UNODA Occasional Papers

No. 30, November 2017

Perspectives on Lethal Autonomous Weapon Systems

UNODA
United Nations Office for
Disarmament Affairs

United Nations

UNODA

United Nations Office for
Disarmament Affairs

UNODA Occasional Papers

No. 30, November 2017

Perspectives on Lethal Autonomous Weapon Systems

United Nations

The United Nations Office for Disarmament Affairs (UNODA) Occasional Papers are a series of ad hoc publications featuring, in edited form, papers or statements made at meetings, symposiums, seminars, workshops or lectures that deal with topical issues in the field of arms limitation, disarmament and international security. They are intended primarily for those concerned with these matters in Government, civil society and in the academic community.

The views expressed in this publication are those of the authors and do not necessarily reflect those of the United Nations or its Member States.

Material in UNODA Occasional Papers may be reprinted without permission, provided the credit line reads "Reprinted from UNODA Occasional Papers" and specifies the number of the Occasional Paper concerned. Notification to the following e-mail address would be highly appreciated: unoda-web@un.org.

Symbols of United Nations documents are composed of capital letters combined with figures. These documents are available in the official languages of the United Nations at http://ods.un.org. Specific disarmament-related documents can also be accessed through the disarmament reference collection at https://www.un.org/disarmament/publications/library/.

This publication is available from

www.un.org/disarmament

UNITED NATIONS PUBLICATION
Sales No. E.17.IX.6

ISBN 978-92-1-142324-2
eISBN 978-92-1-362894-2

Printed at the United Nations, New York

Contents

Foreword

The prospect of the deployment of fully autonomous weapon systems raises a number of troubling questions. As explored in this paper, these questions are multi-dimensional. For example, lethal autonomous weapon systems (LAWS)[1] could seriously test existing legal frameworks by posing novel challenges for attribution and accountability. They also pose ethical and moral quandaries: are we comfortable with outsourcing life and death decisions to machines and what does that say about the value we place on the sanctity of human life? On the security front, such weapons may lower barriers to the use of force and could be particularly attractive to unsophisticated and unscrupulous non-State actors.

These disturbing prospects are not only concerning, but urgently so. This is because there is today no technical barrier to the deployment of LAWS. In fact, autonomous systems are already deployed in limited environments, such as the open seas, generally far from civilian populations.

Although there are no technical barriers to deploying LAWS that could target humans or act in or near civilian areas, there are arguably normative barriers. Through the discussions that have already taken place informally under the auspices

[1] There is no internationally agreed formal definition for LAWS, and each contributor to this paper uses his/her own. Definitions will likely play a key role in international deliberations on this issue.

of the Convention on Certain Conventional Weapons (CCW), there appears to be an emergent consensus around the view that target selection and engagement decisions should not be entirely delegated to machines.

Since its creation more than 70 years ago, the United Nations has been the key forum for Governments to discuss norms with respect to weapons of all types. The discussions on LAWS within the CCW have shown that this remains the case. It is my hope that the next round of formal deliberations will provide an opportunity for States to further build understanding, exchange views and explore areas of normative convergence.

Autonomy in weapons is a cross-cutting issue requiring a cross-disciplinary approach. It is important that government deliberations are informed by relevant expertise. I hope that this publication serves of use to those participating in the important normative talks to come.

Izumi Nakamitsu

Under-Secretary-General
High Representative for Disarmament Affairs

Introduction

Amandeep S. Gill
Ambassador and Permanent Representative of India
to the Conference on Disarmament
Chair of the 2017 Group of Governmental Experts
on lethal autonomous weapon systems

After three years of informal discussions, the High Contracting Parties to the Convention on Certain Conventional Weapons (CCW) decided in December 2016 to establish an open-ended Group of Governmental Experts (GGE) related to emerging technologies in the area of lethal autonomous weapon systems.[1] The Group will hold its first session from 13 to 17 November 2017 in Geneva.

The decision demonstrates a maturing of international interest in the implications for warfare of a new suite of technologies including artificial intelligence and deep machine learning. Use of technology to wage war is nothing new. Every technological revolution from gunpowder and steam ships to nuclear fission and rocketry has given rise to new challenges for international security, arms control and disarmament. What is perhaps different this time is the prospect of losing human control, however imperfect and unwise it has been historically, over the waging of war to machines. The weapon, separate thus far in combat, could in fact fuse with the wielder.

[1] Decision 1, CCW/CONF.V/10.

Then again, autonomy and self-learning in machines, just like all dual-use technologies before, hold the prospect of ameliorating the human condition. The recent Global Summit on AI for Good, held from 7 to 9 June 2017 at the International Telecommunications Union, showcased a range of peaceful applications for artificial intelligence in areas such as medicine, education, mobility and agriculture. The overwhelming sense was that none of the United Nations Sustainable Development Goals would be left untouched by this wave of technology. Policy makers may have to scramble to stay abreast of a rapidly evolving field that exhibits all the characteristics of non-linear change and unpredictability arising out of the interaction of complex systems.

It is therefore timely that a formal and regular consideration of issues related to emerging technologies relevant to lethal autonomous weapon systems (LAWS) has been started under the auspices of the CCW, a unique framework convention that strives for a balance between military necessity and the humanitarian imperative. The flexibility of the CCW and its deep roots in international humanitarian law make it ideally suited to be the framework under which the legal, military and humanitarian issues arising out of the potential development, proliferation and use of LAWS can be tackled. The informal discussions of the past three years, chaired by representatives from France (April 2014)[2] and Germany (April 2015 and April 2016),[3] as well as several in-depth studies carried out by the United Nations Institute for Disarmament Research, the International Committee of the Red Cross, many non-governmental organizations, and academic and industry groups, have laid a sound basis for the start of formal discussions in the GGE in the context of the objectives and purposes of the CCW. Civil society has contributed laudably to raising awareness about the issue and stimulating debate on policy options.

[2] See CCW/MSP/2014/3.

[3] See CCW/MSP/2015/3 and CCW/MSP/2016/3 (advance version).

Based on consultations with High Contracting Parties and as Chair of the GGE, I am of the view that the GGE can make a good start this year by conducting a thorough review of the current state of developments in the technology domain writ large, as well as the status of the specific application of autonomous technologies to the military domain. As Moore's law underlines, 18 months is a long time in technology development and the CCW needs to catch up to enable States to build the foundation of a fact-based discussion. Equally, the GGE could continue earlier examination of legal and ethical issues of relevance to LAWS, including emerging developments in law to regulate the use of civilian autonomous systems such as self-driving vehicles. The interdisciplinary approach adopted so far in the CCW to discuss emerging technologies related to LAWS has proved to be of value. Diplomats, soldiers, technologists, lawyers and ethicists need to continue to pool their insights so that policy makers can better visualize pathways for policy learning.

In this regard, the present publication of the United Nations Office for Disarmament Affairs is timely. Experts in the fields of technology, international law, security policy, human rights and ethics have come together to present their respective perspectives on LAWS. Of course the publication cannot claim to be exhaustive or fully representative of the views of United Nations Member States. It is a toolkit to help policy makers better understand the field. Such understanding would be of value to the GGE work.

The GGE can be the platform the international community needs to begin a much-needed policy debate on the implications of possible lethal autonomy in weapons. High Contracting Parties will own this debate as the primary stakeholders, but other stakeholders, including the private sector that is driving the development of technology, also need to contribute to the success of this effort. Our aim for the immediate future should be twofold: to develop an accurate understanding of the ground truth on technology and its military applications, and to develop common understanding of concepts and other linguistic framing devices.[4] That would enable an objective exploration of policy options to commence in the next stage.

[4] The terms "meaningful human control" or "appropriate human judgement" are two examples.

A legal perspective: Autonomous weapon systems under international humanitarian law

Neil Davison
Scientific and Policy Adviser
Arms Unit, Legal Division
International Committee of the Red Cross

I. Introduction

This chapter reviews the key issues raised by autonomous weapon systems under international humanitarian law (IHL), drawing on previously published documents of the International Committee of the Red Cross (ICRC).[1] For the purpose of this analysis, an **autonomous weapon system is defined as follows**:

> Any weapon system with autonomy in its critical functions—that is, a weapon system that can select (search for, detect, identify, track or select) and attack (use force against, neutralize, damage or destroy) targets without human intervention.

[1] ICRC, *Views of the ICRC on autonomous weapon systems*, paper submitted to the Convention on Certain Conventional Weapons Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS), 11 April 2016, https://www.icrc.org/en/document/views-icrc-autonomous-weapon-system;

After initial launch or activation by a human operator, it is the weapon system itself—using its sensors, computer programming (software) and weaponry—that takes on the targeting functions that would otherwise be controlled by humans. This working definition encompasses any weapon system that can independently select and attack targets, including some existing weapons[2] and potential future systems.

The definition provides a useful basis for a legal analysis by delineating the broad scope of the discussion about autonomous weapon systems without the need to immediately identify the systems that raise legal concerns. In that sense, the definition is not intended to prejudge the level of autonomy in weapon systems that may, or may not, be considered lawful.

Rather, the ICRC has proposed that States determine where these limits must be placed by assessing the **type and degree of human control required in the use of weapon systems to carry out attacks**—at a minimum, for compliance with IHL and, in addition, to satisfy ethical considerations.[3]

ICRC, *Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons*, ICRC, Geneva, September 2016, https://www.icrc.org/en/publication/4283-autonomous-weapons-systems; ICRC*, International Humanitarian Law and the challenges of contemporary armed conflicts*. 32nd International Conference of the Red Cross and Red Crescent, October 2015, 32IC/15/11, p. 44-47, https://www.icrc.org/en/download/file/15061/32ic-report-on-ihl-and-challenges-of-armed-conflicts.pdf.

[2] Examples are missile and rocket defence weapons; vehicle "active protection" weapons; certain missiles, loitering munitions and torpedoes; and some "sentry" weapons. See ICRC, *Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons*, footnote 1, pp. 10-14.

[3] ICRC, *Statement to the Convention on Certain Conventional Weapons (CCW) Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS)*, Geneva, 11 April 2017, https://www.icrc.org/en/document/statement-icrc-lethal-autonomous-weapons-systems.

II. Compliance with international humanitarian law

Autonomous weapon systems, as defined, are not specifically regulated by IHL treaties. However, it is undisputed that any autonomous weapon system must be capable of being used, and must be used, in accordance with IHL. The responsibility for ensuring this rests, first and foremost, with each State that is developing, deploying and using weapons (see also section IV).

While the primary subjects of IHL are the parties to an armed conflict, the rules on the conduct of hostilities—notably the rules of **distinction, proportionality and precautions in attack**—are addressed to those who plan, decide upon and carry out an attack.

The core legal obligations for a commander or operator in the use of weapon systems include the following: to ensure **distinction** between military objectives and civilian objects, combatants and civilians, and active combatants and those hors de combat; to determine whether the attack may be expected to cause incidental civilian casualties and damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated, as required by the rule of **proportionality**; and to cancel or suspend an attack if it becomes apparent that the target is not a military objective or is subject to special protection, or that the attack may be expected to violate the rule of proportionality, as required by the rules on **precautions in attack**.

