TOWARDS A TWO-TIERED APPROACH TO REGULATION OF AUTONOMOUS WEAPON SYSTEMS

Identifying Pathways and Possible Elements

LAURA BRUUN

STOCKHOLM INTERNATIONAL
PEACE RESEARCH INSTITUTE

SIPRI is an independent international institute dedicated to research into conflict, armaments, arms control and disarmament. Established in 1966, SIPRI provides data, analysis and recommendations, based on open sources, to policymakers, researchers, media and the interested public.

The Governing Board is not responsible for the views expressed in the publications of the Institute.

STOCKHOLM INTERNATIONAL
PEACE RESEARCH INSTITUTE
Signalistgatan 9
SE-169 70 Solna, Sweden
Telephone: +46 8 655 97 00
Email: sipri@sipri.org
Internet: www.sipri.org

TOWARDS A TWO-TIERED APPROACH TO REGULATION OF AUTONOMOUS WEAPON SYSTEMS

Identifying Pathways and Possible Elements

LAURA BRUUN

August 2024

sipri

STOCKHOLM INTERNATIONAL
PEACE RESEARCH INSTITUTE

DOI No: 10.55163/LPED7967

Contents

Acknowledgements

SIPRI and the author express their sincere gratitude to the Swedish, Dutch and Swiss ministries for foreign affairs for their generous financial support for this project.

The author thanks those SIPRI colleagues who contributed to this report: first and foremost Marta Bo, as well as Vincent Boulanin, Alexander Blanchard and Dustin Lewis for their invaluable insights and support throughout the project.

In addition, the author thanks all the experts who generously shared their knowledge and perspectives in the scenario exercise SIPRI organized in Stockholm on 22–23 January 2024: Yasmin Afina, Riham Barghouthi, Tabitha Bonney, Michelle Carr, Delfina Caula, Laurence Degraeve, Alexandra Duca, Netta Goussac, Martijn Hadicke, Bassem Hassan, Georgia Hinds, Joanne Kirkham, Quang Li, Hope Reyes, Rizwan Siddique, Michael Siegrist, Jeroen van den Boogaard, Charles P. Trumbull, Ilse Verdiesen, Essate Weldemichael, Rasmus Weschke and Jie Won Yoon.

The author is also grateful for comments, discussions and feedback provided at different stages of the project by Anna Andersson, Sibylle Bauer, Ingvild Bode, Tabitha Bonney, John Borrie, Damian Copeland, Neil Davison, Netta Goussac, Leo Günzel, Martijn Hadicke, Martin Hagström, Richard Moyes, Anna Nadibaidze, Robert Oostendorp, Jules Palayer, Michael Siegrist, Bart van Breda, Jeroen van den Boogaard and Reto Wollenman. Finally, the author acknowledges the invaluable editorial work of Linda Nix and contributions to the visuals by Frank Esparraga.

Responsibility for the views and information set forth in this report lies entirely with the author.

Summary

After years of discussing how to address the challenges posed by autonomous weapon systems (AWS), there is now growing support among states that the normative and operational framework governing AWS needs to be developed further and that one possible way to proceed is through a ‘two-tiered’ approach. Such an approach would, on the one hand, prohibit certain types and uses of AWS and, on the other hand, place limits and requirements on the development and use of all other AWS. A critical task facing states is to generate agreement about what a two-tiered approach should comprise and how it can be implemented.

To help states elaborate on the contours and content of two-tiered regulation of AWS, SIPRI conducted a scenario exercise based on hypothetical designs and use cases of AWS. The scenario exercise aimed to: (*a*) test and assess the value of using concrete scenarios to advance the policy debate; (*b*) explore in greater detail the spectrum of views around a set of questions that are critical for the identification of possible elements of the two tiers; and (*c*) generate insights into how states could further elaborate on the content and contours of a two-tiered approach to the regulation of AWS. This report presents the key findings and recommendations of this exercise, which can be summarized as follows.

First, the scenario exercise appeared to be a useful format for deepening discussions around limits and requirements in the context of AWS. It allowed participants to bypass the ‘context-dependency problem’ and unpack in greater detail what the concerns are and how they could be addressed. Supplemented with other approaches, hypothetical use cases can further states’ understanding of what limits and requirements are needed to ensure permissible use of AWS, that is, use that satisfies legal, ethical and other policy considerations.

Second, the scenario exercise reflected a broad spectrum of views concerning limits on AWS design and use. While fundamental assumptions around the importance of ensuring international humanitarian law (IHL) compliance and accountability for IHL violations were shared, the limits that flow from these assumptions were widely debated, with notable areas of divergence being types of targets, environments of use and certain data process methods. A key source of divergence stems from the different underlying rationales and concerns that informed participants’ views, from different readings of IHL to broader values rooted in different human rights, ethics, security or other policy priorities. However, while broader concerns, notably those related to ethics, were often used to draw lines, they were rarely explained in depth. To advance discussions, states could usefully elaborate in greater detail the nature of their ethical concerns and explore the role these should play in the policy discussion.

Third, the scenario exercise generated useful insights into how states could develop the content and contours of a two-tiered approach to the regulation of AWS. Participants seemed to agree that targeting decisions ultimately lie with humans, not machines, and therefore that technical characteristics or use cases of AWS that prevent the ability for users to exercise agency and ensure accountability should be prohibited or restricted. This suggests that structuring the two tiers around the importance of the exercise of human agency—understood as the ability to reasonably anticipate and control the behaviour and effects of an AWS—could be a viable way forward. That is, the policy conversation could focus on the specific elements—in the form of prohibitions, limits and requirements—needed to ensure that targeting decisions reflect an exercise of human agency and that the use of force involving AWS can be traced back to the responsible agent(s).

1. Introduction

Advances in military artificial intelligence (AI) are increasingly affecting how wars are fought, including how targeting decisions are made. While the technology is developing rapidly and is already being applied in segments of the targeting process, the global consensus on the regulation of AI in warfare remains unestablished. For more than 10 years, the humanitarian and security challenges posed by autonomous weapon systems (AWS)— a weapon category likely enabled by AI (for a working definition of AWS, see box 1)—have been subject to intergovernmental debate. The discussion—which mainly takes place under the auspices of the Convention on Certain Conventional Weapons (CCW Convention) and is led by a group of governmental experts (GGE)—has centred around whether the challenges posed by AWS warrant the adoption of new regulation, for instance in the form of a new protocol under the CCW Convention.[1] The current mandate of the CCW GGE includes elaborating 'by consensus, possible measures' related to the regulation of AWS.[2]

While states continue to express different views on that question, there is now, at least, growing support for the idea that the normative and operational framework governing AWS needs to be developed further and that one possible way to proceed is through a so-called two-tiered approach.[3] A two-tiered approach is common in arms control and is usually used to refer to prohibitions as one tier and restrictions as the other tier (for a discussion of how a two-tier structure might work in the context of AWS, see box 2). However, while there is progress on agreeing on this basic structure, the content of the two tiers remains debated. A critical task for the GGE is to generate agreement about how a two-tiered approach could be enacted. For example, what, if any, technical characteristics of AWS should be made subject to prohibition? And what specific limits and requirements on AWS use are deemed necessary to ensure the use is permissible?[4] It is also unclear which considerations should be used to inform a two-tiered approach to regulation—in particular, whether the two tiers should be grounded only in international humanitarian law (IHL) or also in other bodies of international law, such as human rights, and whether to consider other perspectives related to, for example, ethics.

To help stakeholders in the international policy debate delineate the contours and content of a two-tiered approach to the regulation of AWS, SIPRI invited experts to participate in an in-person scenario exercise. The exercise, which was conducted under Chatham House Rule, served as a platform for the experts to discuss in greater detail the limits and requirements that would apply in specific use cases. The experts were carefully selected to ensure broad representation in terms of geography, gender, positions and areas of expertise, and participants included military lawyers and

[1] Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons which may be Deemed to be Excessively Injurious or to have Indiscriminate Effects (CCW Convention, or 'Inhumane Weapons' Convention), opened for signature 10 Apr. 1981, entered into force 2 Dec. 1983.

[2] CCW Convention, Group of governmental experts on emerging technologies in the area of lethal autonomous weapon systems (CCW GGE on LAWS), Report of the 2023 session, CCW/GGE.1/2023/2, 24 May 2023, para. 13.

[3] See CCW GGE on LAWS, CCW/GGE.1/2023/2 (note 2). See also working papers and statements submitted to the CCW GGE on LAWS meetings in 2023: 'Draft articles on autonomous weapon systems—prohibitions and other regulatory measures on the basis of international humanitarian law ("IHL")', Working paper submitted by Australia, Canada, Japan, South Korea, the United Kingdom and the United States, CCW/GGE.1/2023/WP.4, 6 Mar. 2023; 'Draft protocol on autonomous weapon systems (Protocol VI)', Working paper submitted by Argentina, Colombia, Costa Rica, Ecuador, El Salvador, Guatemala, Kazakhstan, Nigeria, Palestine, Panama, Peru, Philippines, Sierra Leone and Uruguay, CCW/GGE.1/2023/WP.6, 10 May 2023); 'Translating the progress at the GGE LAWS into a substantive outcome', Joint statement delivered by Germany and the Philippines on behalf of 51 states, 15 May 2023.

[4] This report refers to 'limits' throughout as shorthand to refer to possible elements in a two-tiered approach without prejudging whether such limits should be formulated as prohibitions or restrictions on design or use.

Box 1. A working definition of autonomous weapon systems

Autonomous weapon systems (AWS) are commonly understood as weapons that, once activated, can identify, select and apply force to targets without human intervention. Certain distinctive socio-technical characteristics distinguish AWS from other means and methods of warfare:

- AWS function based on pre-programmed target profiles and technical indicators that AWS can 'recognize' through their sensors and software.
- Since AWS are triggered to apply force partly by the environment of use (rather than a user's input), aspects of a decision to apply force can be made further in advance than with traditional weapons.
- A human operator may supervise and retain the possibility of overriding an AWS, but the system's default functioning is that human input is not required to identify and select targets, nor to apply force against them.

These characteristics entail that those who configure and deploy an AWS will not necessarily know the exact targets, location, timing or circumstances of the resulting use of force.

The term 'AWS' is preferred to the term 'LAWS' (i.e. lethal AWS) based on the reasoning that the concept of lethality is not needed for the analysis in this report. This is both because 'lethal' pertains to how the weapon system is used and its effects rather than the way it is designed, and because AWS are still capable of causing harm in the form of material damage or injury.

