artificial Intelligence,” p. 48. The Toronto Declaration on the Right to Equality and Non-Discrimination in Machine Learning System, for example, states that “[a]ll actors, public and private, must prevent and mitigate against discrimination risks in the design, development and application of machine learning technologies. Amnesty International, Access Now, “Toronto Declaration: Protecting the Right to Equality and Non-Discrimination in Machine Learning Systems” (2018), https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf (accessed September 9, 2020), p. 6.

III. Positive Obligations

To further the general obligation and complement the prohibitions, the proposed elements of a new treaty include "specific positive obligations to ensure that meaningful human control is maintained in the use of all other systems that select and engage targets." The restrictions do not address weapons systems already covered by the prohibitions. Instead, they outline affirmative steps states parties would need to take to cover systems that are not *inherently* unacceptable but still have the potential to be *used* to select and engage targets without meaningful human control. The content of the positive obligations should draw on the concept of meaningful human control, which is discussed below. Specific positive requirements would bolster the strength of the treaty by regulating the use of emerging technologies in weapons that are not explicitly captured by the treaty's prohibitions and by being adaptable enough to address future technological developments.

Several international humanitarian law and weapons treaties include positive obligations that elaborate on actions states should take to implement and comply with a treaty's general obligation. Their positive obligations parallel the provisions proposed for the fully autonomous weapons treaty that mandate measures to ensure meaningful human control over the selection and engagement of targets.

Additional Protocol I to the Geneva Conventions enumerates both prohibitions and positive obligations that parties must follow to uphold Article 48 on distinction. Article 57, for example, requires parties to take all feasible precautions to avoid civilian casualties and damage to civilian objects. They must verify targets are not civilian ones, provide effective advanced warning of an attack when possible, and design attacks to "cause the least danger to civilian lives and to civilian objects."

The Arms Trade Treaty similarly enumerates positive obligations designed to operationalize the general principles outlined in Article 1. Articles 7 to 10 provide detailed requirements for states parties involved with the import, export, transit, or brokering of arms. Article 7 obliges exporting states to conduct a risk assessment evaluating the potential that conventional arms covered under the treaty may be used, for example, to commit violations of international humanitarian law or international human rights law, or to commit gender-based violence.[39] The exporting state party shall not authorize transfer if there is an “overriding risk” of these negative consequences, even with available mitigation measures. Article 11 outlines requirements for states parties to take measures to prevent diversion.[40]

Other types of positive obligations common to international humanitarian and human rights law may be useful to include. For example, reporting requirements would promote transparency and facilitate monitoring. Verification and compliance mechanisms could help prevent treaty violations. Regular meetings of states parties would provide opportunities to review the treaty’s status and operation, identify gaps in implementation, and set goals for the future. The details of these provisions are beyond the scope of this report, but there is ample precedent for them in many of the treaties discussed above.

[39] Arms Trade Treaty, adopted April 2, 2013, A/RES/ 67/234B, entered into force December 14, 2014, art. 7(4).

[40] Ibid., art. 11.

IV. Meaningful Human Control

The concept of meaningful human control cuts across all three proposed obligations of a treaty on fully autonomous weapons. The general obligation requires meaningful human control over the use of force. The prohibitions ban the development, production, and use of weapons systems that inherently lack meaningful human control. The positive obligations require states to ensure weapons systems that select and engage targets are used only with meaningful human control.

The concept is fundamental to this instrument because most of the concerns arising from the use of fully autonomous weapons are attributable to the lack of such human control.[41] For example, the use of fully autonomous weapons would undermine human dignity by delegating life-and-death determinations to machines that cannot comprehend the value of human life. Algorithmic bias in systems operating autonomously could lead to discriminatory outcomes. Fully autonomous weapons systems would also be unable to replicate the human judgment necessary to weigh the proportionality of an attack as required under international law. Even if the systems could apply human judgment, the law is designed to be implemented by humans. Finally, the use of fully autonomous weapons would create an accountability gap because it would be legally difficult and possibly unjust to hold a human liable for the actions of a system operating beyond human control. All of these concerns demonstrate the need to maintain meaningful human control over the use of force.