These **IHL rules create obligations for human combatants in the use of weapons to carry out attacks**, and it is combatants who are both responsible for respecting these rules, and who will be held accountable for any violations. As for all obligations under international law, these legal obligations, and accountability for them, cannot be transferred to a machine, computer program or weapon system.

Those who plan, decide upon and carry out an attack using an autonomous weapon system must, therefore, ensure that the weapon system and the way it is used preserve their ability to make these necessary legal judgements, and thereby ensure compliance with IHL. It follows that an autonomous weapon system will raise concerns under IHL if—through its design, performance and/or method of use—it impedes commanders or operators in making these legal judgements. For example, if a mobile autonomous weapon system searches for targets over a wide area and for a long duration, without human supervision and communication, the commander who authorized the launch of the weapon and the operator who activated it will not know exactly where and when an attack will take place. This raises questions of whether they will be able to ensure distinction, judge proportionality or take precautions should the circumstances change.

III. The "principles of humanity" and the "dictates of the public conscience"

The Martens Clause provides a link between ethical considerations and IHL, which makes it particularly relevant to the assessment of autonomous weapon systems. It provides that, in cases not covered by existing treaties, civilians and combatants remain protected by customary IHL, the **principles of humanity, and the dictates of the public conscience**.[4] As such, the principles of humanity are a universal reference point, preventing the assumption that anything not explicitly prohibited is permitted, and thereby addressing new situations and new means and methods of warfare.

With increasing autonomy in weapon systems, a point may be reached where humans are so far removed in time

[4] The "principles of humanity and the dictates of public conscience" are mentioned notably in article 1(2) of Additional Protocol I and in the preamble of Additional Protocol II to the Geneva Conventions, referred to as the Martens Clause.

and space from the acts of selecting and attacking targets that human decision-making is effectively substituted with computer-controlled processes, and life-and-death decisions in armed conflict ceded to machines. This raises profound ethical questions about the role and responsibility of humans in the use of force and the taking of human life, which go beyond questions of IHL compliance in the conduct of hostilities. With respect to the public conscience, there is a sense of deep discomfort with the idea of any weapon system that places the use of force beyond human control.[5]

IV. Legal review of new weapons

The obligation to carry out legal reviews of new weapons under article 36 of Additional Protocol I to the Geneva Conventions is important for ensuring that a State's armed forces are capable of conducting hostilities in accordance with its international obligations.[6]

As with all weapons, assessing the lawfulness of an autonomous weapon system will depend on its specific characteristics and whether, given those characteristics, it can be employed in conformity with the rules of IHL in all circumstances in which it is intended and expected to be used. In

[5] See, for example, ICRC (2015) *Statement to the Convention on Certain Conventional Weapons (CCW) Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS), 13-17 April 2015, Geneva*, https://www.icrc.org/en/document/lethal-autonomous-weapons-systems-LAWS; Future of Life Institute, *Autonomous Weapons: an Open Letter from AI & Robotics Researchers*. International Joint Conference on Artificial Intelligence, 28 July 2015, https://futureoflife.org/open-letter-autonomous-weapons; and Future of Life Institute (2017), *An Open Letter to the United Nations Convention on Certain Conventional Weapons*, 21 August 2017, https://futureoflife.org/autonomous-weapons-open-letter-2017.

[6] ICRC (2006), *A Guide to the Legal Review of New Weapons, Means and Methods of Warfare: Measures to Implement Article 36 of Additional Protocol I of 1977*, Geneva, January 2006, www.icrc.org/eng/assets/files/other/icrc_002_0902.pdf.

particular, the legal review must consider treaty and customary prohibitions and restrictions on specific weapons, as well as the general IHL rules applicable to all weapons, means and methods of warfare. These include the rules aimed at protecting civilians from the indiscriminate effects of weapons and combatants from superfluous injury and unnecessary suffering.

The ability to carry out such a review entails fully understanding the weapon's capabilities and foreseeing its effects, notably through verification and testing. Since the commander or operator must make an assessment of the lawfulness of an attack using an autonomous weapon system at an earlier stage than if the selection and attack of targets were under direct human control, the legal review must demand a very high level of confidence that, once activated, the autonomous weapon system would predictably and reliably operate as intended. This raises unique challenges in ensuring that predictability and reliability are tested and verified for all foreseeable scenarios of use.

Predictability is the ability to "say or estimate that (a specified thing) will happen in the future or will be a consequence of something".[7] Applied to an autonomous weapon system, predictability is knowledge of how it will function in any given circumstances of use, and the effects that will result.[8] **Reliability** is "the quality of being trustworthy or performing consistently well".[9] In this context, reliability is knowledge of how consistently the machine will function as intended—e.g., without failures or unintended effects.[10]

[7] Oxford English Dictionary, third edition, Oxford University Press, 2010, https://en.oxforddictionaries.com/definition/predictability.

[8] ICRC Expert Meeting, *Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons, Versoix, Switzerland, 15-16 March 2016*, p. 9.

[9] Oxford English Dictionary.

[10] ICRC, *Autonomous Weapon Systems*, p. 13.

V. Human control under international humanitarian law

The question remains, however, what limits are needed on autonomy in weapon systems to ensure compliance with IHL?

There is general agreement among Convention on Certain Conventional Weapons (CCW) States Parties that "meaningful" or "effective" human control, or "appropriate levels of human judgement" must be retained over weapon systems and the use of force. The Chair's summary of the April 2016 CCW informal meeting of experts states the following:

> *Views on appropriate human involvement with regard to lethal force and the issue of delegation of its use are of critical importance* to the further consideration of LAWS [lethal autonomous weapon systems].[11]

For its part, the ICRC has called for human control to be maintained over weapon systems and the use of force to satisfy legal and ethical requirements.

A certain level of human control or involvement is inherent in the implementation of the IHL rules on the conduct of hostilities. While IHL creates obligations for States and parties to armed conflicts, IHL rules are ultimately implemented by human subjects who are responsible for complying with these rules in carrying out attacks, and must be held accountable for violations. It follows that some degree of human control over the functioning of an autonomous weapon system, translating the intention of the user into the operation of the weapon system, will always be necessary to ensure compliance with IHL, and this may indeed limit the lawful level of autonomy.

Core components of human control include the following: **predictability** and **reliability** (defined in section IV) of the

[11] United Nations*, Recommendations to the 2016 Review Conference submitted by the Chairperson of the Informal Meeting of Experts*, para. 2 (b); italics added.

weapon system in its intended or expected circumstances of use; **human intervention** in the functioning of the weapon system during its development, activation and operation; **knowledge** and **information** about both the functioning of the weapon system and the environment of its use; and **accountability** for the ultimate operation of the weapon system.

For autonomous weapon systems, as defined, the control exercised by humans can take various forms and degrees at different stages of development, deployment and use, including the following: (a) the development and testing of the weapon system ("development stage"); (b) the decision by the commander or operator to activate the weapon system ("activation stage"); and (c) the operation of the autonomous weapon system during which it independently selects and attacks targets ("operation stage").

A. Development stage

Human control can be exercised at the development stage, including through technical design and programming of the weapon system. Decisions taken during the development stage must ensure that the weapon system can be used in accordance with IHL and other applicable international law in the intended or expected circumstances of use. At this stage, the **predictability** and **reliability** of the weapon system must be verified through testing in realistic environments. Operational parameters on the use of the weapon must be integrated into the military instructions for its use, for instance to limit its use to a specific situation, to constrain its movement in time and space, or to enable human supervision (see activation and operation stages). For example, an existing vehicle "active protection" weapon (which attacks incoming rockets or mortars) will need to be tested against the intended circumstances of use, and operational limits must be set so that the weapon is only activated in situations where its effects will be predictable. Also, the operational requirement and technical mechanism for human supervision, as well as the ability to deactivate the weapon, will need to be established.

B. Activation stage

The second stage at which human control can be exerted is at the point of activation, which involves the decision of the commander or operator to use a particular weapon system for a particular purpose either in a specific attack, or to respond to a general threat over a specific time period (e.g., defending against incoming rockets). This decision on the part of the commander or operator must be based on sufficient knowledge and understanding of the weapon's functioning in the given circumstances to ensure that it will operate as intended and in accordance with IHL. This knowledge must include adequate situational awareness of the operational environment, especially in relation to the potential risks to civilians and civilian objects.

Whether the weapon system will operate within the constraints of IHL once activated will depend on the technical performance of the specific weapon in the specific circumstances of use, especially its predictability and reliability (as determined and tested at the development stage). However, it will also depend on various operational parameters, most of which will be set at the development stage, and some that will be set or adjusted at the activation stage. These include the following:

- The **task** the weapon system is assigned
- The **type of target** the weapon system may attack
- The **type of force** and munitions it employs (and associated effects)
- The **environment** in which the weapon system is to operate
- The **mobility** of the weapon system in space
- The **time frame** of its operation

- The **level of human supervision and ability to intervene after activation**.

There are lessons to be drawn from existing autonomous weapon systems, such as missile and rocket defence systems, where human control is largely exerted through a combination of technical performance and operational constraints, such as limits on targets, limits in geographical space and time frame of operation, physical controls over the environment, and human supervision and ability to deactivate.[12]

C. Operation stage

The risk that IHL might be violated can be reduced by manipulating these operational parameters up to the point of activation. However, in order to ensure compliance with IHL, there may need to be additional human control during the operation stage, when the weapon autonomously selects and attacks targets. The last operational parameter listed above, the **level of human supervision and ability to intervene after activation**, provides a means by which further control can be exerted over an attack.

Where the technical performance of the weapon and operational parameters set during the development and activation stages are insufficient to ensure compliance with IHL in carrying out an attack, it will be necessary to retain the ability for human control and decision-making during the operation stage. An example would be through supervision of the weapon system and the target area and two-way communication links that permit adjustment of the engagement criteria and the ability to cancel an attack. For example, some existing counter-rocket, artillery and mortar weapons retain the ability, even with

[12] ICRC, *Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons*, ICRC, Geneva, September 2016, pp. 10-14.

incoming projectiles, for a human operator to visually verify the projectile on screen and decide to cancel the attack if necessary.

In sum, the type and degree of human control over an autonomous weapon system that is required to ensure compliance with IHL can manifest itself in terms of the following: (a) verified technical performance of the weapon system for its intended use, as determined at the development stage; (b) manipulation of operational parameters at the development and activation stages; and (c) human supervision and potential for intervention and deactivation during the operation stage. This suggests that compliance with IHL requires limits to lawful levels of autonomy in weapon systems.

VI. The importance of predictability for IHL compliance

Predictability in the functioning of a weapon in the intended circumstances of use is central to compliance with IHL (see also definitions in section IV). The commander or operator needs a high level of confidence that, upon activation, an autonomous weapon system will operate predictably, which in turn demands a high degree of predictability in its technical performance, the environment and the interaction of the two. The greater the uncertainty and unpredictability, the greater the risk that IHL might be violated.