Sources: Boulanin, V. and Verbruggen, M., *Mapping the Development of Autonomy in Weapon Systems* (SIPRI: Stockholm, Nov. 2017), pp. 24–27; and the International Committee of the Red Cross (ICRC), 'ICRC position on autonomous weapon systems', 12 May 2021.

policy advisers from governments across six continents, as well as experts from civil society, academia and the International Committee of the Red Cross (ICRC).

This report presents the key takeaways from the scenario exercise. Its general aim is to help generate a deeper understanding of how to structure and frame a two-tiered approach to the regulation of AWS. More specifically, it aims to provide stakeholders in the policy debate on AWS with insights into the following aspects: (*a*) the value of using concrete scenarios to advance the policy debate around the two tiers (chapter 2); (*b*) the spectrum of views and rationales around a set of questions deemed critical for identifying possible elements of the two tiers (chapter 3), (*c*) areas of convergence on the possible elements of the two tiers (chapter 4); and (*d*) key findings and recommendations as to how to advance the policy conversation on the governance of AWS (chapter 5). The specific scenarios and assessment templates used in the scenario exercise are attached as an annex to this report.

Box 2. Possible ways to structure a two-tiered approach in the context of autonomous weapon systems

While autonomous weapon systems (AWS) may be relatively new to arms control, a two-tiered approach to regulation is not. Several existing arms control agreements reflect this structure, with a mix of outright prohibitions and restrictions.[a] However, in the context of AWS, it is not yet settled among states how each of the two tiers should be structured and interrelate. Current proposals include, among others:

- Tier 1 contains prohibitions and limits. Tier 2 contains positive requirements.
- Tier 1 contains limits and requirements in the development phase. Tier 2 contains limits and requirements in the use phase.
- Tier 1 contains prohibitions. Tier 2 contains limits and requirements.
- Tier 1 is a prohibition against AWS that are 'fully' autonomous. Tier 2 contains limits and requirements on AWS that are 'partially' autonomous.

This report adopts the two-tiered approach in the third proposal above. That is, tier 1 prohibits types and uses of AWS that would be unlawful in all circumstances, while tier 2 places limits and requirements on the development and use of all other AWS. This is not intended to prejudge the final structure of a potential two-tiered approach to AWS but to generate a deeper understanding of what elements should be considered off-limits, what elements should be subject to restrictions, and what, if any, positive obligations apply, regardless of the form these elements may take in an instrument.

[a] E.g., the 1993 Convention on the Prohibition of the Development, Production, Stockpiling and Use of Chemical Weapons and their Destruction (Chemical Weapons Convention, CWC Convention) and 3 of the 5 protocols under the 1981 Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons which may be Deemed to be Excessively Injurious or to have Indiscriminate Effects (CCW Convention), namely Amended Protocol II on Mines, Booby-traps and Other Devices, Protocol III on Incendiary Weapons and Protocol IV on Blinding Laser Weapons.

2. Using scenarios as a method to identify limits and requirements for the development and use of AWS

This chapter explains how scenario exercises can help states elaborate the limits and requirements that should be reflected in a two-tiered aproach to regulating AWS. It covers how the specific scenarios used in the SIPRI exercise were chosen and designed, how the discussion was conducted, and the limitations of scenario exercises, both as a general methodology and specific to the SIPRI exercise.

Reasons to use a scenario exercise

Scenario exercises are a well-known methodology within and beyond the context of AWS.[5] In the specific context of elaborating on the content and contours of a two-tiered approach to regulating AWS, there are several reasons why a scenario exercise is a useful approach.

First, it can help *ground the discussion*. When stakeholders involved in the AWS debate think of AWS, they usually have different systems and use cases in mind. Scenarios serve as a concrete basis on which to formulate the concerns that stakeholders associate with AWS, and provide a way to bypass circular debates around definitions and to avoid possible misunderstandings of what the discussion is about.

Second, scenarios can help address the so-called *context-dependency problem*. A critical finding from SIPRI's past research on AWS is that states usually find it difficult to identify limits and requirements in the abstract because these depend heavily on context, such as the characteristics of the systems and the environment of use.[6] Limits and requirements identified through discussion of specific scenarios can, arguably, be generalized and used to help identify wider categories of limits and requirements relevant across multiple contexts.

Third, scenarios can help *clarify the core concerns* and rationales underpinning the various positions, particularly in the context of the current debate about which specific considerations should inform the identification of limits and requirements. The main focus of this debate is whether a two-tiered regulatory framework should be grounded only in IHL or also in other bodies of international law, such as international human rights law, as well as other perspectives related to ethical and policy considerations. Scenarios offer a platform for states to present and compare in more detail what limits and requirements they see flowing from IHL as well as from other legal, ethical, policy and operational considerations.

Factors considered in designing the scenarios

For scenarios to be useful, they must be designed to expose aspects that deserve particular attention and generate discussions that build on—rather than replicate—existing debates around the two tiers. Accordingly, the scenarios designed for this project were

[5] Past scenarios exercises on AWS have been conducted by, among others, SIPRI, the International Committee of the Red Cross (ICRC) and the United Nations Institute for Disarmament Research (UNIDIR). See Boulanin, V. et al., *Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control* (SIPRI/ICRC: Stockholm, June 2020); and Persi Paoli, G., Spazian, A. and Anand, A., *Table-top Exercises on the Human Element and Autonomous Weapons System: Summary Report* (UNIDIR: Geneva, 2021).

[6] Boulanin, V., Bruun, L. and Goussac, N., *Autonomous Weapon Systems and International Humanitarian Law: Identifying Limits and the Required Type and Degree of Human–Machine Interaction* (SIPRI: Stockholm, June 2021), p. 54; and Bruun, L., Bo, M. and Goussac, N., *Compliance with International Humanitarian Law in the Development and Use of Autonomous Weapon Systems: What Does IHL Permit, Prohibit and Require?* (SIPRI: Stockholm, Mar. 2023), p. 24.

based on a review of the current state of the policy discussion on AWS, especially within the CCW. This review revealed two main contentions that are undisputed among states: (*a*) the development and use of AWS must comply with international law, including IHL; and (*b*) AWS that cannot be used within a responsible chain of command and control must be prohibited.[7] Another area of common ground is that, to ensure compliance with IHL, operational constraints may be needed on the development and use of AWS, including limits placed on targets, duration and geographical scope of attack.[8]

However, besides these shared assumptions (which mostly restate existing IHL obligations), a review of the policy debate exposed several threshold questions that must be addressed before states can make progress on the content and contours of a two-tiered approach to regulation. These questions can be summarized as follows: (*a*) What objects and which people should never be made the subject of an attack involving AWS? (*b*) Can AWS ever be used in environments where civilians are present? (*c*) How predictable, controllable and traceable should AWS be in terms of design and use?[9]

The scenarios in the exercise were deliberately designed to support states in elaborating their views on these threshold questions. One scenario revolved around the design and use of a hypothetical anti-armour loitering munition ('Alpha-100') and the other around the design and use of a hypothetical anti-personnel loitering munition ('Beta-100'). Both scenarios included variations, all designed to reflect different permutations of realistic technical characteristics, operational environments and human involvement, with a view to sparking debate around off-limits technical characteristics and uses, specific limits on use, and requirements on users. (The term 'users' in this report is understood in a broad sense, involving the number of actors involved in the development and use of AWS.[10]) A detailed description of the two scenarios is contained in the annex to this report. A high-level overview of each scenario is shown in figures 1 and 2.

Guiding the scenario exercise

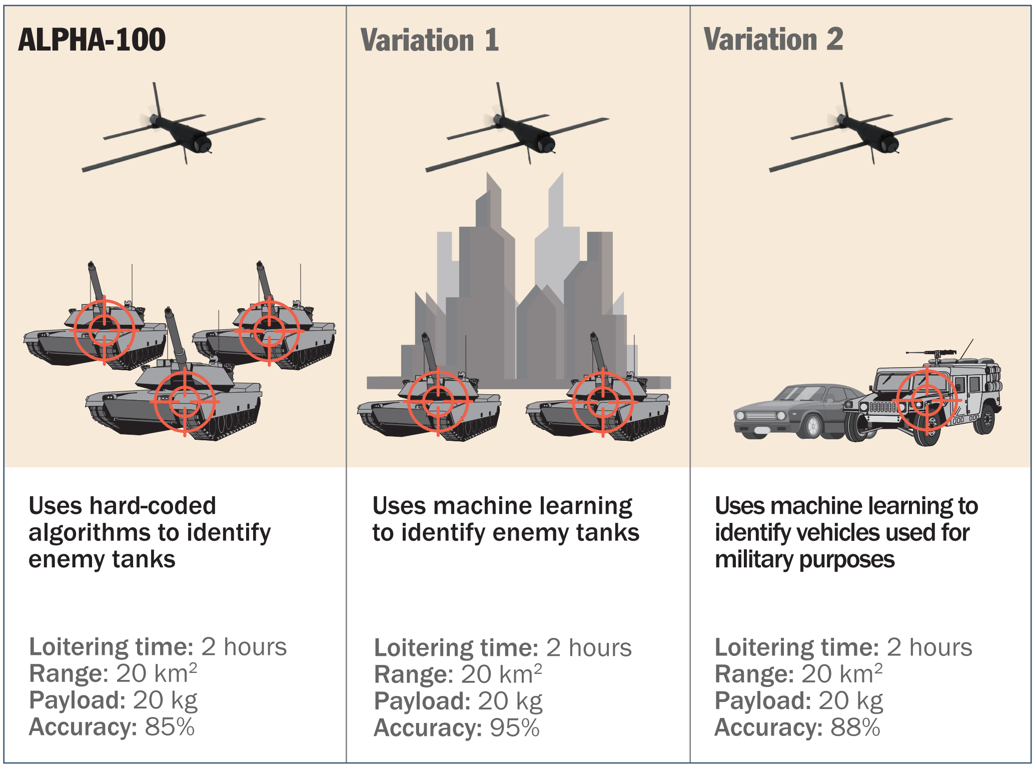
Scenario exercises can be conducted in multiple ways. For this project, the SIPRI team designed the exercise in a way that mirrored the two tiers. That is, each scenario was split into two parts that generated two respective phases of discussions focused on identifying first prohibitions and then restrictions. The first phase was intended to simulate a weapons review process (rather than a strict legal review, so that a broader set of perspectives could inform the assessments) in which participants reviewed the system to identify inherently off-limits technical characteristics and uses. In the second phase, participants were invited to discuss the types of control measures (in the form

[7] This includes the prohibition of the development and use of AWS which: (*a*) are of a nature to cause superfluous injury or unnecessary suffering, (*b*) are by nature indiscriminate, (*c*) cannot be used in compliance with the principles of distinction, proportionality and precautions in attack, and (*d*) cannot be directed at a specific military objective. These proposals are reflected in most national positions: see e.g. the Netherlands, 'Possible answers to the guiding questions of the Chair of the GGE on LAWS', Document submitted to the 2021 session of the GGE CCW on LAWS, Sep. 2021; CCW GGE on LAWS, CCW/GGE.1/2023/2 (note 2); and working papers submitted to the 2023 session of the CCW GGE on LAWS: Australia et al., CCW/GGE.1/2023/WP.4 (note 3) Argentina et al., CCW/GGE.1/2023/WP.6 (note 3); and 'Objectives and purposes of the conventions: proposal for an international legal instrument on LAWS', Working paper submitted by Pakistan, CCW/GGE.1/2023/WP.3, 6 Mar. 2023.

