HUMANITARIAN LAW & POLICY

Three lessons on the regulation of autonomous weapons systems to ensure accountability for violations of IHL

March 2, 2023, Analysis / Autonomous Weapons / Humanitarian Action / Law and Conflict / New Technologies

11 mins read

Vincent Boulanin
Senior Researcher, SIPRI

Marta Bo
Associate Senior Researcher, SIPRI

States have agreed on the principle that machines cannot be held accountable for violations of international humanitarian law (IHL), but how would accountability be ensured in practice?

In this post, Vincent Boulanin and Marta Bo from the Stockholm International Peace Research Institute (SIPRI) argue that looking at how responsibility for IHL violations is currently ascribed under international law provides useful lessons for the regulation of AWS.

ICRC Humanitarian Law & Policy Blog · 3 lessons on the regulation of AWS to ensure accountability for violations of IHL

What if the use of an autonomous weapon system (AWS) during an armed conflict resulted in the death or injury of civilians or damage to civilian objects? Whereas not all harm to civilians is illegal under international humanitarian law (IHL), launching an attack against civilians or civilian objects amounts to a violation. But how could accountability for such a violation be ensured?

We argue that looking at how responsibility for IHL violations is currently ascribed under international law is critical not only to ensuring accountability but also to identifying clearer limits and requirements for the development and use of AWS.

Human responsibility and accountability in the GGE debate

The risks posed by autonomous weapon systems have been the focus of intergovernmental discussions at the UN Convention on Certain Conventional Weapons (UN CCW) since 2013. States still disagree on whether and how the development and use of AWS should be (further) regulated, but they have recognized, among other principles, that *human responsibility for decisions on the use of weapon systems must be retained, since accountability cannot be transferred to machines*.

The question of what this principle entails is critical for the continuation of the policy process on AWS. To date, the expert debate has mainly elaborated on *how human responsibility should be exercised – preventively – to ensure compliance with IHL*. Less attention has been cast on how accountability would be ensured, in practice, in case of IHL violations involving AWS. The Group of Governmental Experts (the GGE) recognized that the rules governing State responsibility for internationally wrongful acts and individual criminal responsibility for war crimes were the relevant legal framework, but discussed their application in the context of AWS only superficially.

In our view, this represents a major gap in policy conversation, first and foremost because preventing and suppressing IHL violations is part of States' obligations under Additional Protocol 1 (AP I) to the Geneva Conventions and customary law. We also found that reviewing how these rules apply to IHL violations involving AWS could provide important lessons for the intergovernmental debate on the regulation of AWS.

Here are three lessons we identified in our most recent report on *Retaining Human Responsibility in the Development and Use of Autonomous Weapons Systems: on Accountability for Violation of International Humanitarian Law involving AWS*.

Three lessons for the regulation of AWS

Lesson 1. Discerning IHL violations in the development and use of AWS will remain challenging without further clarification on what IHL permits, requires, and prohibits

The first lesson is that legal clarification will be needed to ensure that the legal framework governing accountability can be effectively triggered.

The rules governing State responsibility for internationally wrongful acts and individual criminal responsibility for war crimes are linked to IHL. Both the establishment of State responsibility for IHL violations and individual criminal responsibility for war crimes depend on normative standards established by IHL rules. The fact that the *debate on IHL compliance in the development and use of AWS is still unsettled* presents, in that context, a fundamental challenge. Many questions remain about what IHL requires, permits, and prohibits, for instance in terms of human-machine interaction. This means that the basis for establishing that a State or an individual violates IHL is still, in some cases, unclear, or at least subject to different interpretations.

AWS bring also into new light old and unresolved legal disputes around the standards of conduct that would trigger State responsibility or individual criminal responsibility for war crimes (or both). For instance, it has been debated to what extent a violation of the *principle of distinction has to be 'deliberate' for State responsibility to arise*. And it is an open question whether *recklessness or omission satisfy the mental and material elements of perpetrating or participating in the commission of a war crime*. The fact that AWS are pre-programmed weapons, which are ultimately triggered by the interaction with the environment rather than direct user input, gives these *debates new resonance but also new scenarios to deal with*. For instance, would a failure to suspend an attack involving an AWS that is expected to harm civilians be considered a deliberate attack on civilians and amount to a war crime?

These questions and controversies underline the need for the policy process on AWS to achieve more precision and a common understanding of IHL compliance. In particular, they invite the GGE to elaborate on standards of intent, knowledge and behaviour that are demanded on the part of the user(s) of AWS. Clarifying what the user(s) of an AWS should be able to reasonably foresee and do to ensure that the AWS attack is directed at a specific military objective and the effects of the weapon are limited as required by IHL would make it easier to determine whether a violation has been committed intentionally or that the user engaged in risk-taking behaviour that could give rise to State responsibility, and individual criminal responsibility or both.

Lesson 2. Elaboration on what constitutes a 'responsible human chain of command' could help with the attribution of responsibility

The second lesson is that the policy process needs to unpack the notion of 'responsible human chain of command'. Elaboration on how such a chain may look could dramatically facilitate the attribution of responsibility, be it to the State or individual.