Support for maintaining meaningful human control in weapons systems has been widespread in multilateral meetings and expert reports. Almost all states that have spoken on the topic have argued that humans need to play a role in the use of force.[42] CCW states parties also agreed on a guiding principle that says: "Human-machine interaction … should ensure that the potential use of weapons systems based on emerging technologies in the area of lethal autonomous weapons systems is in compliance with applicable

[41] For a more detailed description of the concerns arising from the use of weapons systems without meaningful human control, see Human Rights Watch and IHRC, "Killer Robots and the Concept of Meaningful Human Control."

[42] Human Rights Watch, *Stopping Killer Robots: Country Positions on Banning Fully Autonomous Weapons and Retaining Human Control*, August 2020, https://www.hrw.org/sites/default/files/media_2020/08/arms0820_web_0.pdf.

international law, in particular [international humanitarian law]."[43] The arguments for meaningful human control are bolstered by both international law and documents setting out principles of artificial intelligence. As will be discussed below, these sources provide precedent for the content of the concept and the choice of terminology.

Components of Meaningful Human Control

The proposed treaty elements break meaningful human control into decision-making, technological, and operational components. None of these components would by itself be sufficient to make human control meaningful, but each would promote such control. While the components were initially distilled from the positions of states and experts, they also appear in many of the documents of AI principles analyzed in "Principled Artificial Intelligence." Their inclusion in a variety of AI standards underscores the value that states, private actors, and civil society groups place on the components and support their incorporation in the treaty.

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." The human operator should understand the operational environment and how the system functions, including what it might identify as a target, and should have sufficient time for deliberation.

Ensuring an understanding of how the system functions embodies the principle of "explainability." This principle seeks to counter "[p]erhaps the greatest challenge that AI poses from a governance perspective … [that is,] the complexity and opacity of the technology."[44] According to "Principled Artificial Intelligence," explainability in the AI context refers to the "translation of technical concepts and decision outputs into intelligible, comprehensible formats suitable for evaluation."[45] One of the principles documents, the Think 20 report "Future of Work and Education for the Digital Age,"

[43] Final Report, Meeting of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, Guiding Principles affirmed by the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, CCW/MSP/2019/CRP.2/Rev.1, November 2019, Annex III, para. (c).

[44] Fjeld et al., "Principled Artificial Intelligence," p. 41.

[45] Ibid., pp. 42–43.

highlights the importance of “clear, complete and testable explanations of what the system is doing and why.”[46] The Montreal Declaration for a Responsible Development of Artificial Intelligence finds that a satisfactory explanation “should take the same form as the justification we would demand of a human making the same kind of decision.”[47] Under existing principles, explainability is especially important in systems that cause harm or affect a person’s life.[48] Fully autonomous weapons would exemplify such systems.

Technological Components

Technological components are “embedded features of a weapons system that can enhance meaningful human control.” They can include reliability and predictability, the ability of the system to relay relevant information to the human operator, and the ability of a human to intervene after the activation of the system.

Several technological components are reflected in existing documents laying out AI principles. Reliability, for example, encompasses the widely cited principles of security and safety, which are related but distinct concepts.[49] According to “Principled Artificial Intelligence,” safety “refers to the proper *internal* functioning of an AI system and the avoidance of unintended harms,” while security “addresses *external* threats to an AI system.”[50] AI principles documents also highlight predictability and link that principle to public trust.[51] Predictability generally “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.”[52]

“Principled Artificial Intelligence” also draws a link between human control and the ability of humans to intervene. According to the report, human control “requires that AI systems

[46] Paul Twomey, “Future of Work and Education for the Digital Age,” (Think 20, 2018), https://www.g20-insights.org/wp-content/uploads/2018/07/TF1-1-11-Policy-Briefs_T20ARG_Towards-a-G20-Framework-For-Artificial-Intelligence-in-the-Workplace.pdf (accessed September 3, 2020), p. 7.

[47] University of Montreal, “Montreal Declaration for a Responsible Development of Artificial Intelligence,” 2018, https://www.montrealdeclaration-responsibleai.com/the-declaration (accessed September 3, 2020), p. 12.

[48] Fjeld et al., “Principled Artificial Intelligence,” p. 32.

[49] Ibid., p. 37.

[50] Ibid. (emphasis added).

[51] Ibid., p. 40 (citing, for example, the Beijing AI Principles).

[52] Campaign to Stop Killer Robots, “Key Elements of a Treaty on Fully Autonomous Weapons” (quoting 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), p. 4, n. 4.

are designed and implemented with the capacity for people to intervene in their actions."[53] Indeed the ability to intervene was "the most commonly referenced principle under the theme of Human Control of Technology."[54]

Operational Components

Operational components limit "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 materiel).