[8] CCW GGE on LAWS, CCW/GGE.1/2023/2 (note 2).

[9] 'Traceability' can be understood as referring to the ability to discern, scrutinize and attribute responsibility for violations of IHL. See Bo, M., Bruun, L. and Boulanin, V., *Retaining Human Responsibility in the Development and Use of Autonomous Weapon Systems: On Accountability for Violations of International Humanitarian Law Involving AWS* (SIPRI: Stockholm, Oct. 2022), p. 41.

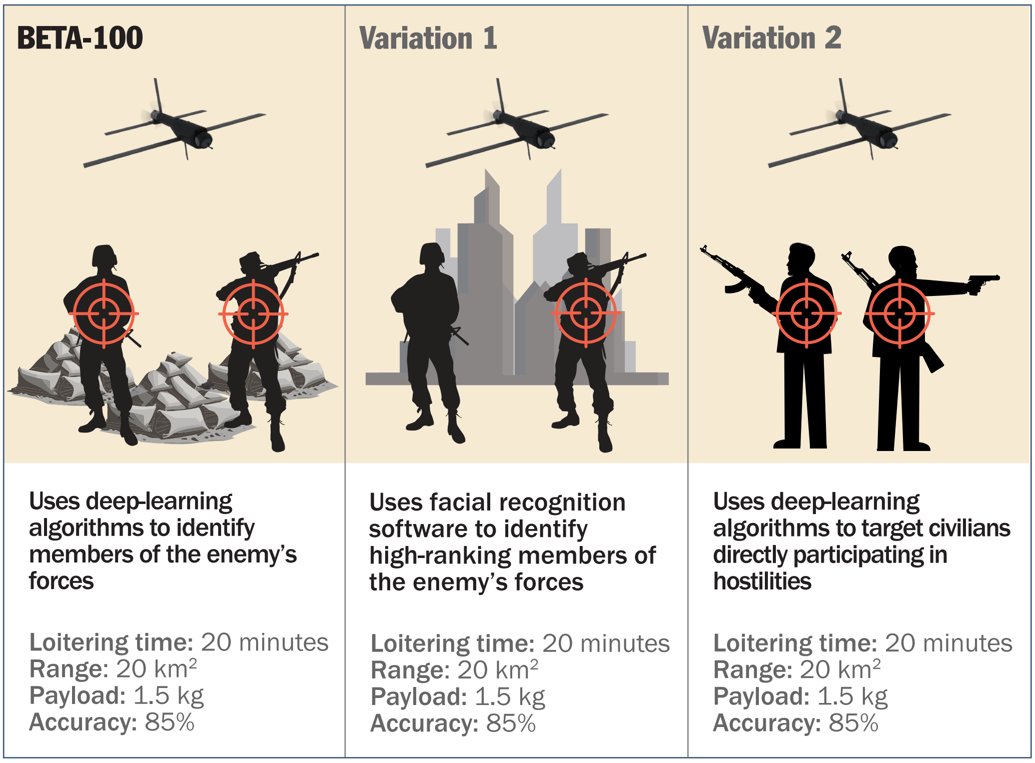
[10] This includes those acting on behalf of states and parties to an armed conflict and those who plan, decide on or carry out military action involving an AWS. It encompasses 'operators' and 'commanders' and the many indivdiduals involved in other stages, such as those who create the algorithmic parameters, those who prepare, select and label the data required by machine-learning algorithms, and those who test and train them.



ALPHA-100	Variation 1	Variation 2
Uses hard-coded algorithms to identify enemy tanks	Uses machine learning to identify enemy tanks	Uses machine learning to identify vehicles used for military purposes
Loitering time: 2 hours **Range:** 20 km^2 **Payload:** 20 kg **Accuracy:** 85%	**Loitering time:** 2 hours **Range:** 20 km^2 **Payload:** 20 kg **Accuracy:** 95%	**Loitering time:** 2 hours **Range:** 20 km^2 **Payload:** 20 kg **Accuracy:** 88%

Figure 1. An overview of the Alpha-100 variations used in the SIPRI scenario exercise

Credit: Vector images from Vecteezy.com.



BETA-100	Variation 1	Variation 2
Uses deep-learning algorithms to identify members of the enemy's forces	Uses facial recognition software to identify high-ranking members of the enemy's forces	Uses deep-learning algorithms to target civilians directly participating in hostilities
Loitering time: 20 minutes **Range:** 20 km^2 **Payload:** 1.5 kg **Accuracy:** 85%	**Loitering time:** 20 minutes **Range:** 20 km^2 **Payload:** 1.5 kg **Accuracy:** 85%	**Loitering time:** 20 minutes **Range:** 20 km^2 **Payload:** 1.5 kg **Accuracy:** 85%

Figure 2. An overview of the Beta-100 variations used in the SIPRI scenario exercise

Credit: Vector images from Vecteezy.com of limits and requirements) they would deem necessary in light of specific operational situations. The participants were provided with an assessment template for each part of both scenarios (see annex).

Potential limitations of scenario exercises

When engaging in a scenario exercise, it is important to be aware of the potential drawbacks and limitations associated with this methodology. Three such aspects became evident in the SIPRI scenario exercise. First, preparing to take part in a scenario exercise demands both time and personnel, presenting a hurdle for states to contribute on an equal footing. Consequently, there is a risk that discussions may be skewed towards those who possess the resources to prepare for exercises more thoroughly (a risk not confined to scenario exercises; it is also apparent in the policy discussions concerning AWS). Second, there is a risk that scenarios could be instrumentalized and used to normalize the use of AWS. That is, prompting discussions on limits and restrictions can be perceived as permitting certain systems and use cases that, according to some, should not exist at all. Third, there is a risk that the aspects that the scenarios seek to explore (e.g. legal, ethical, operational) might not be reflected in the composition of participants. For example, if the majority of participants are IHL lawyers, as was the case with the SIPRI scenario exercise, the focus will likely be on legal concerns, while other aspects, such as ethical or human rights perspectives, recede into the background, because the expertise to address them is potentially lacking. With these limitations in mind, the scenario exercise did generate lengthy, nuanced discussions around some of the most contested aspects of the AWS debate. Key takeaways from the discussions are presented in the following chapter.

3. Mapping the spectrum of views on limits and requirements in the design and use of AWS

While states are increasingly supporting a two-tiered approach to the regulation of AWS, they express different views as to the structure and contents of these tiers. The differences revolve around three threshold questions in particular: (*a*) which objects and people should never be made the subject of attack involving AWS; (*b*) whether AWS should ever be used in environments where civilians are present; and (*c*) how predictable, controllable and traceable should AWS be in terms of the design and use? As described in chapter 2, these questions informed the design of the scenario exercise, to help advance discussions around the possible elements of the two tiers. This chapter maps in greater detail the spectrum of views around these limits and requirements revealed in the scenario exercise, while also reflecting on their underlying rationales and concerns.

Which objects and people should never be made the subject of attack involving AWS?

States agree that to ensure IHL compliance in the context of AWS, especially with the principle of distinction, it is important, when 'necessary', to 'Limit the types of targets that the system can engage'.[11] This can be understood as a reflection of existing IHL, which already limits certain types of targets by prohibiting direct attacks against civilians, civilian objects and combatants *hors de combat* (see box 3).[12] However, states are discussing whether additional specific limits on types of targets are needed in the context of AWS.[13] A main dividing line is whether AWS can *ever* be used to direct attacks against persons, or whether their use should be limited to attacks against objects, with some states and the ICRC further arguing that AWS should only be used in attacks against objectives that are military by nature.[14] To advance discussions in this respect, both scenarios were designed around a variety of target types, from enemy tanks and pick-up trucks used for military purposes, to combatants and civilians directly participating in hostilities (DPH). In summary, most participants agreed on the outer limits: AWS used in attacks against military objectives by nature is the least problematic case, while AWS should in principle never be used to target civilians DPH. However, whether AWS could ever be used against other types of military objectives or against persons at all divided the room, bringing ethical and policy concerns associated with the use of AWS to the forefront. A more detailed overview of the spectrum of views captured during the scenario exercise, and their underlying rationales, is set out in the following sections.

[11] CCW GGE on LAWS, CCW/GGE.1/2023/2 (note 2), para. 22(a).

[12] Protocol Additional to the 1949 Geneva Conventions, and relating to the Protection of Victims of International Armed Conflicts (AP I), opened for signature 12 Dec. 1977, entered into force 7 Dec. 1978, Arts 41 and 48.

[13] See e.g. CCW GGE on LAWS, 'Principles and good practices on emerging technologies in the area of LAWS', Proposal submitted by Australia, Canada, Japan, Korea, the UK and the USA, Mar. 2022; and 'Roadmap towards new protocol on AWS', Proposal submitted by Argentina, Costa Rica, Guatemala, Kazakhstan, Nigeria, Panama, Philippines, Sierra Leone, Palestine and Uruguay, Mar. 2022.

[14] Among others, Palestine, Austria and the ICRC argue that AWS must never be used in attacks against humans: CCW GGE on LAWS, 'State of Palestine's proposal for the normative and operational framework on autonomous weapon systems', Working paper submitted by Palestine, CCW/GGE.1/2023/WP.2, 3 Mar. 2023; Revised working paper submitted by Austria, CCW/GGE.1/2023/WP.1, 3 Mar. 2023; and ICRC, 'ICRC position on autonomous weapon systems', 12 May 2021.