Predicting the outcome of using autonomous weapon systems will become increasingly difficult if such systems become very complex in their functioning (e.g., hardware sensors and software algorithms) and/or are given significant freedom of operation in tasks, and over time and space. For example, in the legal assessment of an autonomous weapon system that carries out a single task against a specific type of target in a simple environment, that is stationary and limited in the duration of its operation, and that is supervised by a human operator with the potential to intervene at all times (e.g., existing missile and rocket defence systems), it may be concluded that there is an acceptable level of predictability, allowing for a human operator to ensure IHL compliance. However, the conclusion may be very different for an autonomous weapon system that carries out multiple tasks or adapts its functioning against different types of targets in a complex environment, that searches for targets over a wide area and/or for a long duration, and that is unsupervised.

Increased flexibility in tasks or mobility over time and space would increase uncertainty about when and where specific attacks would take place and unpredictability in the environment encountered. Increased complexity, such as systems controlled by software incorporating artificial intelligence algorithms to set its own goals or to "learn" and adapt its functioning, would arguably be inherently unpredictable, especially when combined with an often unpredictable and hostile environment.

VII. Accountability for violations of IHL

There have been questions raised about whether the use of autonomous weapon systems may lead to a legal "accountability gap" in case of violations of IHL. While there will always be a human involved in the decision to deploy and activate a weapon to whom accountability could be attributed, the nature of autonomy in weapon systems means that the lines of responsibility may not always be clear.

Under the law of **State responsibility**, a State could be held liable for violations of IHL resulting from the use of an autonomous weapon system. Indeed, under general international law governing the responsibility of States, they would be held responsible for internationally wrongful acts, such as violations of IHL committed by their armed forces using an autonomous weapon system. A State would also be responsible if it were to use an autonomous weapon system that has not been adequately tested or reviewed prior to deployment.

Under IHL and **international criminal law**, the limits of human control over an autonomous weapon system could make it difficult to find individuals involved in the programming (development stage) and deployment (activation stage) of the weapon liable for serious violations of IHL in some circumstances. Humans that have programmed or activated the weapon systems may not have the knowledge or intent required to be found liable, owing to the fact that the machine, once activated, can select and attack targets independently. Programmers might not have knowledge of the concrete situations in which, at a later stage, the weapon system might be deployed and in which IHL violations could occur and, at the point of activation, commanders may not know the exact time and location where an attack would take place.

On the other hand, a programmer who intentionally programmes an autonomous weapon to operate in violation of IHL or a commander who activates a weapon that is incapable of functioning lawfully in that environment would certainly be criminally liable for a resulting violation. Likewise, a commander who knowingly decides to activate an autonomous weapon system whose performance and effects they cannot reasonably predict in a particular situation may be held criminally responsible for any serious violations of IHL that result, to the extent that their decision to deploy the weapon is deemed reckless under the circumstances.

Furthermore, under the laws of **product liability**, manufacturers and programmers might also be held accountable for errors in programming or for the malfunction of an autonomous weapon system.

VIII. Conclusion

IHL rules on the conduct of hostilities—notably the rules of distinction, proportionality and precautions in attack—are addressed to those who plan, decide upon and carry out an attack in armed conflict. These rules create obligations for human combatants in the use of all weapons to ensure compliance with IHL. The lawful use of autonomous weapon systems, as broadly defined, will therefore require that combatants retain a level of human control over their functioning in carrying out an attack.

Examining the way in which—and at which stages of their development, activation and operation—human control is currently exerted over autonomous weapon systems, through technical characteristics and operational parameters, can provide insights into the type and degree of human control necessary for IHL compliance, including standards of predictability, operational constraints, and human supervision and ability to intervene.

Overall, this analysis indicates that, under IHL, there will be limits to lawful levels of autonomy in weapon systems. States should now begin to determine where internationally agreed limits must be placed by assessing the type and degree of human control required, in the use of weapons to carry out attacks, to ensure compliance with IHL. This assessment should also consider the level of human control required to satisfy ethical considerations, which may call for additional limitations.

A security perspective: Security concerns and possible arms control approaches

Paul Scharre
Senior Fellow
Center for a New American Security

Autonomous weapons and stability

In addition to legal and ethical concerns, a number of scholars have raised questions about how autonomous weapons might affect stability.[1] During the cold war, the concept of

Note: Paul Scharre is also the author of the forthcoming book *Army of None: Autonomous Weapons and the Future of War*, to be published in April 2018.

[1] John Borrie, "Safety aspects of 'meaningful human control': Catastrophic accidents in complex systems" (statement delivered at the UNIDIR conference entitled "Weapons, Technology, and Human Control", New York, New York, 16 October 2014); Michael Carl Haas, "Autonomous Weapon Systems: The Military's Smartest Toys?", *The National Interest*, 20 November 2014, http://nationalinterest.org/feature/autonomous-weapon-systems-the-militarys-smartest-toys-11708; Alexander Velez-Green, "The Foreign Policy Essay: The South Korean Sentry—A 'Killer Robot' to Prevent War", *Lawfare*, 1 March 2015, http://www.lawfareblog.com/2015/03/the-foreign-policy-essay-the-south-korean-sentry-a-killer-robot-to-prevent-war/; John Borrie, "Unintentional risks", (statement delivered at

"stability" emerged as an important factor in evaluating new weapon technologies.[2] Stability was seen as a good thing, because it meant maintaining the status quo: peace. Instability was seen as dangerous, because it could lead to war.[3]

There are a number of variants of this concept. First-strike instability refers to the idea that some weapons or deployment postures might give an advantage to whichever side struck first, thus incentivizing nations to launch a preemptive attack in a crisis.[4] A stable situation is one in which neither side can gain an advantage by striking first. Over time, major nuclear powers adapted their forces so that they could survive a nuclear first

the UNIDIR conference entitled "Understanding Different Types of Risks", Geneva, Switzerland, 11 April 2016); Alexander Velez-Green, "When 'Killer Robots' Declare War", *Defense One*, 12 April 2015, http://www.defenseone.com/ideas/2015/04/when-killer-robots-declare-war/109882/; Paul Scharre and Michael C. Horowitz, "Keeping Killer Robots on a Tight Leash", *Defense One*, 14 April 2015, http://www.defenseone.com/ideas/2015/04/keeping-killer-robots-tight-leash/110164/?oref=d-river; Jean-Marc Rickli, "Some Considerations of the Impact of LAWS on International Security: Strategic Stability, Non-State Actors and Future Prospects", paper presented to the Meeting of Experts on Lethal Autonomous Weapons Systems of the Convention on Certain Conventional Weapons, 6 April 2015, Geneva, Switzerland, http://www.unog.ch/80256EDD006B8954/(httpAssets)/B6E6B974512402BEC1257E2E0036AAF1/$file/2015_LAWS_MX_Rickli_Corr.pdf; United Nations Institute for Disarmament Research, "Safety, Unintentional Risk and Accidents in the Weaponization of Increasingly Autonomous Technologies", 2016, http://www.unidir.org/files/publications/pdfs/safety-unintentional-risk-and-accidents-en-668.pdf; and Jürgen Altmann and Frank Sauer, "Autonomous Weapon Systems and Strategic Stability". *Survival*, vol. 59, No. 5 (2017), pp. 117-142, http://dx.doi.org/10.1080/00396338.2017.1375263.

[2] When in reference to nuclear weapons, the term "strategic stability" is often used.

[3] For an overview of the concept of stability, see Elbridge A. Colby, "Defining Strategic Stability: Reconciling Stability and Deterrence", *Strategic Stability: Contending Interpretations* (Strategic Studies Institute and U.S. Army War College Press, 2013), Elbridge A. Colby and Michael S. Gerson, eds., pp. 48-49. See also Thomas C. Schelling and Morton H. Halperin, *Strategy and Arms Control* (New York, The Twentieth Century Fund, 1961).

[4] This is sometimes called "first-mover advantage".

strike and still respond, for example by placing nuclear weapons aboard submarines or mobile missile launchers that would be hard to find and destroy. This was seen as stabilizing, since it reduced incentives for a first strike.

A related concept is crisis stability, which is concerned with avoiding crisis escalation through miscalculation, accidents or a lack of effective control over military forces. False alarms, misunderstandings and the fog of war often contribute to instability in crises. Safety protocols that reduce the risk of accidents and increase national leaders' control over their own forces can increase stability. Similarly, measures that increase communication and transparency between nations, such as crisis hotlines, can increase stability.

Controlling escalation and war termination are also important components of stability. Just as national leaders should have complete control over the decision to go to war, they should also have control over escalation in conflicts and ending wars once they are under way.[5] For example, destroying all communication links between an enemy's leadership and their deployed forces might have tactical advantages, but could make war termination harder if enemy leaders cannot communicate to their forces a decision to surrender.

There are many ways in which robotics and autonomy could affect stability.[6] It is important, however, to distinguish

[5] Thomas C. Schelling, *Arms and Influence* (New Haven and London,Yale University Press, 1966), pp. 203-208; and Fred Ikle, *Every War Must End* (New York, Columbia Classics, 2005).

[6] For example, some have raised concerns that using robots would reduce the risk of casualties and therefore lower the threshold for countries to go to war (Peter M. Asaro, "How Just Could a Robot War Be?", http://peterasaro.org/writing/Asaro%20Just%20Robot%20War.pdf).
Others have raised concerns that robotic swarms could alter the offence-defence balance between nations, thus incentivizing conflict by making territorial aggression easier (Rickli, "Some Considerations of the Impact of LAWS on International Security: Strategic Stability, Non-State Actors and Future Prospects"). These concerns, which may or may not be

between how robotics and autonomy *in general* might affect war and the implications of autonomous weapons *specifically*. The key issues under discussion at the Convention on Certain Conventional Weapons meetings are not *all* military applications of autonomy, but rather autonomy in lethal force decisions.

Human control over war

Because the essence of autonomous weapons is that humans have delegated lethal force decision-making to the machine, one question is whether autonomous weapons would increase or decrease human control in war. If autonomous weapons led to greater human control over war initiation, escalation and termination, then that would be desirable. Greater control would be stabilizing; it would make accidents and miscalculation less likely. If autonomous weapons led to less human control, then that could increase the risk of unintended escalation and would be undesirable.[7]

It might seem counter-intuitive that autonomy could increase human control, since the essence of autonomy is delegating a task to a machine. Sometimes, however, allowing machines to perform a task autonomously can increase human control over the outcome. For example, allowing a thermostat to turn the heat on at night makes it more likely that a home will be at the desired temperature in the morning. Automobile collision avoidance systems take immediate control from the driver, but do so in order to achieve the driver's desired outcome—avoiding hitting another car.

One way that autonomous weapons could potentially increase stability would be if they increased national leaders' control over how their forces behave in crises. People are idiosyncratic and can deviate from orders. Autonomous systems will do precisely what they are programmed to do. In theory,

justified in certain cases, are beyond the scope of this paper, which is about the role of autonomy in lethal force decisions.

[7] In either case, legal, ethical and other considerations would still apply.

this might make autonomous systems more predictable than humans in a crisis. An autonomous system could be the perfect soldier, never violating its orders.