Such restrictions promote human control. As discussed in "Principled Artificial Intelligence," the Alisomar AI Principles state that that to ensure human control, "[h]umans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives."[55] The Montreal Declaration says that "AI systems should be built and used … 'with the goal of increasing people's control over their surroundings.'"[56] Operational constraints allow humans to control their surroundings and set the parameters for when they can delegate decisions.

Term "Meaningful Human Control"

While the human role in the use of force has been described in different terms, the phrase "meaningful human control" has several benefits. First, it has already been employed by a large number of states, international organizations, nongovernmental organizations, and other experts. Second, the specific words are well-suited to addressing the problems of fully autonomous weapons. According to the group Article 36, "meaningful," compared to other potential qualifiers, is "general rather than context specific (e.g. appropriate) [and] derives from an overarching principle rather than being outcome driven (e.g. effective, sufficient)."[57] "Control" is broader than alternative terms like judgment and intervention

[53] Fjeld et al., "Principled Artificial Intelligence," p. 54.

[54] Ibid.

[55] Future of Life Institute, "Asilomar AI Principles" (2017), https://futureoflife.org/ai-principles/?cn-reloaded=1 (accessed September 3, 2020); see also Fjeld et al., "Principled Artificial Intelligence," p. 54.

[56] Fjeld et al., "Principled Artificial Intelligence," p. 55.

[57] Article 36, "Key Elements of Meaningful Human Control," April 2016, http://www.article36.org/wp-content/uploads/2016/04/MHC-2016-FINAL.pdf (accessed September 3, 2020), p. 5.

because it encompasses both the application of human reasoning and actions to ensure human intention is followed. Third, and most relevant for this report, the concept of control is frequently used in international law and AI principles to promote accountability and reduce harm.[58]

International law often requires "control" to ensure legal responsibility. In international criminal law, under the concept of command responsibility, commanders or other superiors can be held criminally liable for the actions of subordinates over whom they have "effective command *and control*."[59] Additionally, in *jus ad bellum* and *jus in bello*, legal accountability often requires "effective control"[60] or "overall control."[61] An adjective such as "effective" or "overall" is included to emphasize that only substantive levels of control qualify. Mandating meaningful human control over the use of force or weapons systems that select and engage target objects similarly helps hold humans accountable for violations of international humanitarian and human rights law.

International law also imposes requirements to maintain control in order to reduce harm. Although they do not employ the word "control," as described above, the Mine Ban Treaty, the Biological Weapons Convention, and the Chemical Weapons Convention were motivated by desire to avoid harm caused by a lack of control. International environmental law explicitly requires states to "control" pollution and other causes of environmental damage in order to prevent and minimize harm to the environment.[62] For example, the United Nations Convention on the Law of the Sea devotes a section to states' obligations to "prevent, reduce and control pollution of the marine environment."[63] A number of other treaties, many of which use the word "control" in their titles, exist specifically to control the transboundary movement of hazardous wastes and contaminants. International

[58] Human Rights Watch and IHRC, "Killer Robots and the Concept of Meaningful Human Control."

[59] 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 (emphasis added) ("[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.").

[60] Military and Paramilitary Activities in and against Nicaragua (Nicaragua v. United States of America), International Court of Justice, Judgment, June 24, 1986.

[61] *Prosecutor v. Tadic*, International Criminal Tribunal for the Former Yugoslavia Appeals Chamber, Case No. IT-94-1-AR72, 35 ILM 32 (1996).

[62] Human Rights Watch and IHRC, "Killer Robots and the Concept of Meaningful Human Control," pp. 15-16.

[63] United Nations Convention on the Law of the Sea (UNCLOS), adopted December 10, 1982, entered into force November 16, 1994, 1833 UNTS 3, part XII, section 5 (entitled "International Rules and National Legislation to Prevent, Reduce and Control Pollution of the Marine Environment").

environmental law thus shows that a state's obligation to exercise control has found acceptance within the international community and offers a promising model for a comparable duty in the context of regulating the use of weapons.[64] Like state control over pollution, human control over weapons serves to prevent harm to unintended victims and across national borders. The notion of meaningful human control over weapons systems and the use of force thus fits neatly into existing international law frameworks.