Box 3. Objects and persons against which attacks may be directed under international humanitarian law

International humanitarian law (IHL) prohibits making civilians, civilian objects and combatants *hors de combat* the object of an attacks. The categories of objects and people that may lawfully be object of an attack under IHL include the following:

Objects

IHL distinguishes four types of objects that constitute lawful targets insofar as (*a*) they make an effective contribution to military action and (*b*) their total or partial destruction, capture or neutralization, in the circumstances ruling at the time, offers a definite military advantage.[a] These are:

- **military objectives by nature**, defined as an object of intrinsic military character, such as missiles, highly explosive warheads and other weapons, military equipment and military headquarters
- **military objectives by location**, defined as a specific area of tactical importance such as mountain passes, bridgeheads, footbridges, crossroads, hills and tunnels
- **military objectives by purpose**, defined as an object which is not inherently military but whose use could make an effective contribution to the military action if there is evidence that a party to a conflict intends to use it for military action
- **military objectives by use**, defined as an object which is not inherently military but which becomes a lawful target when used to support the military action, such as bridges and vehicles.

Persons

IHL also distinguishes between different categories of people who may be made the object of attack, including combatants, civilians directly participating in hostilities (DPH) and members of organized armed groups.

- **Combatants**: In international armed conflicts, combatants are members of all organized armed forces, groups or units under a command responsible to a state party to the conflict. Medical and religious personnel assuming exclusively humanitarian functions are excepted from the definition. The only weapon-bearers who may be regarded as combatants without being members of the armed forces are participants in a *levée en masse* (mass conscription).[b]
- **Civilians DPH**: If and for as long as civilians are directly participating in hostilities, they lose their protection against attack and become targetable.[c] The criteria amounting to DPH are debated but have been interpreted to cover both direct participation in combat and also active participation in other military activities, such as scouting, spying and sabotage.[d]
- **Members of organized armed groups**: In non-international armed conflicts, members of organized armed groups that constitute the organized fighting forces of a non-state party to a conflict are targetable.[e]

[a] Protocol Additional to the 1949 Geneva Conventions, and relating to the Protection of Victims of International Armed Conflicts (AP I), opened for signature 12 Dec. 1977, entered into force 7 Dec. 1978, Art. 52(2).

[b] AP I, Art 43(1) and (2); and International Committee of the Red Cross (ICRC), 'Rules', Customary IHL Database, [n.d.], Rule 3.

[c] AP I, Art. 51(3); and Protocol Additional to the 1949 Geneva Conventions, and relating to the Protection of Victims of Non-International Armed Conflicts (AP II), opened for signature 8 June 1977, entered into force 7 Dec. 1978, Art. 13(3). See also Sassòli, M. et al., 'Direct participation in hostilities', *How Does Law Protect in War?*, Online Casebook (ICRC: Geneva, 2014).

[d] Melzer, N., *Interpretive Guidance on the Notion of Direct Participation in Hostilities under International Humanitarian Law* (ICRC: Geneva, 2009); and Schmitt, M., 'The interpretive guidance on the notion of direct participation in hostilities: a critical analysis', *Harvard National Security Journal*, vol. 1 (2010).

[e] Melzer, N., *International Humanitarian Law: A Comprehensive Introduction* (ICRC: Geneva, 2022), pp. 126 and 127.

Targeting of objects

To advance the discussion around what limits, if any, should be placed on objects made subject to attack by AWS, participants were asked to assess the permissibility of designing and using loitering munitions to direct attacks against enemy tanks and pick-up trucks used for military purposes, representing examples of military objectives by nature and use, respectively.

Military objective by nature. Most participants tolerated the scenario involving a loitering munition designed to direct attacks against enemy tanks in communication-denied environments (i.e. with no option for direct supervision or intervention once the system is activated), free of civilians and with a loitering time of two hours (Alpha-100).[15] A main rationale was that the target's status as a lawful military target could be presumed to remain valid throughout the attack, and thus allow users to make IHL-mandated decisions, notably around distinction, a longer time in advance. However, participants disagreed about the need to further differentiate between different *types* of military objectives by nature that may be attacked with AWS. The disagreement arose mainly from concerns around complying with the principle of distinction, where some argued that to reduce the risk of misidentification and attacking protected persons, AWS should only be used in attacks against military objectives by nature if they are: (*a*) large in size; (*b*) not moving; and (*c*) presenting a significant military advantage. In addition, a few argued on the basis of ethical concerns that using AWS to target objects inhabited by persons, such as crewed tanks, is problematic and potentially off-limits. This view flows from the assumption that for an AWS to inflict harm on people, even indirectly, is 'dehumanizing' and should thus be avoided.

Military objective by use. When presented with the scenario of a loitering munition designed to direct attacks against pick-up trucks used for military purposes in communication-denied environments where the presence of civilians could not be ruled out (Alpha-100, variation 2), all participants called for the need to place significantly stricter control measures on the use of AWS. However, whether the use of AWS against objects whose status as lawful targets depends on the context (as in the scenario) should be categorically ruled out divided participants: more than half argued that AWS should *never* be designed as capable of and used for directing attacks against military objectives by use—or location or purpose for that matter. This view was rooted in the particular concern that the risk of misidentifying targets is too high when directing attacks against military objectives whose status as lawful targets may change after activation.[16] Some also argued that the risk of misidentification was too high from a technical perspective, questioning the ability to develop a system that could sufficiently distinguish between a mounted turret and a similarly shaped civilian object. Others argued against a categorical prohibition because the permissibility of using AWS in attacks against military objectives by use necessarily depends on the context. These participants argued that the use of AWS to direct attacks against military objectives by use (and location, as mentioned by some) can be considered permissible, and even in some cases preferable, for both legal and operational reasons, depending on: (*a*) the target's location; (*b*) its expected military advantage; and (*c*) what the alternative to using an AWS is. They also argued that concerns related to IHL compliance—notably whether assessments man-

[15] The potential to carry out attacks in communication-denied environments is one of the key military advantages associated with AWS because it makes operations resistant to jamming and enables attacks in environments that are too dangerous for troops to otherwise enter. However, the use of AWS in communication-denied environments raises questions about what is required in terms of control during use, an aspect explored later in this chapter.

[16] CCW GGE on LAWS, Pakistan, CCW/GGE.1/2023/WP.3 (note 7); Palestine, CCW/GGE.1/2023/WP.2 (note 14); and ICRC, 'ICRC position on autonomous weapon systems' (note 14).

dated by the principle of distinction would remain valid after AWS activation—could be addressed by placing strict limits on the environment, duration, scale and scope of the attack. To this end, some mentioned, by way of example, that AWS should never be used to target military objects by use in environments where civilians and civilian objects may be present.

Targeting of persons

To explore views around whether AWS can ever be used to target persons (and if so, what kind of persons), participants were invited to discuss the permissibility of anti-personnel loitering munitions designed for and used in attacks against combatants hiding in trenches and against civilians directly participating in hostilities (DPH), among other possible use cases.

Combatants. Participants assessed the permissibility of using loitering munitions to target combatants hiding in trenches located in communication-denied environments where it is not possible or feasible to deploy remote-controlled systems (Beta-100). The scenario, involving a loitering munition, which could loiter for up to 20 minutes without human intervention, divided participants equally as to whether its use was permissible, with both sides invoking different legal, ethical and policy reasons for their view.

First, from a legal perspective, particularly that of compliance with the principle of distinction, participants disagreed about the permissibility of this anti-personnel system. Those who argued that the system was impermissible argued that even if an AWS is deployed in an observable environment free of civilians, the risk of targeting combatants *hors de combat* (such as those who are wounded or who are surrendering and which are protected under IHL) is too high, and for that reason alone such systems should never be used. It was further argued that it would even not be possible, from a technical perspective, to translate *hors de combat* identifiers into data points because gestures of surrender and discernible signs of being wounded or sick are considered extremely context-dependent. Others acknowledged this concern but reminded the group that it is not the AWS that assesses *hors de combat* status, but the users, upon activation of the AWS. This group argued that, based on existing practice, there is no requirement to recognize and assess *hors de combat* status until the very moment of force application, and that some time lag is permissible, arguably allowing the use of AWS in communication-denied environments. However, these participants stressed that very strict control measures, notably limits on the geographical and temporal scope of the attack, would be needed to ensure compliance with IHL in such situations.

Several participants argued that even if there is no strong basis in IHL to prohibit anti-personnel AWS, they would not use such systems for reasons related to ethics, human rights and broader policy concerns. For example, taking the point of departure in concerns around those affected by the use of AWS, some argued that reducing a human to data points is ‘dehumanizing’ and contravenes principles to preserve human dignity. Others argued that this issue goes beyond IHL as it touches on the relationship between humans and machines in general, forcing policymakers to think deeply about how much ‘power’ they want to give machines. From a broader policy and security perspective, a handful of participants cautioned against permitting the design and use of anti-personnel AWS because of risks related to proliferation and escalation of violence, and the risk that they would be used by malicious actors.

Civilians DPH. The final scenario presented to participants revolved around the design and use of a loitering munition to direct attacks against civilians DPH (Beta-100, variation 2). While the criteria for ‘direct participation in hostilities’ are generally debated, participants assessed the use of AWS to target civilians based on their intent to use force. In the scenario, an individual's 'intent' would be determined by whether the person carried a weapon and whether their body language indicated an 'offensive posture'. Here, all participants agreed that this scenario was unacceptable. Most pointed out that, from an IHL perspective, assessing whether a civilian is DPH is extremely context-based and that real-time assessments are difficult to make, even for humans. Many participants also argued that from a technical perspective it would be difficult, if at all possible, to translate DPH status into data points. A few noted that their assessment might change if the targeting criteria were based on direct actions (such as active fire engagement), arguing that using AWS to target civilians DPH could potentially be used lawfully, if the environment of use was sufficiently predictable and the technology was advanced enough. However, almost all maintained that regardless of the context and targeting criteria, AWS should never be designed and used to engage civilians DPH.

Can AWS ever be used in environments where civilians are present?

Another question that divides states is whether the use of AWS should be limited to environments where civilians are not present. Several states, such as Costa Rica, Palestine and Pakistan, as well as the ICRC, share a concern that the presence of civilians in the area of operations of an AWS may unintentionally trigger the application of force by the AWS.[17] Also, the ability to make valid assessments about the circumstances ruling at the time, including expected incidental harm (necessary for compliance with the principle of proportionality), may be more difficult to make in advance in areas characterized by civilian movement. At the same time, other states, such as the United States, have argued that imposing such a restriction could prove 'counterproductive from a humanitarian perspective'; for example, when defending against an attack in a populated area, use of an AWS with increased accuracy and reduced payload potentially poses fewer risks to civilians and civilian objects than non-autonomous weapons.[18] To build on existing discussions, participants were asked to consider to what extent their assessments changed if civilian presence was added to the scenarios (Alpha-100, variation 1 and 2; Beta-100, variation 1).