The value of human judgment in crises

Unfortunately, this inflexibility has a downside. Autonomous systems can be “brittle”. They may perform some tasks very well—often better than humans—but if the context for their action changes, they often lack the flexibility to adapt.[8] This brittleness could be a major problem in controlling escalation. There are many examples of humans using their judgment to avoid escalation in crises. In 1983, Soviet Lieutenant Colonel Stanislav Petrov ignored information from early warning satellites that the United States had launched a surprise attack because the information did not fit the broader context. He later said about the missile warning, “I had a funny feeling in my gut”.[9] During the 1962 Cuban missile crisis, Soviet Navy Captain Vasili Arkhipov refused to authorize the launch of a nuclear torpedo against United States naval forces that were harassing a submarine under his command with signaling depth charges, even though he was authorized to do so and the submarine commander had ordered it.[10]

Autonomous weapons would strip away the potential for human judgment to consider the specific context for an action. Militaries have a concept of “commander’s intent”.[11] Subordinates must understand not only their orders, but also the

[8] John Launchbury, “A DARPA Perspective on Artificial Intelligence” (video), 15 February 2017, https://www.youtube.com/watch?v=-O01G3tSYpU.

[9] David Hoffman, “‘I Had a Funny Feeling in My Gut’”, *Washington Post*, 10 February 1999, http://www.washingtonpost.com/wp-srv/inatl/longterm/coldwar/shatter021099b.htm.

[10] “The Submarines of October”, accessed 17 June 2017, http://nsarchive.gwu.edu/NSAEBB/NSAEBB75/; and “The Cuban Missile Crisis, 1962: Press Release, 11 October 2002, 5:00 PM”, accessed 17 June 2017, http://nsarchive.gwu.edu/nsa/cuba_mis_cri/press3.htm.

[11] Headquarters, Department of the Army, Field Manual 100-5 (June 1993), 6-6.

outcome their commander intends to achieve.[12] More generally, humans have the ability to imagine what their leaders would want, given the situation they are in.[13] This allows humans to deviate from their instructions if necessary to comply with their leadership's intent. Machines—at least given the current state of artificial intelligence—cannot understand human intent.[14] The brittleness and inflexibility of autonomous systems could take away an important safety valve in crises—human judgment—thereby increasing the risk of escalation.[15]

Controlling escalation and war termination

Once a war is under way, autonomous weapons could also be detrimental to stability if they decreased human control over how the war is conducted. Accidents have, in the past, led to tit-for-tat exchanges that resulted in conflicts escalating to levels of destruction that both sides found undesirable. For example, in the Second World War, Germany and the United Kingdom initially refrained from attacking each others' cities. This changed after several German bombers strayed off course from their military targets in the dark and bombed London by mistake.[16] Britain retaliated by bombing Berlin and, in response, Germany launched the London Blitz.[17] Both humans and

[12] Lawrence G. Shattuck, "Communicating Intent and Imparting Presence", *Military Review* (March-April 2000), http://www.au.af.mil/au/awc/awcgate/milreview/shattuck.pdf, 66.

[13] Charles C. Krulak, "The Strategic Corporal: Leadership in the Three Block War", *Marines Magazine* (January 1999).

[14] Launchbury, "A DARPA Perspective on Artificial Intelligence".

[15] Paul Scharre, "Autonomous Weapons and Operational Risk", Center for a New American Security, February 2016, 53, http://www.cnas.org/sites/default/files/publications-pdf/CNAS_Autonomous-weapons-operational-risk.pdf.

[16] This attack occurred on 24 August 1940.

[17] Following the attack on Berlin, Hitler declared in a public speech, "If they declare that they will attack our cities on a large scale—we will eradicate their cities." ("Hitlers Bombenterror: 'Wir Werden Sie Ausradieren'", *Spiegel Online*, accessed 1 April 2003, http://www.spiegel.de/spiegelspecial/a-290080.html).

machines can cause accidents, but autonomy allows operation at greater scale, which can increase the consequences of accidents. An autonomous weapon that malfunctioned or was hacked might continue attacking the wrong targets until it ran out of ammunition. Moreover, since the same software would be replicated in other autonomous weapons of the same type, many systems could fail at the same time, leading to large-scale accidents.[18] Even if humans eventually regained control over such systems, it may be difficult to decrease tensions if the autonomous weapons had caused significant destruction.

War termination could also be a concern if autonomous weapons were operating for extended periods of time without communication links to human controllers, such as undersea where communications is challenging. If humans decided to end the war, they might be not be able to recall autonomous weapons for some period of time.[19]

[18] For more on the risks and consequences of accidents with autonomous weapons, see John Borrie, "Safety aspects of 'meaningful human control': Catastrophic accidents in complex systems"; John Borrie, "Unintentional risks"; Scharre, "Autonomous Weapons and Operational Risk"; and United Nations Institute for Disarmament Research, "Safety, Unintentional Risk and Accidents in the Weaponization of Increasingly Autonomous Technologies".

[19] Michael Carl Haas has noted, "Recallability and loss of control, clearly, are major concerns. While strike systems along the lines of the X-47B UCAS could initially be employed under close human supervision, it is difficult to see how they could realize their full potential in those scenarios where they offer by far the greatest value added: intelligence, surveillance and reconnaissance (ISR) and strike missions deep inside well-defended territory, where communications will likely be degraded and the electronic emissions produced by keeping a human constantly in the loop could be a dead giveaway. … During these stints inside the defended zone, [autonomous weapons] might not be fully recallable or reprogrammable, even if the political situation changes, which presents a risk of undesirable escalation and could undermine political initiatives" (Haas, "Autonomous Weapon Systems: The Military's Smartest Toys?"). For this reason, the official United States Department of Defense policy on autonomy in weapon systems requires that autonomous and semi-autonomous weapons be designed to "complete engagements in

The pace of battle

Even if there were no accidents with autonomous weapons, they could undermine stability if they accelerated the pace of battle beyond human reaction times. Autonomy is already used by over 30 nations to defend against rocket and missile attacks that could overwhelm human operators. When used purely in a defensive context, these applications of autonomy would likely increase stability, since they make it harder for an adversary to gain an advantage by striking first. If both sides were to use autonomous weapons offensively, however, the outcome could be a military "singularity", where the speed of action on the battlefield would eclipse the speed of human decision-making. This could undermine stability if national leaders lose the ability to effectively control a conflict. Strategist Thomas Schelling observed in *Arms and Influence*:

> The premium on haste—the advantage, in case of war, in being the one to launch it or in being a quick second in retaliation if the other side gets off the first blow—is undoubtedly the greatest piece of mischief that can be introduced into military forces, and the greatest source of danger that peace will explode into all-out war.[20]

Leaders need time to step back from the pressures of a conflict to use their judgment. This is true both during crises, so that leaders can avoid war, and once wars are under way, so that leaders can find a way to terminate hostilities.[21] Even if humans were still technically in control, autonomous weapons could undermine stability if they accelerated the speed of

a timeframe consistent with commander and operator intentions and, if unable to do so, terminate engagements or seek additional human operator input before continuing the engagement" (United States Department of Defense, "Directive 3000.09: Autonomy in Weapon Systems", 8 May 2017, http://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodd/300009p.pdf, 2).

[20] Schelling, *Arms and Influence,* 227.

[21] Ikle, *Every War Must End.*

war such that humans no longer felt that they had the time to fully consider their actions. A world in which humans had less control over war, whether because of the pace of battle or a lack of communication with military forces, would be undesirable. Less human control over the initiation of conflict, escalation and war termination could make wars more likely longer and more destructive.[22]

Possible arms control approaches

Autonomous weapons are not the first new military technology that humanity has grappled with. Prohibitions on weapons date back to antiquity.[23] There are many examples of arms control successes and failures throughout history.[24] The examples below point to several common themes.

[22] There is a concept known as the "stability-instability paradox", which suggests that too much stability can be a bad thing and that some degree of instability can be a good thing if it induces leaders to be more cautious. In general, though, instability is seen as undesirable (Michael Krepon, "The Stability-Instability Paradox, Misperception, and Escalation Control in South Asia", The Stimson Center, 2003, https://www.stimson.org/sites/default/files/file-attachments/stability-instability-paradox-south-asia.pdfl; and B. H. Liddell Hart, *Deterrent or Defence* (London, Stevens and Sons, 1960), 23).

[23] The ancient Hindu Laws of Manu, Dharmaśāstras and Mahābhārata contain prohibitions against concealed weapons and weapons tipped with poison or fire. There were also attempts to ban the crossbow in Europe in the twelfth century and a short-lived attempt to ban firearms in England in the early 1500s. These unsuccessful attempts are contrasted with Japan's successful relinquishment of firearms from the early 1600s until the mid-1800s, an unusual historical case.

[24] Notable arms control failures include attempts to ban the crossbow in Europe in the twelfth century and early twentieth century attempts to ban aerial attacks on cities, regulate submarine warfare and the use of poison gas in the First World War. Examples of arms control successes include European restraint from battlefield use of poison gas in the Second World War and bans on chemical and biological weapons, using the environment as a weapon, deploying nuclear weapons on the seabed or in space, placing weapons on the moon, weapons that cause fragments not detectable by X-ray, blinding lasers, landmines and cluster munitions. The latter part of the cold war also saw a number of successful bilateral arms control treaties between the United States

When countries refrain from building weapons, they do so for strategic reasons, not merely because weapons are prohibited. International law can be a useful coordinating mechanism among nations. Treaties help nations communicate which weapons they believe are unacceptable. But nations have often violated treaties in war.[25] At the same time, nations have sometimes refrained from deploying or using certain weapons even without formal treaties.[26] The most significant factor in inducing restraint is a fear of reciprocity, that if one nation takes an action, another will retaliate.[27]

Mutual restraint can take many forms. Non-proliferation regimes aim to limit access to dangerous technologies.[28] Some

and the Soviet Union limiting intermediate-range nuclear missiles, ballistic missile defences, and nuclear weapons. Since the end of the cold war, some of these bilateral treaties have collapsed or begun to fray.

[25] For example, chemical weapons and the aerial bombardment of undefended cities were prohibited prior to the First World War, but all sides used them. Aerial attacks on undefended cities were also banned prior to the Second World War, but were widely used on all sides. Rules prohibiting surprise attacks by submarines collapsed in both the First and Second World Wars. More recently, chemical weapons have been used by Saddam Hussein and Bashar al-Assad against their own civilian populations, despite being prohibited under both the 1925 Geneva Gas and Bacteriological Protocol and the Chemical Weapons Convention.

[26] In the First World War, Germany unilaterally withdrew use of the sawback bayonet because of a perception among opposing troops that it could cause horrific injuries. During the cold war, the United States and the Soviet Union refrained from large-scale deployment of anti-satellite weapons and neutron bombs, despite the lack of any formal agreement.

[27] The history of chemical weapons use in the Second World War and afterwards points to the importance of reciprocity as a major factor driving restraint. All the major powers in the Second World War had chemical weapons, but they refrained from using them against one another for fear of retaliation. Some countries did use chemical weapons against groups who could not retaliate, however. Japan used them in small amounts against China, who did not have chemical weapons. Germany used poison gas systematically and at a large scale in the Holocaust. In recent years, dictators have used chemical weapons against unprotected civilian populations, but generally not against opposing military forces who had protective equipment and who could retaliate.