Documents setting out AI principles provide additional support for using human control as a guiding principle for the development, deployment, and governance of AI-driven technology. In fact, the principle of human control of technology itself is present in 64 percent of the documents studied in "Principled Artificial Intelligence."[65] The UNI Global Union, for example, states that AI systems need to maintain "the legal status of tools, and legal persons [need to] retain control over, and responsibility for, these machines at all times."[66]

[64] Other treaties on pollution control include the Convention on the Control of Transboundary Movements of Hazardous Wastes and their Disposal, adopted March 22, 1989, 1673 UNTS 57, entered into force May 15, 1992; Convention to Ban the Importation into Forum Island Countries of Hazardous and Radioactive Wastes and to Control the Transboundary Movement and Management of Hazardous Wastes within the South Pacific Region, adopted 1995, entered into force 2001; Convention on the Ban of the Import into Africa and the Control of Transboundary Movements and Management of Hazardous Wastes within Africa, adopted January 30, 1991, entered into force April 22, 1998, 30 ILM 773.

[65] Fjeld et al., "Principled Artificial Intelligence," p. 53.

[66] UNI Global Union, "Top 10 Principles for Ethical Artificial Intelligence" (2017), http://www.thefutureworldofwork.org/media/35420/uni_ethical_ai.pdf (accessed September 3, 2020), p. 8.

V. Conclusion

Negotiating a new treaty on fully autonomous weapons is a challenging but feasible endeavor. While states will have to tailor the instrument to address the distinctive characteristics of this emerging technology, they will not be in uncharted territory. The elements of a new treaty provide a starting point for discussion and offer an effective model for addressing the concerns raised by fully autonomous weapons. Drafters can also draw on and adapt existing international law and AI principles when developing the structure and content of the treaty. Finally, the precedent discussed in this report can help generate political support by showing states that they have adopted similar norms in the past. States that wish to preserve meaningful human control over the use of force and prevent the next dangerous revolution in warfare should not be swayed by skeptics who say these goals are too difficult to accomplish. States have successfully governed unacceptable weapons in the past. They can and, given the high stakes, should do so again.

Acknowledgments

The lead writer of this report was Bonnie Docherty, senior researcher in the Arms Division at Human Rights Watch and associate director of armed conflict and civilian protection at the Harvard Law School International Human Rights Clinic (IHRC). Alev Erhan, Ayako Fujihara, Daniel Moubayed, and Shaiba Rather, IHRC students, contributed significantly to the research, analysis, and writing of this report.

This report was reviewed and edited by Arms Division executive director Steve Goose and advocacy director Mary Wareham. James Ross, legal and policy director, and Tom Porteous, deputy program director, provided legal and programmatic review. This report was prepared for publication by Jacqulyn Kantack, associate in the arms division, Fitzroy Hepkins, senior administrative manager, and José Martinez, administrative officer.

NEW WEAPONS, PROVEN PRECEDENT

Elements of and Models for a Treaty on Killer Robots

The rapid development of fully autonomous weapons, also known as "killer robots," presents one of the most urgent challenges facing the world today. These systems, which would select and engage targets without meaningful human control, raise a host of moral, legal, accountability, and security concerns. Human Rights Watch and the Harvard Law School International Human Rights Clinic are calling for a new treaty to prevent the delegation of life-and-death decisions to machines. A majority of countries party to the Convention on Conventional Weapons have adopted a similar position.

New Weapons, Proven Precedent identifies key elements of a new treaty to maintain meaningful human control over the use of force and prohibit weapons systems that operate without such control. The treaty should lay out positive obligations and prohibitions and elaborate on the components of meaningful human control.

hile the cutting-edge technology of fully autonomous weapons raises distinct issues, creating new law from scratch would unnecessarily slow negotiations. This report provides precedent for each of its proposed treaty elements and shows that constructing a legally binding instrument does not require an entirely fresh approach. Existing international law and principles of artificial intelligence can inform the structure and content of a future instrument. The existence of relevant models should make it legally, politically, and practically feasible to develop a new treaty in a timely way.

Following the lead of legal precedent, a new treaty on killer robots should ensure meaningful human control over the use of force and ban weapons operating without such control.

© 2020 Brian Stauffer for Human Rights Watch

IHRC | INTERNATIONAL HUMAN RIGHTS CLINIC
HUMAN RIGHTS PROGRAM AT HARVARD LAW SCHOOL

hrw.org

http://hrp.law.harvard.edu/clinic