Overall, the presence of civilians (or the mere possibility thereof) was extremely relevant to all participants' assessments: when the scenarios were unfolding in environments where the presence of civilians 'could not be ruled out', the need for control measures significantly increased. But whether the use of AWS in areas where civilians are present should be ruled out altogether divided participants. Those who argued that AWS should *never* be used in environments where civilians are present expressed the same types of concerns around the ability to comply with the principle of proportionality as already expressed by Costa Rica, Palestine, Pakistan, the ICRC and others. Many argued that the use of AWS in populated environments could only be considered permissible if the system is operated with real-time supervision and human intervention (which, however, would take it outside the category of AWS as defined by SIPRI for this scenario exercise). Moreover, some argued that even if there are circumstances where the use of AWS in populated environments can, in theory, be used

[17] For example, the ICRC warns that civilian objects like cars or buses could trigger an AWS programmed to recognize and attack military vehicles or personnel carriers. ICRC, 'ICRC position on autonomous weapon systems' (note 14). See also CCW GGE on LAWS, Written contribution for the Chair of the 2021 meeting of the GGE on LAWS, Submitted by Argentina, Costa Rica, Ecuador, El Salvador, Panamá, Palestine, Perú, the Philippines, Sierra Leone, and Uruguay, Sep. 2021; and CCW GGE on LAWS, Working papers submitted at the 2023 meeting by Palestine, CCW/GGE.1/2023/WP.2 (note 14) and Pakistan, CCW/GGE.1/2023/WP.3 (note 7).

[18] United States, Statement delivered at the First Session of the CCW GGE on LAWS, Geneva, 8 Mar. 2023, UN WebTV, 00:22:43–00:26:15.

in compliance with IHL, the general risk of inflicting excessive harm outweighs such narrow circumstances. These participants suggested that, for humanitarian reasons, using AWS in populated areas should be avoided altogether.

Other participants argued that the use of AWS in areas where civilians are present is not off-limits per se. They suggested that the challenges posed by environments containing civilians can be accounted for through measures other than prohibition, for example by placing stricter limits on the duration and geographical scope of the attack. It was also argued that the use could be allowed on the condition that the target is a high-value target and thus associated with significant military advantage. The underlying reasoning is that in those cases, users can lawfully accept higher risks of civilian harm because of the high value of the target and the expected military advantage associated with its destruction or neutralization (in compliance with the principle of proportionality).

Despite the different views on where to draw the lines, all concerns seemed to stem from the same source, namely whether environments where civilians may be present would allow AWS users to possess *sufficient situational awareness* in relation to the use of the AWS. In this context, situational awareness was understood as referring to users' understanding and knowledge about the environment of use, including knowledge about possible civilian movements. It appeared undisputed that for legal and operational reasons, users should be reasonably certain that the environment is not materially changing after AWS activation to make sure that their assessments remain valid. But while everyone agreed on the importance of ensuring situational awareness, participants disagreed on whether that categorically rules out the use of AWS in environments where civilians are present.

How predictable, controllable and traceable should AWS be in terms of design and use?

When using AWS, users do not necessarily know when, where, or against whom or what force will be applied.[19] These characteristics (among others) have generated debate about what is legally (or otherwise) required in terms of predictability, controllability and traceability of AWS and to what extent, if any, such requirements should be used to identify off-limits technical characteristics or use cases. Through different variables, the scenarios tested views around the limits that should be placed on technical characteristics or use cases on the basis of predictability, controllability and traceability. The spectrum of views that emerged during these discussions is captured below.

Unpredictability as the basis for prohibiting use of an AWS?

Several states, including the USA, Japan, Austria and Palestine, as well as the ICRC, have expressed the view that AWS whose effects cannot reasonably be anticipated are prohibited because they cannot be used in compliance with IHL.[20] However, what that means in practice—that is, what limits to place on the design and use of AWS—is unclear. For example, Palestine and the ICRC, argue that AWS whose data processing is based on machine learning (ML) should be 'avoided' due to the unpredictability introduced by such processes, but this view remains contested.[21] To provide more details to the

[19] ICRC, 'ICRC position on autonomous weapon systems' (note 14).

[20] See e.g. CCW GGE on LAWS, 2023 meeting, Working papers submitted by Australia et al., CCW/GGE.1/2023/WP.4 (note 3), and Palestine, CCW/GGE.1/2023/WP.2 (note 14), and Written contribution submitted by Argentina et al. (note 17); CCW GGE on LAWS, 2021 meeting, Document submitted by the Netherlands (note 7); and ICRC, 'ICRC position on autonomous weapon systems' (note 14).

[21] CCW GGE on LAWS, Palestine, CCW/GGE.1/2023/WP.2 (note 14), p. 10; and ICRC, 'ICRC position on autonomous weapon systems' (note 14).

Box 4. Machine learning and predictability

Machine learning (ML) is an approach to software development that consists of building a system that can learn and then teaching it what to do using a variety of methods (e.g. supervised learning, reinforcement learning or unsupervised learning). ML removes the need for hand-coded programming (i.e. humans hard-coding software features into the systems) and creates opportunities for improvement in many military applications, from navigation to target recognition.

While ML has the potential to improve system performance by optimizing behaviour through environmental interactions, these improvements can lead to the system generating unforeseen behaviours, diminishing system predictability. ML systems operate like 'black box' systems, where the inputs and outputs are observable but the process leading from input to output is unknown or difficult to understand, which potentially also impacts system predictability. Unless the system's learning algorithm is 'frozen' at the end of the training phase (which currently is the case with most known examples of ML used in the military), once deployed, it might 'learn' something it is not intended to learn or do something that humans do not want it to do. This is what is usually referred to as 'online learning'.

Sources: Boulanin, V., 'Artificial intelligence: A primer', V. Boulanin et al., *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk—Volume I: Euro-Atlantic Perspectives* (SIPRI: Stockholm, May 2019), pp. 15 and 19; and Boulanin, V. and Verbruggen, M., *Mapping the Development of Autonomy in Weapon Systems* (SIPRI: Stockholm, Nov. 2017), p. 16.

debate, the scenarios tested views around when, if ever, an AWS if off-limits on the basis of unpredictability through two variables: targeting accuracy (in terms of target recognition) and use of ML-based data processing for target recognition (see box 4).

Accuracy rates. Participants were asked to assess the permissibility of loitering munitions with different accuracy rates, for example an anti-tank loitering munition described as having an 85% accuracy 'when deployed in uncluttered, observable environments and under clear weather conditions' (Alpha-100) and an anti-vehicle loitering munition, intended to be used against civilian pick-up trucks used for military purposes, with an accuracy rate of 88% (Alpha-100, variation 1). Many participants said that an 85% accuracy rate was too low regardless of the context and considered it an off-limits characteristic. Others took a less categorical approach, arguing that it could be used permissibly, depending on the target, environment, other means available and information about how it would fail *if* it failed. For example, some said that a 15% fail rate would be permissible if such AWS are only used in situations where, in case of failure, the system would just hit the ground and not inflict excessive harm on protected persons or objects.

The main source of divergence among participants seemed to stem from different views about the extent to which strict operational constraints could compensate for the unpredictability associated with 'low' accuracy rates, for example limiting use to environments where no civilians are present and limiting targets to military objectives by nature. However, it was common ground that AWS with a 50% or lower accuracy rate would likely make it inherently indiscriminate in all circumstances.

Machine learning (ML). To further test when, if ever, certain data process methods would make an AWS off-limits, participants were asked to first assess the permissibility of an anti-tank loitering munition that used rule-based programming for target recognition (Alpha-100), and then whether their assessment changed if the system's target recognition capabilities had been updated, now developed with ML in the training phase (Alpha-100, variation 1). While the update meant that the system's accuracy rate had increased from 85% to 95%, it came with a trade-off in terms of predictability (see box 4).

According to about half of the participants, the mere use of ML-based processing was considered an off-limits technical characteristic even if the system's accuracy was enhanced. It was argued that such complex processes undermine users' ability to understand how an AWS functions and consequently to sufficiently anticipate and control the consequences of its use. The other half argued that, as with accuracy rates, ML is not an inherently off-limits feature, as the associated unpredictability can be compensated through other means, notably by placing strict limits on the temporal and geographical scope of an attack.

However, most participants agreed that AWS which use 'online learning' (see box 4) for target recognition would constitute an inherently impermissible technical characteristic on the basis of unpredictability. In contrast to the scenarios involving ML where the 'learning' took place in the training phase, everyone seemed to agree that an AWS that 'learns' after activation, which may include that it changes parameters without the knowledge of its users, would, regardless of the context, be off-limits because it would prevent users from reasonably anticipating the system's behaviour and effects.

Controllability as the basis for prohibiting use of an AWS?

Many states, as well as the ICRC, have expressed the view that AWS whose effects cannot reasonably be controlled should be prohibited as they cannot be used in compliance with IHL.[22] However, the implications this may have on the design and use of AWS is debated. Some argue that an AWS which by design prevents end-users from supervising, intervening, suspending or deactivating the system once activated, should be subject to strict temporal and geographical limits, if used at all.[23] Others argue that *constant* human supervision is not required by IHL (what matters is that users can exercise 'situational judgement'during the attack) and, therefore, AWS that are designed to operate without human intervention, are not necessarily off-limits.[24]

To explore what limits flow from the need to control the behaviour and effects of AWS, an inherent feature in all scenarios was that the systems were designed to be used in a communication-denied environment. That is, the AWS in the scenarios did not require constant, real-time human supervision and intervention (also referred to as AWS operating with 'humans out of the loop'). A number of participants argued that such a feature would only be permissible in very narrow situations where the risks of misidentification and inflicting excessive harm on civilians were minimal. An example of such use would be the targeting of a military objective that is both stationary and large, such as a military installation, in areas where civilians are not present. Thus, grounded in legal, but also ethical and operational concerns, the position of many participants was that controlling the behaviour and effects of AWS entails the ability to supervise, intervene, suspend and cancel an attack right up until the 'last minute'.

