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.