Some States and experts have expressed the concern that, in the case of a harmful incident involving an AWS for instance, it could be difficult to identify whose conduct is blameworthy given that the operation, performance and effect of an AWS were determined in part by decisions and actions of multiple individuals involved in the development and use of the systems; as well as the interaction of the system with the environment.

We argue in this context that it would be extremely useful if States could elaborate on what a scheme of responsibility for the development and use of AWS could look like. Such a scheme would provide more clarity on how the roles and responsibilities for IHL compliance may or may not be distributed in practice: who should do what, when and where the roles and responsibilities of the different individuals start and end and how might these interact with one another. Such an effort would be doubly beneficial. On the one hand it would strengthen IHL compliance by providing clearer expectations for the users of AWS. On the other, it could make it easier to detect who engaged in unlawful conduct that could give rise to State responsibility, individual criminal responsibility (or both).

Lesson 3. Traceability is a critical component for the regulation of AWS

The third lesson is that traceability – understood here as the ability to trace the operation, performance and effect of an AWS back to people involved in its development and use – should be regarded as a critical component of further regulation of AWS. It should inform the identification of new limits and requirements on the design and use of AWS – for two reasons.

First and foremost, traceability is a practical requirement for complying with States' obligations under international law. Under AP I, States are obliged to repress war crimes, including searching for individuals responsible, and suppressing any other violations of IHL. To be able to perform these obligations,

States need to be able to determine whether illegal conduct took place and, if so, identify blameworthy individuals. Second, it is also a practical requirement to assess and impose State responsibility, individual criminal responsibility or both.

If an attack involving an AWS results in the deaths of civilians – both the States with jurisdiction over the incident*[1]* and other States and institutions that are entitled to investigate the incident, such as the ICC or fact-finding commissions, would need to determine whether the deaths were caused by a technical failure or unlawful conduct on the part of the user(s) and/or developers of the AWS. This demands a practical ability to scrutinize the operation, performance and effect of AWS and trace back whether and how these result from decisions and actions made by people involved in the development and use of AWS.

Certain emerging technologies in the area of AWS, such as certain approaches to *artificial intelligence and machine learning (ML), could make the task of investigating the cause of an incident difficult*. Machine learning methods, such as deep learning, could offer military benefits but they are also opaque in their functioning. As they stand, current ML techniques used in target recognition software are not explainable which means that a programmer or a user cannot fully understand how they learn to recognize a target type. This opacity could make it difficult to determine after the fact what caused a system to strike civilians or civilian objects. Even in situations where a technical problem can be excluded, attribution problems could also emerge as the operation, performance and effect of the AWS is determined by decisions made by multiple people at different points in time and, in part, depends on the interaction of the AWS with the environment. Tracing back whose conduct is blameworthy could be difficult.

The takeaway here for the regulation of AWS is two-fold. Should States decide to explicitly prohibit AWS that are incompatible with IHL or otherwise posing unacceptable risks to civilians and other protected persons, such a prohibition should make explicit that technical characteristics and forms of human-machine interaction that preclude the ability to trace back the cause of a harmful incident are off-limits. That could include unexplainable machine learning algorithms. Efforts to codify lawful uses of AWS could, on the other hand, make traceability a critical requirement for the design and use of an AWS. On the technical side, that could entail that algorithms on which the targeting functions are based should be transparent, explainable, and interrogable enough to identify legal/illegal conduct and blameworthy individuals. On the organisational side, that could entail, as suggested, developing and using, a scheme of responsibility, but also mechanisms to record and trace back decisions in the development and use of AWS.

Exploring how accountability for IHL violations involving AWS would be ensured may seem to some actors premature if not irrelevant, as the use of AWS, is – depending on one's understanding of an AWS – not yet an operational reality. With this post we hope to have demonstrated that it is a useful and much-needed exercise for the policy process on AWS, as it provides a lens to explore what is, or should be, demanded, permitted, and prohibited in the development and use of AWS.

[1] On the basis of the territoriality, active personality, and universal jurisdiction principles.

See also

- Laura Bruun, *Autonomous weapon systems: what the law says – and does not say – about the human role in the use of force*, November 11, 2021
- Frank Sauer, *Autonomy in weapons systems: playing catch up with technology*, September 29, 2021
- Neil C. Renic, *Autonomous Weapons Systems: When is the right time to regulate?*, September 26, 2019
- ICRC, Neil Davison, *Autonomous weapon systems: An ethical basis for human control?* April 3, 2018

Tags: autonomous weapon systems, autonomous weapons, AWS, GGE, IHL, international humanitarian law, international law, UN Convention on Certain Conventional Weapons

You may also be interested in:

Preventing and eradicating the deadly legacy of explosive remnants of war

13 mins read

Analysis / Autonomous Weapons / Humanitarian Action / Law and Conflict / New Technologies Eirini Giorgou

Humanitarian neutrality in contemporary armed conflict: a conversation with Nils Melzer

13 mins read

Analysis / Autonomous Weapons / Humanitarian Action / Law and Conflict / New Technologies Nils Melzer & Elizabeth Rushing

The deadly legacy of armed conflict continues to claim lives long after the fighting is …

As with many humanitarian crises in the past, the international armed conflict between Russia and …