Other participants held that real-time and constant supervision and the ability for humans to intervene are 'desirable' features but are not legally required and should only

[22] See e.g. CCW GGE on LAWS, 2021 meeting, Document submitted by the Netherlands (note 7) and Written contribution submitted by Argentina et al. (note 17) CCW GGE on LAWS, 2023 meeting, Working papers submitted by Palestine, CCW/GGE.1/2023/WP.2 (note 14), Pakistan CCW/GGE.1/2023/WP.3 (note 7), Austria, CCW/GGE.1/2023/WP.1 (note 14) and Australia et al., CCW/GGE.1/2023/WP.4 (note 3); and ICRC, 'ICRC position on autonomous weapon systems' (note 14).

[23] ICRC, 'ICRC position on autonomous weapon systems' (note 14); CCW GGE on LAWS, 2023 meeting, Palestine, CCW/GGE.1/2023/WP.2 (note 14); and CCW GGE, 2022 meeting, 'Protocol VI', Working paper submitted by Argentina, Ecuador, Costa Rica, Nigeria, Panama, the Philippines, Sierra Leone and Uruguay, July 2022.

[24] United States, Statement delivered at the Second Session of the CCW GGE on LAWS, Geneva, 17 May, 2023, UN Web TV, 00:47:00; and CCW GGE, 2022 meeting, 'United Kingdom proposal for a GGE document on the application of international humanitarian law to emerging technologies in the area of lethal autonomous weapon systems (LAWS)', Proposal submitted by the UK, Mar. 2022, annex A.

be in place if they are feasible in light of the larger context of the military operation. They argued that what matters for legal compliance is for users to maintain an ability to exercise situational judgement, which they considered possible without users having constant supervision of the system.

Despite these differences in views, there was broad agreement among participants that users must be able to reasonably control the behaviour and effects of AWS. Participants cited IHL rules, notably the principles of distinction, proportionality and precautions in attack, because compliance with these rules presumes an ability to control the effects of the system. Also, from an operational standpoint, it was stressed that commanders would necessarily want to exercise some form of control over a system's functions and effects, and that maintaining control of the system is considered important both to ensure operational safety and to minimize the risk of unintended harm as well as escalatory effects. However, whether control of an AWS can be exercised in ways other than through direct supervision and direct control over the system was, as in the policy debate, highly debatable.[25]

Traceability as the basis for prohibiting use of an AWS?

Another key question in the policy debate is whether to formulate a prohibition on the basis of traceability. Some, such as Austria, frame the issue in positive terms, arguing that those authorizing the use of AWS 'must be able to trace back the outcome of the use of force to human agency', while others frame the issue in the negative, arguing for prohibition of AWS whose use 'cannot be traced'.[26] The underlying concern is that AWS which are not traceable undermine the ability to ensure that AWS are used within a responsible chain of command and control, and the ability to assign responsibility in case of a breach of IHL. However, others, such as the USA, counter these suggestions with the argument that there is no requirement in IHL to trace back.[27]

To unpack views around when, if ever, AWS would be impermissible on the basis of traceability, the scenarios involved loitering munitions whose automatic target-recognition capabilities were trained using ML-based algorithms (for example Alpha-100, variation 1). Since this introduced more opacity about how targeting decisions are produced (i.e. 'the black box'), the intent was to test whether technical 'explainability'—the ability to understand the computational processes embedded in a certain algorithm—is needed to ensure the ability to trace back.[28] For about half of the participants, that was the case: they held that the AWS used in the scenario should be prohibited because of the inability to investigate and trace back possible violations of IHL. The other half opposed this approach to traceability, arguing that explainability is neither required by IHL nor necessary for tracing behaviour to the agents responsible. Instead, these participants reiterated that tracing back responsibility for the use of ML-based systems can be ensured through, for example, limits on how and where the systems are used and through clearly established roles and responsibilities.

Despite different views around whether AWS *by design* can be deemed impermissible on the basis of traceability, everyone seemed to agree that AWS that *by use* undermine

[25] For a discussion on the ways in which control can be exercised, see Boulanin et al., *Limits on Autonomy in Weapon Systems* (note 5), p. 27.

[26] CCW GGW on LAWS, 2023 meeting, Revised working paper submitted by Austria, CCW/GGE.1/2023/WP.1 (note 14); and Working paper submitted by Argentina et al., CCW/GGE.1/2023/WP.6 (note 3).

[27] United States, Statement delivered at the Second Session of the CCW GGE on LAWS (note 24) 00:47:00.

[28] There is no consensus about the notion of 'explainability' but one way of understanding it is as a 'condition where a human designer, user or auditor can understand the workings of an algorithm and thereby make predictions or deductions as to how it will behave or why it did behave in a certain way'. Kwik, J. and van Engers, T., 'Performance or explainability? A law of armed conflict perspective', A. Kornilakis et al. (eds), *Artificial Intelligence and Normative Challenges: International and Comparative Legal Perspectives*, Law, Governance and Technology Series, vol. 59 (Springer: Cham, 2023), p. 259.

the ability to trace back are highly problematic. Examples include using an AWS without having sufficient organizational measures in place to trace back decisions made in earlier phases, including programming and testing. Beyond legal obligations to investigate IHL violations, several participants argued that traceability is also critical to ensure political accountability. It was argued that since governments are accountable to their citizens in terms of how and where they use force and as such are subject to public scrutiny, they have a broader interest in ensuring that their use of AWS is traceable.

Summary of views

Limits on the design and use of AWS are needed to ensure permissible use

The views expressed in the scenario exercise reflected agreement that to ensure permissible use of AWS, limits are needed on *what/who* AWS can target, *where* they can be used and *how* they identify targets. What became particularly clear during the scenario exercise is the importance of considering *combinations* of variables and the unpredictability associated with these when identifying such limits and requirements (figure 3). In several cases, the question for many participants was not whether a single variable was off-limits in all circumstances but the combination of variables that would be.[29] The underlying rationale is that the unpredictability introduced by one variable (e.g. fluid status of the target) must be 'compensated' for by other, more predictable variables (e.g. environments free of civilians). If several variables related to an attack are unpredictable, the ability for users to comply with their obligations under IHL, particularly to ensure that distinction and proportionality assessments remain valid after activation, is likely to be undermined. Drawing from the three main variables used in the scenario exercise, it appeared that once a scenario entailed more than one unpredictable variable, most participants considered the combination to be problematic, if not impermissible.

But fundamentally different concerns appear to be the main barrier to agreeing on where to draw limits

Despite agreement that limits are needed, participants were still divided when it came to identifying particular limits. What perhaps stood out as the biggest area of divergence concerned the lack of agreement around categorical limits, that is, whether there are types of targets, environments or technical characteristics that are prohibited regardless of the context and combination with other variables. A key reason *why* participants arrived at different conclusions stems from the fundamentally different standpoints on which they evaluated the risks associated with AWS.[30]

First, the discussion revealed that participants considered legal aspects in radically different manners. Some evaluated the permissibility of AWS using a utilitarian, results-oriented perspective on IHL, with a focus on AWS as a means for potentially achieving 'better' compliance compared with alternative means and methods of warfare. Others adopted a more principle-oriented approach, arguing that, regardless of how sophisticated technology becomes, IHL rules generally (and, in particular, the rules on

[29] The importance of considering combinations of variables (i.e. different contexts) when identifying limits on autonomy is also reflected in Guiding Principle C, adopted by the GGE as well as in previous research findings. See CCW GGE on LAWS, Report of the 2019 session, CCW/GGE.1/2019/3, 25 Sep. 2019, annex IV, 'Guiding principles'; and Boulanin, V. et al., *Limits on Autonomy in Weapon Systems* (note 5).

[30] It is worth noting that different modes of analysis adopted by participants should also be understood in relation to their different starting points. That is, what participants consider a desired outcome for political, ethical or operational reasons was likely to determine their mode of analysis.

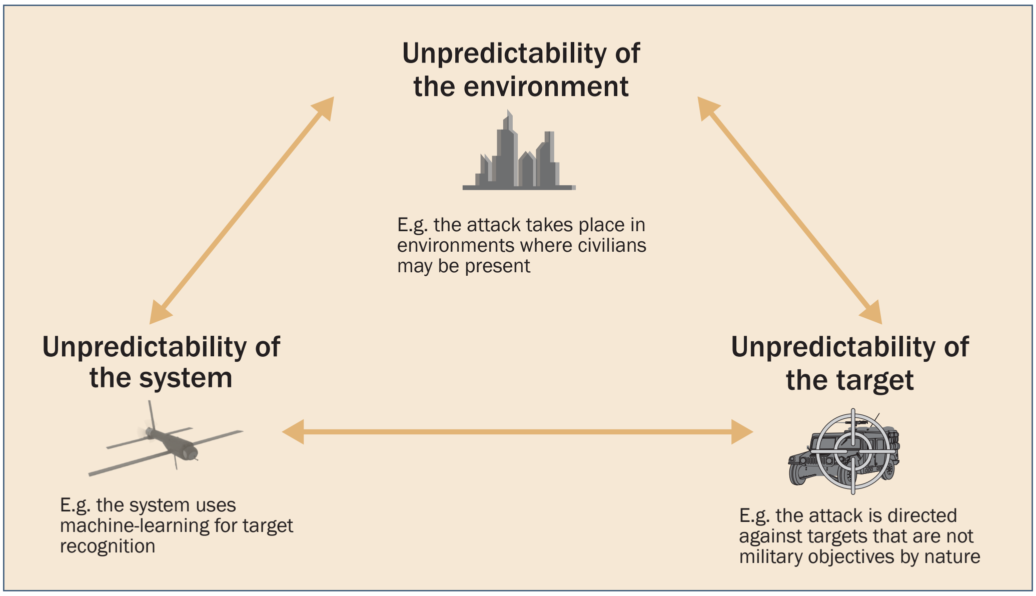

Unpredictability of the environment
E.g. the attack takes place in environments where civilians may be present
Unpredictability of the system
E.g. the system uses machine-learning for target recognition
Unpredictability of the target
E.g. the attack is directed against targets that are not military objectives by nature

Figure 3. Combinations of unpredictable variables used in the SIPRI scenario exercise

Credit: Vector images from Vecteezy.com

distinction, proportionality and precautions in attack) require human evaluation and judgement throughout AWS use.[31]

Second, a number of participants grounded their assessments in reasons beyond IHL, notably ethical considerations and disarmament concerns. The exercise showed that these participants were likely to arrive at different conclusions than those who, for example, informed their risk assessments from a results-oriented approach to IHL. That is, those who evaluated risks based on larger socio-technical questions around the relationship between humans and machines and government accountability, concerns around proliferation and misuse, and deontological concerns were more likely to draw more categorical lines, for example against the use of AWS in populated areas or against persons. However, while ethical concerns were often mentioned, they were rarely unpacked. In several cases, notably when discussing anti-personnel AWS, participants argued that although these systems may be permissible from a legal standpoint, they would not use them due to ethical concerns, but without elaborating further. While this possibly reflects shortcomings connected with the workshop methodology (the scenarios were largely designed to generate debate around IHL questions and most of the participants were IHL lawyers), it suggests that non-legal concerns appear relevant but under-explored in this context.