[28] Examples of non-proliferation regimes include the Nuclear Non-Proliferation Treaty, the Australia Group, the Missile Technology Control Regime, the

treaties ban the production and stockpiling of certain weapons.[29] Arms limitation treaties limit the quantities of certain weapons that nations are allowed to possess in peacetime.[30] Some treaties permit weapons to be used in some circumstances in war, but not others.[31]

Mutual restraint is impossible without clarity about which weapons or methods of war are prohibited. Some treaties have very detailed definitions proscribing certain weapons, while other treaties focus on the intent behind the weapon.[32] Simple rules have the best track record of success. Complicated rules that allow the development in peacetime

Wassenaar Arrangement and the Hague Code of Conduct. Some of these regimes are not legally binding.

[29] Many early weapons bans in the modern era, such as those in the 1899 and 1907 Hague Declarations and the 1925 Geneva Gas and Bacteriological Protocol banned use of the weapon, but not developing, producing and stockpiling weapons. This meant that nations could develop banned weapons in peacetime and have them in their arsenal, ready for use if necessary. Often, in the heat of war, these prohibitions on use collapsed. This has led some bans in more recent years to prohibit the development, production, stockpiling and use of weapons. This is the case today for biological and chemical weapons, the Mine Ban Convention and the Convention on Cluster Munitions.

[30] Examples of arms limitation treaties include the Anti-Ballistic Missile Treaty (ABM Treaty) (now terminated) and the bilateral nuclear arms control agreements Strategic Arms Limitation Talks (SALT) I, SALT II, Strategic Arms Reduction Treaty (START), Strategic Offensive Reductions Treaty (SORT) and New START.

[31] There are a wide range of examples of treaties that permit weapons to be used in some circumstances, but not others. Some treaties prohibit weapons from certain geographic areas. The Outer Space Treaty, the Sea-Bed Treaty and the Treaties of Tlatelolco, Rarotonga, Bangkok, and Pelindaba all declare certain areas off-limits for nuclear weapons. The Outer Space Treaty and the Antarctic Treaty declare the Moon and Antarctica off-limits from weapons of any kind. Other treaties put procedures in place to try to prohibit the use of weapons near populated areas. These include the 1907 Hague Declaration rules against aerial attacks on undefended cities and the Convention on Certain Conventional Weapons rules on using landmines and incendiary weapons.

[32] As an example of a detailed definition proscribing certain weapons and allowing others, see the definition of “cluster munition” in the Convention on Cluster Munitions. By contrast, the blinding laser ban in the Convention on Certain Conventional Weapons doesn’t specify which technical performance characteristics of lasers are allowable. Instead, it focuses on the intent behind

of certain weapons but not others have had some success.[33] Complicated rules for how weapons may be used in wartime—allowing some uses but not others—have been less successful.[34]

The number of actors needed for cooperation matters. There are many examples of treaties collapsing because one or two countries refused to abide by their terms, even though most countries were in favour of the treaty.[35] This means that it is generally easier to regulate or ban weapons that only a few countries can actually produce.

A major factor in a proposed ban's likelihood of success is the balance between the perceived horribleness of the weapon and its perceived military value. Many weapons that have been successfully banned have only marginal military value but can cause great suffering or are seen as destabilizing.[36] It is significantly harder to achieve restraint with weapons that are perceived to be terrible but also give a decisive advantage in war.[37]

the weapon, prohibiting lasers "specifically designed, as their sole combat function or as one of their combat functions, to cause permanent blindness".

[33] Examples of complicated rules for what weapons nations can produce in peacetime that have generally been reasonably successful (with some exceptions) include the Sea-Bed Treaty, the Outer Space Treaty, the ABM Treaty, the Intermediate-Range Nuclear Forces (INF) Treaty, the SALT I, the SALT II, the START, the SORT, the New START, the Mine Ban Convention and the Convention on Cluster Munitions.

[34] Examples of complicated rules for how militaries can use certain weapons in wartime that have not been particularly successful include the 1907 Hague Declaration rules on aerial attacks and submarine warfare and the rules of the Convention on Certain Conventional Weapons on use of landmines.

[35] Examples of treaties that collapsed because of a failure of cooperation include the 1922 Washington Naval Treaty and, more recently, the collapse of the ABM Treaty and concerns about the INF Treaty in the multi-polar era.

[36] Banned weapons that have limited military value include exploding bullets, weapons on the Moon or in Antarctica, chemical and biological weapons, weapons that cause fragments not detectable by X-ray and blinding lasers.

[37] Contrast, for example, the largely successful prohibition of chemical weapons with the continued persistence—and indeed proliferation—of nuclear weapons. Nuclear weapons are far more horrible and destructive than chemical weapons, but they give a decisive edge on the battlefield that chemical weapons do not.

Even the most successful treaties do not have 100 per cent compliance. Particularly if the technology to build a prohibited weapon is widely available, some terrorists or rogue nations are likely to build them, regardless of the degree of international condemnation. This means that the military value of a weapon (and potential countermeasures) is paramount in whether a ban succeeds or unravels because of cheating.

Formal verification regimes are not necessarily required for success, but transparency is. Particularly if a weapon is seen to be valuable, then nations will want to know that potential adversaries are not cheating. Some treaties have formal verification regimes with inspections to ensure that all parties are adhering to a treaty.[38] Formal verification and inspections may not be needed if nations can observe others' compliance from a distance, for example by using satellites.[39] This method is most effective when the prohibited weapons are large assets that cannot be easily hidden, such as ships, submarines or missile installations.[40] There have been examples of nations cheating and researching prohibited weapons when

[38] The Treaty on the Non-Proliferation of Nuclear Weapons, the Chemical Weapons Convention, the INF Treaty, the START and the New START all have procedures for inspections to verify compliance. The Outer Space Treaty has a de facto inspection regime, requiring States to permit others to view space launches and visit any facilities on the moon. The Mine Ban Convention and the Convention on Cluster Munitions do not have inspection regimes, but do require transparency from States on their stockpile elimination.

[39] The Sea-Bed Treaty, the SALT I and the SALT II treaties, and the now-defunct ABM Treaty, all rely on States to verify others' compliance through their own observations. There are no verification provisions for the 1899 ban on expanding bullets, the 1925 Geneva Gas and Bacteriological Protocol, the Convention on Certain Conventional Weapons, the SORT, the Environmental Modification Convention, the Biological Weapons Convention or the Outer Space Treaty's ban on putting weapons of mass destruction in orbit.

[40] The 1922 Washington Naval Treaty that regulated the size and number of ships that major powers could produce did not include any verification procedures, presumably because it would be hard to hide capital ships.

treaties lack robust monitoring and verification regimes.[41] Without transparency, the fear that potential adversaries could be building prohibited weapons in secret could incentivize countries to start their own secret development programme.[42] Transparency could be a challenge for any potential arms control regime for autonomous weapons if the difference between a prohibited weapon and a permitted one is in the software, which is not observable from the outside.

Pre-emptive bans or regulations have certain unique challenges. On the one hand, it may be easier to ban weapons that countries do not yet depend on for their defence.[43] On the other hand, the fact that the weapon does not yet exist may mean that its military value and horribleness are both in question. Chemical weapons were banned prior to the First World War, but both France and Germany developed them in the hopes that they might prove to be a war-winning weapon. By the Second World War, European nations had learned that poison gas caused significant suffering without yielding a decisive advantage and refrained from battlefield use. Another challenge with pre-

[41] For example, there have been numerous reports of a secret Soviet biological weapons programme after the Soviet Union signed the Biological Weapons Convention in 1972 (Tim Weiner, "Soviet Defector Warns of Biological Weapons", *The New York Times*, 24 February 1998; Milton Leitenberg, Raymond A. Zilinskas, Jens H. Kuhn, *The Soviet Biological Weapons Program: A History* (Cambridge: Harvard University Press, 2012); Ken Alibek, "Biohazard: The Chilling True Story of the Largest Covert Biological Weapons Program in the World—Told from Inside by the Man Who Ran It", (Delta, 2000); and Raymond A. Zilinskas, "The Soviet Biological Weapons Program and Its Legacy in Today's Russia". CSWMD Occasional Paper 11, 18 July 2016).

[42] According to some reports, a factor driving the secret Soviet biological weapons programme was the assumption that the United States similarly had their own secret programme.

[43] The preemptive bans on biological weapons (Biological Weapons Convention) and using the environment as a weapon (Environmental Modification Convention) are examples of preemptive bans of this type, where nations agreed to forgo whole classes of weapons that may or may not have some military value, but which were seen as generally difficult to control and susceptible to indiscriminate use.

emptive bans is achieving a definition that stands the test of time as technology evolves. The most successful pre-emptive bans are simple and focus on the intent behind a weapon, rather than complicated rules about specific characteristics of the weapon.[44]

[44] The ban on blinding lasers, which focuses on the intent behind a weapon rather than technical characteristics, stands as a contrast to other weapons bans that failed to foresee the specific path of technology development. The 1899 Hague declarations banned gas-filled projectiles, but not poison gas in canisters, a technicality that Germany exploited in the First World War in its defence of its first large-scale poison gas attack at Ypres. Similarly, the 1907 Hague rules prohibited aerial attacks against "undefended" cities, a concept that failed to foresee a future in which air defences were largely futile in preventing large-scale bomber raids. The surprising ways that technology can evolve is a major challenge for preemptive bans and regulations.

A robotocist's perspective on lethal autonomous weapon systems

Ronald C. Arkin
School of Interactive Computing
Georgia Institute of Technology

I. Background on lethal autonomous military robotics

Lethal weapon systems are relatively easy to define. Adding autonomy complicates matters significantly. To a philosopher, autonomy adds moral agency and free will to a robotic system, something that does not yet exist and will not for quite some time, if ever. To a roboticist, however, it simply involves the delegation of decision-making to a machine that has been pre-programmed by a human. This chapter will use the following definition for lethal autonomy:

> The ability to "pull the trigger"—to attack a selected target without human initiation nor confirmation, both in case of target choice or attack command (Foss, 2008).

Note: Portions of this paper have appeared in Arkin, R. C., *Governing Lethal Behavior in Autonomous Systems,* Chapman and Hall Imprint, Taylor and Francis Group, Spring 2009 and are reproduced with permission.

This is restricted only in the same sense as a soldier is restricted: the robot soldier must be given a mission to accomplish and any lethal action must be conducted only in support of that mission. At the highest level, a human is still in the loop, so to speak—commanders must define the mission for the autonomous agent, whether it be a human soldier or a robot. The warfighter, robot or human, must then abide by the rules of engagement and laws of war as prescribed from their training or encoding. Autonomy in this sense is limited when compared to a philosopher's point of view.

Confounding this discussion are those who would delineate levels of autonomy as a basis for discussion. There are many different points of view regarding the terms automation versus autonomy, semi-autonomy, teleautonomy, supervised autonomy, on-the-loop versus in-the-loop, mixed initiative, and on and on. It reached such a level of confusion that a recent defence science board report recommended that none of these terms be used. The specific recommendation was that "the DoD [Department of Defense] should abandon the debate over definitions of levels of autonomy"[1] for a "trade space" approach: a method of analysis of trade-offs over multiple stakeholders and objectives. Here we will not try and map individual systems onto particular levels of autonomy other than to say that all of them involve human involvement to some degree—they are not agents with free will to do whatever they want, and are not systems that are likely to be moral agents anytime soon.