[31] See Boulanin, Goussac and Bruun (note 6).

4. Potential implications for developing a two-tiered approach to regulation of AWS

This chapter considers whether, and if so how, takeaways from the scenario exercise can be generalized into insights that could help develop the content and contours of a two-tiered approach to regulation of AWS. The focus is on areas of convergence, rather than divergence, with the aim of identifying a direction for the policy debate. The elements outlined in the chapter are thus intended as 'pointers' for policymakers, not definitive answers.

The centrality of human agency

Despite the wide spectrum of views, there was one element that connected most, if not all, assessments: targeting decisions ultimately lie with humans, not machines, and the ability to exercise human agency in targeting decisions must therefore be preserved. Consequently, there must be limits on technical characteristics, types of target and environments of use that alone or in combination undermine the ability to exercise human agency. Recognising that the concept of 'human agency' is complex and subject to different meanings, its use here reflects the agreement among participants that no matter what, users of an AWS must be able to reasonably anticipate and control the behaviour and effects of the system.[32]

While everyone seemed to agree on the centrality of human agency, they arrived at this conclusion for different reasons. Most mentioned that IHL compliance requires the ability to anticipate, limit and control the effects of the use of force—for example, the principle of proportionality specifically refers to the ability to *anticipate* the effects of the attack.[33] Moreover, participants' views reflected that the ability to reasonably anticipate and control the behaviour and effects of AWS is also critical for reasons related to tracing responsibility in case of a breach. The centrality of ensuring predictabilty and controllability was also grounded in broader policy considerations, including: ensuring accountability and minimizing the risk of conflict escalation; operational concerns about losing control over the use of force; and, last but not least, ethical principles of taking into account those affected by the use of force. The importance of ensuring users' ability to reasonably anticipate and control the behaviour and effects of AWS is a useful lesson that builds on what many states have already expressed in the policy debate, as well as on findings of past research conducted by SIPRI.[34]

Translating human agency into two tiers

This insight suggests that grounding regulation of AWS in the centrality of human agency could be a viable way to identify limits and requirements under a two-tiered approach. The importance of thinking holistically about human agency—*who* should anticipate and control *what*, *where* and *when*—could usefully be structured as complementary tiers. Tier 1 could contain a prohibition on AWS design and use that in all circumstances undermine the ability of users to exercise agency in targeting decisions; and tier 2 could elaborate in more detail what users are required to do—and what oper-

[32] Recalling the scope of 'users' in this report as set out in chapter 2 and note 10.

[33] Compliance with this principle depends on the 'concrete and direct military advantage *anticipated*': AP I (note 12) Arts 51(5)(b) and 57 (emphasis added). See also ICRC, 'Rules', Customary IHL Database, [n.d.], Rule 14.

[34] See e.g. CCW GGE on LAWS, 2022 meeting, Working papers submitted by Argentina et al. (note 23), and China, July 2022, and Proposal submitted by Australia et al. (note 13); CCW GGE on LAWS, 2023 meeting, CCW/GGE.1/2023/2 (note 2); and Boulanin, Goussac and Bruun (note 6.)

ational limits are needed—to ensure that users can reasonably anticipate and control the behaviour and effects of AWS. Based on areas of convergence identified at the scenario exercise, the following considers possible elements around a prohibition and around limits and requirements that flow from the centrality of human agency.

Elements of convergence around a prohibition (tier 1)

The views expressed in the scenario exercise suggest the viability of formulating a prohibition around the design and use of AWS whose behaviour and effects cannot be reasonably anticipated, controlled and traced back, which aligns with what many states have already suggested.[35] However, there is not much agreement on what, if any, technical characteristics or uses should be subject to a prohibition on these bases. Based on the options provided in the scenarios, the closest participants came to an agreement was around *prohibiting AWS that use online learning for target recognition*. The main difficulty in agreeing on prohibitions around technical characteristics of AWS appears to be that, according to many, few, if any, technical characteristics are inherently unlawful, at least from an IHL perspective.

Elements of convergence around limits and requirements (tier 2)

The difficulties with agreeing on inherently off-limits technical characteristics suggest that there may be more room to establish specific limits and requirements. Such elements, potentially forming part of the second tier, could thus be used as a vehicle to specify in more detail what users are required to do—and what limits are needed—to ensure the ability to exercise human agency and trace back responsibility. To this end, the scenario exercise reflected an emerging consensus among participants around a set of requirements, as well as some limits, that could usefully be further explored as elements needed to ensure the exercise of human agency in targeting decisions.

Requirements

According to the participants, there are a number of positive actions that users of an AWS should take to ensure the ability to anticipate, control and trace back the behaviour and effects of the AWS. Without prejudging whether these should be considered legally binding—for example, as a specification of the obligation to take feasible precautions in attack, or as elements of good practice—these actions can be summarised as follows.

Users should possess sufficient technical knowledge about the system. To ensure that users of an AWS can exercise agency, users should possess sufficient technical knowledge about the system. In particular, users should be able to understand what conditions will trigger an application of force, the system's expected performance (including fail rates) in relation to different environments, and the system's ability to distinguish between positive and false IDs. This stresses the importance of at least two sets of requirements. First, those developing and acquiring an AWS should ensure that the system undergoes extensive testing in simulations that reflect, as closely as possible, the intended environments of use. Second, those deploying and operating an AWS should receive sufficient training on how the system works, including its expected performance and behaviours in different environmental conditions and in interaction with other

[35] See e.g. CCW GGE on LAWS, 2023 meeting, Working papers submitted by Australia et al., CCW/GGE.1/2023/WP.4 (note 3), Austria, CCW/GGE.1/2023/WP.1 (note 14) and Palestine, CCW/GGE.1/2023/WP.2 (note 14), and 'Additional text proposals', collected by the Chair, 18 May 2023; CCW GGE on LAWS, 2021 meeting, Document submitted by the Netherlands (note 7) and Written contribution submitted by Argentina et al. (note 17); and ICRC, 'ICRC position on autonomous weapon systems' (note 14).

systems. Because technical knowledge is also relevant to traceability, the required user knowledge would include knowledge of how, and on what data, the system was trained and programmed.

Users should have extensive knowledge of the environment of use. To be able to reasonably anticipate and control the behaviour and effects of AWS, everyone agreed on the importance of users possessing extensive knowledge of the environments in which the systems are deployed. While a requirement around knowledge about the environment of use is not unique to AWS, discussions indicated that this requirement becomes particularly relevant in the AWS context because, when using AWS, force is triggered by sensor inputs from the environment. To ensure that legal assessments made upon activation remain valid throughout the attack, it is essential that users understand the environment and reasonably anticipate possible changes within the target area. The question is, then, what level and type of information about the environment should users possess to be reasonably certain that their assessments about distinction and proportionality remain valid after activation? The views expressed in the scenario exercise suggest that knowledge about the environment of use should be understood in the broadest sense possible and would entail reliable, specific and extensive knowledge about movements of civilians in the wider area (beyond just the target area) and up-to-date information about different movements within the target area.[36]

The roles and responsibilities of those involved in developing and using AWS should be clear. Most participants considered that the ability to reasonably anticipate, control and trace back the behaviour and effects of AWS applies to actors beyond just end users. That is, ensuring permissible use of AWS is a collective responsibility, involving programmers, data labellers, and decision makers at the procurement and operation planning levels.[37] The role of those involved in the development phase was considered especially important as the preprogrammed nature of AWS means that many decisions are made at this early stage.[38] Consequently, participants' views reflected a consensus that the challenges posed by AWS cannot solely be addressed in the use phase and that states must have institutional measures in place to account for this collective responsibility. Such measures should be structured in a way that aims to clearly establish, clarify and communicate the roles and responsibilities of those involved in the design and use of AWS. Such a requirement could aim to flesh out what the notion of a 'responsible chain of command and control' entails, for example by detailing what developers are required to know about the intended environment of use.[39]

Limits

In addition to requirements, it also appears undisputed that maintaining the ability to exercise human agency in targeting decisions necessitates there be limits on AWS, including on types of targets, environments of use and technical characteristics. While IHL already places limits on the means and methods of warfare, insights from the scenario exercise suggest that there may be scope to further specify what these limits entail in the context of AWS. Based on areas of convergence among participants,

[36] See also Bruun, Bo and Goussac (note 6), pp. 17–19.

[37] For a breakdown of AWS lifecycle stages see Blanchard, A., Thomas, C. and Taddeo, M., 'Ethical governance of artificial intelligence for defence: Normative tradeoffs for principle to practice guidance', *AI & Society*, 19 Feb. 2024, p. 7.

[38] Indeed, the complex network of agents involved in the use of AWS may have implications for how, where and by whom agency is exercised, and prompts the need to look closer at the larger decision-making structures, extending beyond the armed forces. See Bo, Bruun and Boulanin (note 9), p. 20.

[39] See also Boulanin, V. and Bo, M., 'Three lessons on the regulation of autonomous weapons systems to ensure accountability for violations of IHL', ICRC Humanitarian Law and Policy Blog, 2 Mar. 2023.

there may be particular scope to place limits on certain *combinations* of target types, environments and technical characteristics in cases where a combination is considered to undermine users' ability to reasonably anticipate, control and trace back the behaviour and effects of AWS. Establishing limits on *combinations* of variables (e.g. as part of 'second tier' restrictions) is an approach that has been used in two-tiered regulations before, including within the CCW. For example, Amended Protocol II on Mines, Booby-traps and Other Devices includes restrictions against the use of anti-personnel mines in areas where civilians are present.[40] Similarly, some of the restrictions set out in Protocol III on Incendiary Weapons are structured around the combination of (*a*) 'making any military objective located within a concentration of civilians the object of attack' (*b*) 'by air-delivered incendiary weapons'.[41]

This observation suggests that one way forward could be for states to elaborate on problematic combinations and explore the value of using potential consensus elements as a basis for establishing limits. The guiding principle for such an exercise would be to explore what combinations of (for example) target types, environments and technical characteristics are considered to undermine users' ability to exercise agency in targeting decisions and then delineate limits from there. Narrowing down and identifying the contours of problematic contexts and combinations could also help allow states to bypass the 'context-dependency' problem.