Primary motivators for the use of autonomous, robotic or unmanned systems in the battlefield include the following:

- *Force multiplication.* With robots, fewer soldiers are needed for a given mission and an individual soldier can now do the job that took many before.

[1] Department of Defense, Defense Science Board Task Force Report, "The Role of Autonomy in DOD Systems", July 2012, p. 3.

- *Expanding the battle space.* Robots allow combat to be conducted over larger areas than was previously possible.
- *Extending the warfighter's reach.* Robotics enable an individual soldier to reach deeper into the battle space by, for example, seeing or striking farther.
- *Casualty reduction.* Robots permit removing soldiers from the most dangerous and life-threatening missions.

The initial generation of military robots generally operates under direct human control, such as the "drone" or unmanned aerial vehicles being used by the United States military for air attacks (Singer, 2009; Bergen and Tiedemann, 2009). However, as robotics technology continues to advance, a number of factors are pushing many robotic military systems towards increased autonomy. One factor is that as robotic systems perform a larger and more central role in military operations, there is a need to have them continue to function just as a human soldier would if communication channels are disrupted. In addition, as the complexity and speed of these systems grow, it will be increasingly limiting and problematic for performance levels to have to interject relatively slow human decision-making into the process. As one commentator recently put it, "military systems (including weapons) now on the horizon will be too fast, too small, too numerous, and will create an environment too complex for humans to direct" (Adams, 2002).

Based on these trends, many experts believe that autonomous, and in particular lethal autonomous, robots are an inevitable and imminent development (e.g., Arkin, 2009). Indeed, many military robotic-automation systems already operate at the level where the human is still in charge and responsible for the deployment of lethal force, but not in a directly supervisory manner, as detailed below.[2] Examples

[2] At least 30 nations employ or have in development at least one system of this type, including Australia, Bahrain, Belgium, Canada, Chile, China, Egypt, France, Germany, Greece, India, Israel, Japan, Kuwait, the Netherlands, New Zealand, Norway, Pakistan, Poland, Portugal, Qatar,

generally include close-in weapon systems, anti-submarine weapons, cruise missiles, surface-to-air missiles, fire-and-forget missile systems and anti-personnel and other mines.[3]

These devices are considered to be robotic by most definitions, as they are all capable of sensing their environment and actuating through the application of lethal force.

As early as the end of the First World War, the precursors of autonomous unmanned weapons appeared in a project on unpiloted aircraft conducted by the United States Navy and the Sperry Gyroscope Company (Everett, 2015). Numerous unmanned weaponized robotic systems that employ lethal force and have varying degrees of autonomy are already being developed or are in use.

For a complete listing of weaponized robotic platforms past and present, see Arkin, 2009, chap. 2; Everett, 2015; Roff, 2017; and Human Rights Watch, 2012. A recent United States report stated, “New and powerful robotics systems will be used to perform complex actions, make autonomous systems, deliver lethal force, provide ISR [intelligence, surveillance and reconnaissance] coverage, and speed response times over wider areas of the globe.”[4]

II. Ethical autonomy

The development of autonomous, lethal robotics raises questions regarding if and how these systems can adhere to the existing laws of war as well as or better than soldiers. This is

the Russian Federation, Saudi Arabia, South Africa, South Korea, Spain, Taiwan, the United Arab Emirates, the United Kingdom and the United States (Scharre and Horowitz, 2015, p. 12).

[3] Antipersonnel mines have been banned by the Ottawa Treaty, although China, the Russian Federation, the United States and 34 other nations are not party to that agreement.

[4] United States Joint Force Development, “Joint Operating Environment 2035: The Joint Force in a Contested and Disordered World”, 14 July 2016, p. 17.

no simple task. In the fog of war, it is hard enough for a human to effectively determine whether or not a target is legitimate. Despite the current state of the art, it may be anticipated however that, in the future, autonomous robots may be able to perform better than humans under these conditions for the following reasons:

- The ability to act conservatively; i.e., they do not need to protect themselves in cases of low certainty of target identification. Autonomous, armed robotic vehicles do not need to have self-preservation as a foremost drive, if at all. They can be used in a self-sacrificing manner if needed and without reservation.
- The eventual development and use of a broad range of robotic sensors better equipped for battlefield observations than humans currently possess.
- The absence of emotions, which can cloud human judgment or result in anger and frustration with ongoing battlefield events. In addition, "fear and hysteria are always latent in combat, often real, and they press us toward fearful measures" (Walzer, 1977).
- The avoidance of the human, psychological problem of "scenario fulfillment", a factor believed partly contributing to the downing of an Iranian airliner by the USS *Vincennes* in 1988 (Sagan, 1991). This phenomenon leads to the distortion or neglect of contradictory information in stressful situations, where humans use new incoming information in ways that fit their pre-existing belief patterns, a form of premature cognitive closure. Robots can be developed so that they are not vulnerable to such patterns of behaviour.
- The ability of robots to integrate more information from more sources far faster before responding with lethal force than a human possibly could in real time. These data can arise from multiple remote sensors and intelligence (including human) sources.

- When working in a team of combined human soldiers and autonomous systems as an embedded asset, the potential capability of independently and objectively monitoring ethical behaviour in the battlefield by all parties and reporting infractions that might be observed. This presence alone might possibly lead to a reduction in human ethical infractions.

Considerable research is ongoing in terms of endowing intelligent machines with ethical reasoning or the ability to adhere to moral codes as discussed below (Lin and Bekey, 2014). While "there is every reason to believe that ethically sensitive machines can be created" (Anderson, et al., 2004), there is also widespread acknowledgment regarding the difficulty associated with machine ethics (Moor, 2006; McLaren, 2005 and 2006):

1. Ethical laws, codes, or principles are almost always provided in a highly conceptual, abstract level.
2. Their conditions, premises or clauses are not precise, are subject to interpretation and may have different meanings in different contexts.
3. The actions or conclusions following from the rules are often abstract as well, so, even if the rule is known to apply, the ethically appropriate action may be difficult to execute due to its vagueness.
4. These abstract rules often conflict with each other in specific situations. If more than one rule applies, it is not often clear how to resolve the conflict.

In addition, controversy exists about the correct ethical framework to use in the first place, given the multiplicity of philosophies that exist. In the case of international humanitarian law, the just war theory is agreed upon as the basis for ethical behaviour in the battlefield.

A small sampling of recent and ongoing research on ethical software systems designed to work on autonomous systems is reviewed below. This is by no means comprehensive but, rather, is intended to provide a snapshot of the current state of the art.

1. Ethical governors

One specific approach has been used in two very different cases for seeking to ensure or guide ethical responses from intelligent robotic systems: the ethical governor. The ethical governor was originally developed as a prototype for use in the application of lethal force in war by an intelligent autonomous robot. It was designed to ensure that these systems comply with international humanitarian law and the rules of engagement—the guidelines for the conduct of warfare. It did so through the application of negative constraints (prohibitions) derived from international humanitarian law and the rules of engagement, ensuring that no laws of war are violated, and the assurance that a positive constraint (obligation) derived from a human commander was present before an attack was permitted. The design and function of this system is well documented elsewhere (Arkin, et al., 2012; Arkin, 2009).

Recently the same underlying approach has been extended to health care—specifically for the management of patient-caregiver relationships in early-stage Parkinson's disease (Shim, et al., 2017). An intervening ethical governor has been designed to help provide a restorative force when this human-human relationship starts to veer beyond acceptable bounds. The intervening ethical governor uses rules derived from occupational therapy manuals, so that a small humanoid robot can intervene when required, as would be the case for a human occupational therapist.

The broad applicability of the ethical governor for enforcing either legal or social norms in a range of applications for autonomous robots should now be apparent. Others such as Welsh (2017) have extended the concept of the ethical governor using deontic logic, the moral logic of obligations, permissions and prohibitions, to a variety of new domains.

2. Ethical autonomous unmanned undersea vehicles

An example from the United States Naval Postgraduate School involves unmanned undersea vehicles using constraints for "runtime ethics" (Brutzman, et al., 2012 and 2013). Similar to the ethical governor (Arkin, 2009), they use these constraints to monitor the actual execution of the mission for ethical constraint violations before they occur, thus observing the rules of engagement during mission conduct. Their approach entails developing a set of plans using ethical reasoning and then validates them for correctness. Their system is tested in the context of ethical unmanned undersea vehicle search, ensuring that regions that are off-limits to the robot are avoided while still successfully conducting the higher-level mission goals (Davis, et al., 2016).

3. Verifiably ethical autonomous systems

To ensure that ethical behavior is actually obtained, formal verification methods are crucial. Research in the United Kingdom (Dennis, et al., 2013, 2015 and 2016) specifically addresses this area using a Beliefs-Desires-Intentions rational agent architecture with ethical checking to ensure that it selects the most ethical plan available. As in many other pragmatic systems, the ethical principles come from existing rules from society. In this system, the rules are represented in the context of airmanship for unmanned aircraft in civilian aviation, addressing, for example, concerns that arise from low fuel or erratic intruders into common airspace. Their architecture seems readily generalizable to other domains, such as driverless cars and beyond.

4. Case-based ethics for robots

Researchers have investigated using a small humanoid robot to assist in eldercare (Anderson, et al., 2016; Anderson, et al., 2017), using a "case-supported principle-based behaviour paradigm", initially tested only in simulation. The robot identifies the situation it is in, looks at a set of possible actions and then selects the most ethically preferable one (as determined by human ethicists' evaluations a priori). The action predicates are associated with duty satisfaction/violation values, where these duties include rights that serve as guiding principles, such as minimizing harm, respecting autonomy, preventing immobility and the like.

5. Ethical robot architecture

Research in Bristol (Vanderelst and Winfield, 2016) has led to the development of an implemented ethical robot architecture. The system incorporates a discrete ethical layer sitting atop the more traditional robot controller, incorporating a set of ethical rules to determine appropriate courses of action for specific goals. This layer verifies behaviours with respect to ethical performance that are forwarded by the robot controller and can suggest others that are more ethically suitable. Prediction of the consequences of the goals and tasks is then undertaken, followed by evaluation of the predictions, leading to more ethical behaviour than would be achieved otherwise by the robot controller alone. The system was tested on two small humanoid robots to demonstrate an interpretation of Asimov's laws with respect to self-preservation, obedience and human safety. The approach is consequentialist, as it is judged by outcomes rather than inherent duties.

In all these cases, the field of ethical autonomy is still in very early stages of basic research and, although there are hopeful examples that this technology may someday feasibly apply in the battle space, this is likely a decade or two away.

Given the pressing rate of progress in robotics/autonomy as a whole and its rapid penetration in society, it is important that the field move forward post-haste to ensure the safe and ethical deployment of intelligent autonomous robots, especially in the context of armed conflict.

Concurrently, there are major efforts being conducted worldwide aiming to develop policies and standards for the development of these systems. One notable effort is the Institute of Electrical and Electronic Engineers Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems.[5] This strongly interdisciplinary effort and other related ones require worldwide involvement to ensure that the systems we create meet our ethical and societal expectations.

References

Adams, T. (2002). Future Warfare and the Decline of Human Decisionmaking. *Parameters*. U.S. Army War College Quarterly, Winter 2001-02, pp. 57-71.

Anderson, M., S. Anderson and C. Armen (2004). Towards Machine Ethics. *AAAI-04 Workshop on Agent Organizations: Theory and Practice.* San Jose, CA.

Anderson, M., S. Anderson and V. Berenz (2016). Ensuring Ethical Behavior from Autonomous Systems. *Proc. AAAI Workshop: Artificial Intelligence Applied to Assistive Technologies and Smart Environments*. Available from http://www.aaai.org/ocs/index.php/WS/AAAIW16/paper/view/12555.

___________ (2017). A Value Driven Agent: Instantiation of a Case-Supported Principle-Based Behavior Paradigm.

Arkin, R. C. (2009). *Governing Lethal Behavior in Autonomous Robots*. Taylor-Francis.

[5] Available from http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html.

Arkin, R.C., P. Ulam and A. R. Wagner (2012). Moral Decision-making in Autonomous Systems: Enforcement, Moral Emotions, Dignity, Trust and Deception. *Proceedings of the IEEE*, vol. 100, no. 3, pp. 571-589.

Bergen, P. and K. Tiedemann (2009). *Revenge of the Drones: An Analysis of Drone Strikes in Pakistan*. New America Foundation.

Brutzman, D., D. Davis, G. Lucas and R. McGhee (2013). Run-time Ethics Checking for Autonomous Unmanned Vehicles: Developing a Practical Approach. *Proc. 18th International Symposium on Unmanned Untethered Submersible Technology*. Portsmouth, NH.

Brutzman, D., R. McGhee and D. Davis (2012). An implemented universal mission controller with run time ethics checking for autonomous unmanned vehicles—A UUV example. *Autonomous Underwater Vehicles (AUV), 2012 IEEE/OES*. Institute of Electrical and Electronic Engineers.

Davis, D., D. Brutzman, C. Blais and R. McGhee (2016). Ethical mission definition and execution for maritime robotic vehicles: A practical approach. *OCEANS 2016 MTS/IEEE Monterey*, pp. 1-10.

Dennis, Louise, et al. (2013). Ethical choice in unforeseen circumstances. *Conference Towards Autonomous Robotic Systems*. Springer Berlin Heidelberg.

Dennis, Louise A., M. Fisher and A. Winfield (2015). Towards verifiably ethical robot behaviour, arXiv preprint arXiv:1504.03592.

Dennis, Louise, et al. (2016). Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems* 77: 1-14.

Everett, B. (2015). *Unmanned Systems of World Wars I and II.* MIT Press.

Foss, M. (2008). What are Autonomous Weapon Systems and What Ethical Issues do they Raise.

Hawkley, J. (2017). Patriot Wars: Automation and the Patriot Air and Missile Defense System. CNAS Ethical Autonomy Series.

Human Rights Watch (2012). Losing Humanity: The Case Against Killer Robots.

Lin, P., K. Abney and G. Bekey (2014), eds. *Robot Ethics: The Ethical and Social Implications of Robotics*. MIT Press.

Moor, J. (2006). The Nature, Importance, and Difficulty of Machine Ethics. *IEEE Intelligent Systems*, July/August, pp. 18-21.

McLaren, B. (2005). Lessons in Machine Ethics from the Perspective of Two Computational Models of Ethical Reasoning. *2005 AAAI Fall Symposium on Machine Ethics*. AAAI Technical Report FS-05-06.

___________ (2006). Computational Models of Ethical Reasoning: Challenges, Initial Steps, and Future Directions. *IEEE Intelligent Systems*, July/August, pp. 29-37.

Roff, H. Dataset: Survey of Autonomous Weapons Systems. Available from https://globalsecurity.asu.edu/robotics-autonomy (accessed on 13 June 2017).

Sagan, S. (1991). Rules of Engagement. *Avoiding War: Problems of Crisis Management*, A. George, ed.. Westview Press.

Scharre, P. and M. Horowitz (2015). An Introduction to Autonomy in Weapon Systems. CNAS Working Paper.

Shim, J., R. C. Arkin and M. Pettinati (2017). An Intervening Ethical Governor for a robot mediator in patient-caregiver relationship: Implementation and Evaluation. *Proc. ICRA 2017*. Singapore.

Singer, P. W. (2009). *Wired for War.* Penguin.

Walzer, M. (1977). *Just and Unjust Wars,* 4th edition. Basic Books.

Welsh, S. (2017). Moral Code: Programming the Ethical Robot, PhD dissertation (draft). University of Canterbury.

Vanderelst, D. and A. Winfield (2016). An Architecture for Ethical Robots, arXiv:1609.02931 (cs.RO).

An ethical perspective on autonomous weapon systems

Karolina Zawieska
Industrial Research Institute for Automation and Measurements

As designers endow more weapon systems with degrees of autonomy and other characteristics that aim to equal and even surpass human capacities for discernment, serious ethical concerns persist over the possibility of delegating life-and-death decisions to autonomous weapon systems. Responses to such concerns often emphasize a system's ability or inability to comply with principles of international humanitarian law—distinguishing civilians from military targets while ensuring that harm to the former is in acceptable proportion to the value of the latter—but such analyses regularly fall victim to one of two blind spots: either taking ethical conduct by humans for granted ("humans are ethical, and robots are not") or giving up on human morality altogether ("humans fail to act ethically, so we need ethical robots"). Rather than addressing the specific characteristics of prospective autonomous weapon systems, this chapter will examine concepts such as the human-machine analogy in considering what ethical principles should inform an inherently human act.

The international debate on autonomous weapon systems embraces a variety of perspectives and disciplines, but the current discussion contains certain unexamined assumptions about human nature that unnecessarily limit our visions for how to control such systems. The following pages challenge these assumptions from a humanistic perspective, ascribing to human beings sufficient will to determine how they will apply ethical principles to the potential use of these systems.

On the "inevitability" of autonomous weapon systems

Commentators generally describe autonomous weapon systems in the context of emerging technologies rather than existing weapons, but they often characterize the development and use of such systems as "inevitable" much as they tend to take the continued occurrence of war as a given. For example, one author argued, "Warfare will continue and autonomous robots will ultimately be deployed in its conduct."[1] Another analyst said that "autonomous weapon systems are the next logical and seemingly inevitable step in the continuing evolution of military technologies".[2] Proponents of this line of thinking see technological innovation as "a must" in the development of armed forces,[3] and they often cite the current use and development of semi-autonomous weapon systems to make the case that it is "too late" to stop the use of autonomous weapon systems.

[1] R. Arkin, *Governing Lethal Behavior in Autonomous Robots* (CRC Press, 2009), p. 29.

[2] J. M. Beard, "Autonomous Weapons and Human Responsibilities", *Georgetown Journal of International Law,* vol. 45 (2014), p. 620.

[3] Advisory Council on International Affairs (AIV) and Advisory Committee on Issues of Public International Law (CAAV), *Autonomous weapon systems: the need for meaningful human control*, document No. 97 AIV/No. 26 CAAV, October 2015.

The inevitability of these systems is not a matter of consensus, however.[4,5] In fact, such military and technological determinism deserves our firm rejection.

The decision to deploy autonomous weapon systems, in particular lethal autonomous weapon systems, is a choice that has yet to be made, and parameters for their potential deployment have yet to be agreed upon. Recognition of this fact leaves room to discuss not only *how* to use and manage weapon systems endowed with different degrees of autonomy (for example, what functions should be subject to meaningful human control), but also *if* using such weapons is justified at all.

The supposed ability of autonomous weapon systems to strengthen the application of humanitarian principles in armed conflict is commonly cited as a justification for the development and use of such systems, but this argument merits close scrutiny. The design and use of autonomous weapon systems to reduce fatalities, unnecessary suffering and the risk of war crimes would, in realistic terms, make war more palatable for its perpetrators by diminishing their sense of direct responsibility for victims. The potential for autonomous weapon systems to further distance humans from their violent actions constitutes one of the most compelling arguments against such systems. Furthermore, applying humanitarian principles in the pursuit of military objectives has historically proven difficult, and the ability of autonomous weapon systems to adhere to ethical standards of conduct is even less certain. Because tactical superiority is the main goal of military research[6] and technology

[4] United Nations Institute for Disarmament Research (UNIDIR), “Framing Discussions on the Weaponization of Increasingly Autonomous Technologies”, UNIDIR Resources, No. 1 (Geneva, 2014).

[5] N. E. Sharkey, “The evitability of autonomous robot warfare”, *International Review of the Red Cross* 94, No. 886 (2012), pp. 787-799.

[6] R. Arkin, *Governing Lethal Behavior in Autonomous Robots* (CRC Press, 2009), p. 41.

plays a key role in ensuring military competitiveness,[7] designers of autonomous weapon systems may see little practical reason to restrain their lethal potential. The use of autonomous weapons could, in other words, lead to "inhumanely efficient" wars.[8]

The prospect of humans waging war on such profoundly impersonal terms underscores the need to address the wider context of military and civilian technological progress in discussing whether to limit or prohibit the use of autonomous weapon systems. The consequences of the debate over autonomous weapon systems will extend far beyond the use of particular weapons, making it critical for participants to apply underlying ethical principles in a manner that deliberately avoids defining human beings in increasingly machine-like terms. A rigorous examination of the ethical framework for this debate may ultimately demand steps to strengthen and even redefine existing human rights protections.

On the human-machine analogy

To date, these discussions have referred to their subject alternately as autonomous weapon systems (AWS), lethal autonomous weapon systems (LAWS), lethal autonomous robots (LARs) and increasingly autonomous weapon systems. What all of these terms share is, of course, a focus on autonomy.

The concept of robot autonomy, along with popular references to "machine decision-making" and "machine learning", reflect the human tendency to project human characteristics onto non-human objects, phenomena and entities. Such anthropomorphizing language is useful to explain the complex technical processes and components of autonomous systems, but it reveals an increasingly widespread conviction that human qualities can literally be reproduced in a machine, as

[7] A. Krishnan, *Killer robots: legality and ethicality of autonomous weapons* (Ashgate Publishing, Ltd., 2009), p. 120.

[8] Ibid., p. 130.

well as an underlying belief that human and non-human systems are different only by a matter of degree.

The growing perception that humans and machines share substantial similarities has encouraged the use of engineering and computer science terms to describe human beings, for example, as "complex systems",[9] "goal-based mechanisms"[10] or "collections of components".[11] Such thinking can also extend to the field of human ethics, where an ethical principle might be described as "an ethically consistent protocol" or as "ethical processing"[12] to imply that human and non-human agents can produce the same ethical results.[13] Such language can contribute to the perception of humans as analogous to machines, even when used to argue that fully autonomous weapon systems would lack human qualities.[14]

It is now possible to debate whether machines can uphold human ethical principles not because autonomous systems have become more like humans, but because human beings increasingly view themselves as they view machines. The validity of the human-machine or brain-computer analogy often goes unquestioned, much like the assumptions that the deployment of autonomous weapon systems and the perpetuation of warfare are inevitable. The power that this analogy already exerts over our world provides a clear warning: whether or not we allow autonomous systems to

[9] T. Fong, I.R. Nourbakhsh and K. Dautenhahn, "A survey of socially interactive robots"*, Robotics and Autonomous Systems*, vol. 42 (3-4) (2003), pp. 143-166.

[10] S. Šabanović, "Emotion in Robot Cultures: Cultural Models of Affect in Social Robot Design", *Proceedings of the 7th International Conference on Design and Emotion* (Chicago, October 2010).

[11] R. A. Brooks, *Flesh and Machines: How Robots Will Change Us* (New York, Pantheon Books, 2002).

[12] R. Arkin, *Governing Lethal Behavior in Autonomous Robots* (CRC Press, 2009), p. 94.

[13] Ibid.

[14] B. Docherty, *Losing humanity: The case against killer robots* (November 2012).

make life-and-death decisions on our behalf will shape not just political and military outcomes, but also our self-conception as humans. Taking humans out of the loop of battle risks "losing humanity"[15] in a broader sense; in the words of one statement to the United Nations Human Rights Council, "Taking humans out of the loop also risks taking humanity out of the loop."[16]

This is not the only way forward. By approaching the ethical debate over autonomous weapon systems from the perspective of humans rather than machines, it is possible to acknowledge human distinctiveness and complexity rather than dismissing or oversimplifying it. From this perspective, machine autonomy and machine intelligence are problematic terms, as they refer to functions that are radically different from the human processes that we describe with similar language. Therefore, when reflecting on the use and characteristics of autonomous weapon systems, we should use related terminology carefully, shaping it in a way that emphasizes rather than obscures the difference between what is human and what is only human-like.

On demands for improvement

Proponents of autonomous weapon systems often claim that such systems may equal and eventually outperform human beings, not just at narrowly defined tasks, but also at complex processes like the application of ethics. The implication is that human capacities and conduct would benefit from augmentation with autonomous technologies, or through substitution with fully autonomous weapon systems. While such an approach may be well intentioned, with aims such as minimizing harm on the battlefield, it places in question both the core morality

[15] B. Docherty, *Losing humanity: The case against killer robots* (November 2012).

[16] C. Heyns, "Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions", United Nations Human Rights Council, United Nations document A/HRC/23/47, 9 April 2013, p. 17.

of humans and the relative shortcomings in their performance (a word commonly employed to blur the distinction between human and machine standards of achievement).[17] The perception that autonomous systems can achieve perfection or at least surpass humans at many tasks may contribute to a mistaken belief that, in areas such as ethical behaviour, real improvement is only achievable through the use of autonomous systems.

While it is indisputable that humans do not act ethically in every circumstance, some proponents of autonomous weapon systems have advanced their argument with the fallacious suggestion that humans behave unethically or inadequately as a general rule. This stance does not reflect actual human moral or ethical experience, and it would be a mistake to expect machines to be "a better version of human beings".[18] Rather, we should acknowledge the inherently human nature of ethical values and take responsibility for ethical conduct. The fact that such values and human rights often remain in the realm of aspirations rather than actual conduct should not stop us from pursuing ethical principles, and hence, a fuller expression of our own humanity.

Conclusions

The developmental status of autonomous weapon systems complicates efforts to understand their ethical implications, but, even after their potential development and deployment, the technology would require continuous monitoring and analysis to follow its evolution and accompanying challenges to existing legal and socio-cultural frameworks. A similar approach applies to ethical principles: while universal human values may appear well defined and widely accepted, ethical principles require constant reflection and our ongoing commitment to conduct ourselves in accordance with them. We should therefore never

[17] R. A. Brooks, *Flesh and Machines: How Robots Will Change Us* (New York, Pantheon Books, 2002).

[18] M. Anderson and S. L. Anderson, "Machine ethics: Creating an ethical intelligent agent", *AI Magazine,* vol. 28, No. 4 (2007), p. 15.

stop attempting to actively shape the direction of development in autonomous weapon systems, just as we should never cease in our efforts to refine our conception of distinctively human ethics.

A military perspective on lethal autonomous weapon systems

Ajey Lele
Group Captain (retired)
Institute for Defence Studies and Analyses

War is considered as a last resort after all peaceful options to end a dispute have been exhausted. Nations are expected to refrain from targeting innocent civilian populations. Furthermore, disproportionate military action and collateral damage should be avoided with the ultimate aim being the restoration of peace. The conflicts of the twentieth and twenty-first centuries have shown that, on occasion, avoiding civilian casualties becomes challenging owing to the blurring of lines between soldiers, non-State actors and non-combatants.

History demonstrates that technology can shape how war is fought. Advancements in various weapons and weapon systems influence the methods of waging war and nation-States evolve their doctrines accordingly. With advancements in the autonomy of modern-day weapon systems, there is a need to establish a context for their military applicability. This chapter identifies

Note: This chapter is informed by, among others, the ICRC publication *Autonomous Weapon Systems, Expert Meeting, Versoix, Switzerland, 15-16 March 2016.*

and analyses the possible military applications and implications of lethal autonomous weapon systems (LAWS).

Autonomous technology can function on its own, without human intervention. LAWS are weapon systems that identify and attack a target without human intervention. At present, such systems are mostly so-called “fire-and-forget” systems, which, once activated, select and engage targets on their own without any human intervention.

LAWS could be considered as either offensive or defensive systems. However, today, the only deployed fully autonomous systems belong to the defensive category. This may be because it is arguably impossible for a weapon system to choose a target on its own, since no machine can decide why, when, where and how to start a conflict unless, and until, it is programmed to do so. Hence, LAWS, as they are deployed today, could be considered as defensive weapon systems, which are programmed to respond to incoming threats.

Almost all the prevailing autonomous weapon systems (mainly used in missile and air defence roles) are designed as point defence or area defence weapon systems. Such systems respond to incoming missile threats, but are not capable of launching an attack independently. To this author’s knowledge, no weapon system to date has been designed and programmed that can decide to engage a (human) target on its own.

The United States has operated armed ground robots like the Special Weapons Observation Reconnaissance Direct-action System (SWORDS), which has been deployed in Afghanistan for detecting and disabling improvised explosive devices. SWORDS was the first weaponized unmanned ground vehicle. Such robotic systems have limited inbuilt artificial intelligence and are remotely operated by a soldier. Such systems indicate that similar systems, with a capability of firing without human intervention or oversight, could be designed and developed. Such systems would be categorized as offensive LAWS.

Israel has developed a loitering munition designed to target radio emissions, the Harop, which can loiter for up to six hours. It autonomously homes in on radio emissions, using itself as a munition.

A further example could be the sentry robot SGR-1, which has been developed by Samsung Techwin. Presently, this system has been used by the Republic of Korea along its border with the Democratic People's Republic of Korea. This robot can detect targets from a distance of around 3.5 km, however, the final order to fire is currently given by a human operator.

At present, the only fully autonomous weapon systems that are completely operational are counter-rocket, artillery and mortar systems, such as the so-called "Iron Dome"; anti-missile systems, such as the Terminal High Altitude Area Defense (THAAD); and anti-aircraft systems, such as the S-400. In addition, there are systems based on robotic technologies, like drones and unmanned ground and underwater vehicles, which are able to navigate, but not select and engage targets, autonomously.

A close-in weapon system is a point defence system used for defence against short-range anti-ship missiles. These systems are also useful for engaging enemy aircraft that have successfully infiltrated outer defences to approach the target (normally, a battle ship or tanker ship) with high speed. Land-based close-in weapon systems can also address threats like shell bombardment and rocket fire. All major maritime forces in the world are equipped with close-in weapon systems. These systems could also be used on land to protect military bases. Such systems have both gun-based and missile variants. The gun-based system comprises multiple-barrel, rotary rapid-fire cannons placed on a rotating gun mount. Both variants require various types of passive and active radar units for providing terminal guidance.

The Iron Dome has proved its effectiveness for short-range applications. It is a system conceptualized by Israel and jointly funded by the United States. It is a counter-rocket, artillery and mortar system capable of intercepting multiple targets from any direction. The Iron Dome uses an autonomous guidance and control system capable of intercepting specific targets that represent a high-priority threat according to the system configuration. The Tamir Adir system is a sea-based variant of an Iron Dome missile battery, developed as an autonomous maritime missile interception system. Israel conducted a successful test of this system in May 2016. The Tamir Adir system is capable of engaging and destroying airborne targets from a moving platform.

THAAD is designed to defend against short- and medium-range ballistic missiles. This system claims to have a 100 per cent intercept test success rate. The entire system architecture relies on other important elements like radars and satellites. The system operates in a fully autonomous mode whereby an infrared satellite detects an incoming missile's heat signature and sends an early warning and other real-time tracking data to the ground-based system through a communications satellite. When the threat is confirmed, based on an assessment carried out from inputs received from various early warning systems, a suitable command is conveyed to the sensors and weapon systems (such commands put the weapon system into active mode). Subsequently, the long-range radar detects and tracks the missile for some time to further improve the accuracy. The tracking data helps to compute the trajectory of the incoming threat missile. Among the group of batteries available to address the threat, the most effective interceptor battery is engaged and carries out the interception. The complete process of identifying, engaging and destroying the missile is fully autonomous in nature and is known to have very high efficiency.

An ongoing project by the United States defence establishment is the development of armed drone swarms, unmanned flying units that fly in formation to achieve a given task. For example, the Perdix system consists of autonomous drones operating as cooperative swarms of 20 or more flying units. The drones are launched to achieve a specific goal, are expected to engage in collective decision-making and are known to possess swarm self-healing abilities, whereby, in case one or more drone units are forced to dropout, the entire system reconfigures itself automatically for mission completion. It should be noted that the autonomy of this system relates to navigation, not target selection or engagement. Work on this system began in 2013. Since then, several testing missions have been launched and the system's software is currently being upgraded.

Other LAWS, which are still either at the drawing board level or in the realm of theoretical possibilities, also warrant attention. These include space-based autonomous systems, which could be used to target space-based systems, as well as targets on Earth. There exists a possibility that, with the overall growth in the technology sector globally, some capable States could seek to develop such systems in the near future.

The nature of warfare is ever evolving. An increasingly automated battlefield is expected to add another dimension to warfare, which will have a mixed impact on militaries. States are bound to develop countermeasures (and counter-countermeasures) to LAWS. In general, LAWS are likely to continue to have relevance both as tactical and strategic weapon systems. It is important to note that autonomy cannot be thought of in absolute terms; there may be either low or high levels of autonomy. Arguably, militaries will be required to keep these weapons under their effective control and decide about the contexts in which they can be deployed, as well as the nature and degree of autonomy allowed for any given deployment. Militaries will also be required to effectively navigate the various legal challenges, arms control considerations and moral issues related to LAWS in order to continue to keep these weapons in their arsenals. Today, LAWS provide both opportunities and vulnerabilities for militaries. Hence, it is necessary for militaries to incorporate such weapons into their war-fighting doctrines with due diligence.

ISBN 978-92-1-142324-2

9 789211 423242

17-18665