[40] CCW Convention (note 1), Amended Protocol II on Mines, Booby-traps and Other Devices, opened for signature 10 Apr. 1981, entered into force 3 Dec. 1998, Arts 3 and 5(6)(b).

[41] CCW Convention (note 1), Protocol on Prohibitions or Restrictions on the Use of Incendiary Weapons (Protocol III), opened for signature 10 Apr. 1981, entered into force 2 Dec. 1983, Art. 2(2).

5. Key findings and recommendations

This report aims to help stakeholders in the international policy debate on AWS to advance consensus on the contours and content of a two-tiered approach to regulation of AWS. It is informed by a scenario exercise that invited selected experts to elaborate on the limits and requirements they would deem necessary in specific use cases. This chapter summarizes the key findings concerning (*a*) the value of the methodology; (*b*) the landscape of views and rationales around what limits and requirements are needed; and (*c*) areas of convergence and their potential implications for the development of a two-tiered regulatory approach. The chapter concludes with a series of recommendations to stakeholders, including governmental experts and representatives from academia and civil society, involved in the intergovernmental policy discussions on AWS.

Key findings

Scenarios provide a useful methodology for generating insights into limits and requirements in the context of AWS

The scenario exercise appeared to be a useful format for generating deeper discussions around the limits and requirements for the development and use of AWS. It provided an opportunity to focus on some of the most contested issues and understand in greater detail where (and why) views diverge. The specific use cases presented in the scenarios were useful vehicles for exploring how specific concerns arising from a use case can generate more general insights about limits and requirements that are needed in the context of AWS. However, limitations to this methodology included the inability of participants to prepare and engage on an equal footing; the risk that scenarios could be instrumentalized to normalize the use of AWS in ways that most participants did not want; and the imbalance in expertise, with most participants being IHL lawyers, leading to other concerns and perspectives being under-explored.

There is little agreement about what categorical limits are needed, partly due to different understandings about the role of ethical concerns

The scenario exercise revealed a wide spectrum of views concerning limits needed in the design and use of AWS to ensure permissible use. Notably, when discussing potential elements under tier 1, there was little consensus about which categories of targets, environments or technical characteristics, if any, were off-limits in all circumstances. A key reason for divergence was the different perspectives and concerns that informed participants' assessments. Some approached the questions mainly, if not solely, through the lens of IHL, while others mobilized a wider set of considerations grounded in concerns related to ethics and human rights. When participants considered concerns beyond IHL, their assessment would become more restrictive, often including categorical limits. However, while such concerns, especially those related to ethics, were often used to identify prohibitions, they were rarely unpacked. And even when elaborated upon, the discussion indicated that participants had very different understandings of ethical concerns in the context of AWS and what role they should play in identifying limits.

Insights from the scenarios suggest structuring a two-tiered regulation of AWS around ensuring the exercise of human agency

Across the wide spectrum of views, the centrality of ensuring the exercise of human agency in targeting decisions, including the ability to reasonably anticipate and control the behaviour and effects of an AWS, operated as the starting point for most participants when identifying limits and requirements. This insight suggests the viability of grounding a two-tiered approach to regulation of AWS in the centrality of human agency. Tier 1 could contain a prohibition on design or use that undermines the ability of users to exercise agency in targeting decisions, while tier 2 could establish what users are required to do—and what limits are needed—to ensure that users exercise agency. However, telling from the divergence in views observed in the scenario exercise, it may be difficult for states to agree on specific elements under tier 1; that is, which (if any) technical characteristics or use cases should be prohibited because they undermine the exercise of human agency in all circumstances. Instead, the scenario discussion indicated that it may be more viable for states to seek consensus around limits and requirements under a possible second tier. It especially appeared easier for participants to agree on what those involved in both development and use are *required* to do and know, to ensure the ability to exercise human agency in targeting decisions and trace back responsibility.

Recommendations

Use scenarios alongside other methods to further unpack limits and requirements for the development and use of AWS

States could usefully use scenarios as a vehicle to further deepen their understanding of what, if any, technical characteristics or uses of an AWS are off-limits alone or in combination, and what restrictions are needed to ensure its permissible use. Such an exercise could be conducted internally as a vehicle for states to elaborate on positions for their own purposes, and also in broader international forums. The scenarios that served as a baseline for this report could be used as a starting point for such efforts. However, it is important to note that scenario exercises should not stand alone in efforts to identify limits and requirements, and should ideally be combined with other methods, including more principle-based top-down approaches.

Further explore the role of ethics when identifying limits and requirements for AWS

Compared to progress made on legal aspects, the scenario exercise showed that the nature and role of 'non-legal' concerns, especially broader ethical concerns, are under-explored in the AWS debate. Several participants recognised this shortcoming and argued it would be useful for states to further explore perspectives beyond IHL, especially in efforts to identify elements of a prohibition. Insights from the scenario exercise highlight the particular need to clarify ethical aspects in the regulatory debate. To this end, states could usefully elaborate on the arguments and rationales underpinning ethics-based interventions, and subsequently explore the role such concerns should play for the identification of limits and requirements applicable to AWS.

Further elaborate on the specific technical characteristics and uses that, alone or in combination, undermine the ability to exercise human agency in targeting decisions

One way for states to advance their (common) understanding of the content and contours of a two-tiered regulation of AWS would be to ground the conversation in the importance of preserving the ability to exercise human agency. That is, the policy conversation could more systematically explore what prohibitions, limits and requirements are needed to help ensure human users are able to reasonably anticipate and control the behaviour of AWS and trace back responsibility. To this end, states could specifically explore which variables that, alone or in combination, undermine the ability to exercise agency in targeting decisions. Variables that deserve dedicated discussion in this context include (but are not limited to) AWS that use ML for target recognition, environments where civilians are present, and targets whose status as lawful targets depends on the context. Grounding elements in preserving the ability to exercise of human agency would also help ensure that a regulatory instrument or framework remains relevant irrespective of technological developments.

Annex

This annex contains a description of the two scenarios that were used as a basis for the scenario exercise SIPRI conducted in Stockholm, in January 2024. Both scenarios contain two parts, one designed as a basis for identifying off-limits characteristics and uses of a hypothetical autonomous weapon system (AWS), and one designed as a basis for identifying types of context-specific limits and requirements to be applied to the AWS. While the scenarios aimed for technological and military plausibility, certain aspects were deliberately simplified due to practical limitations. Following the two scenarios are the assessment templates provided to participants for each part of the scenarios.

Scenario A: Anti-armour loitering munition (‘Alpha-100’)

Part 1: Is the system’s design and intended use permissible?

<table>
<tr><td><b>Capabilities and intended use</b></td><td>Alpha-100 is intended to neutralize enemy tanks or other armoured vehicles in communication-denied environments where remote control is not possible for friendly combatants to access. Once activated, the system can, without further human intervention, loiter over a 20 km^2 area for up to two hours to identify, track, prioritize, select and strike targets on the ground. The system’s maximum speed is 300 km p/h and the maximum payload is 20 kg. It is designed with armour-piercing warheads with a minimal blasting effect (3 meters).</td></tr>
<tr><td><b>How the system recognizes targets</b></td><td>The automatic target recognition (ATR) software recognizes tanks and armoured vehicles based on their shape, heat signatures and height. To do so, it relies on computer vision using sensory input from infrared cameras and lidar systems. The system applies predefined rules to the incoming sensor data to determine whether the observed object matches the characteristics of a target and then prioritizes possible actions. The library of target signatures has been developed based on intelligence, surveillance and reconnaissance (ISR) footage from previous operational missions. The library includes information about the target (all models of tanks and armoured vehicle that are known to be used by the enemy forces) and different operational environments.</td></tr>
<tr><td><b>Performance</b></td><td>The system has been tested in computer simulations and in mock operational situations. Tests shows an accuracy rate (in terms of target recognition) of 85% when deployed in uncluttered, observable environments and under clear weather conditions (in the sense of daylight and no rain).</td></tr>
<tr><td><b>Limitations</b></td><td>The system can recognize predefined target profiles only based on those that exist in its target library. It is not trained to identify negative IDs (i.e. objects that must <b>not</b> be made the object of attack, such as civilian objects), and it cannot recognize humans. Test shows that the system’s accuracy rate (in terms of target recognition) decreases to 65% when deployed in cluttered environments.</td></tr>
<tr><td><b>Human–machine interaction</b></td><td>The system can be supervised by an operator in real-time in situations where communication is possible. When operating in communication-denied environments, which is the primary use case, a human will not be able to intervene, suspend or cancel an operation.</td></tr>
<tr><td colspan="2"><i>Variation 1: The system’s ATR software involves machine-learning</i></td></tr>
<tr><td colspan="2">Rebranded as <b>Alpha-500</b>, the system has been updated with a new ATR system that has been developed with machine-learning. The system has been trained via supervised learning methods, i.e. where it is trained by labelled data and ‘learns’ by comparing inputs with desired outputs. When tested in controlled and uncluttered environments, the system’s accuracy rate is estimated to increase to 95%. The ATR has been trained to distinguish between military objects and civilian objects by nature (e.g. a tank versus a bus). The system is programmed to self-deactivate if at least one civilian object is detected within a 20-meter radius of a potential target location. Intended use: The neutralization or destruction of enemy tanks in communication-denied areas where the presence of civilians and/or civilian objects cannot be ruled out.</td></tr>
<tr><td colspan="2"><i>Variation 2: The system is being trained to also identify vehicles used for military purposes</i></td></tr>
<tr><td colspan="2">Alpha-500’s ATR software is further developed and is now trained to also recognize pick-up trucks with a mounted turret. For this, additional predefined criteria—related to shape, heat and height, and notably whether turrets are detected on the vehicle—are included and the system is programmed to apply force only once both criteria are satisfied, i.e. civilian trucks and turrets. The system performs with an 88% accuracy rate when tested in uncluttered environments and under clear weather conditions (in the sense of daylight and no rain). Intended use: The neutralization or destruction of adversary capabilities in communication-denied environments where the presence of civilians and/or civilian objects cannot be ruled out.</td></tr>
</table>

Part 2: What restrictions are needed to ensure that the use of Alpha-100 in the described situation is permissible?

Mission description

- *Mission objective:* Neutralize or destroy enemy tanks and armoured vehicles.
- *Context*: In a high-intensity international armed conflict, State A seeks to regain control of a border area taken over by State B. The two states have been in direct combat since State B, two months ago, launched an offensive military campaign to gain control over territory that it claims the historical right over. State A, which acts in self-defence to protect its borders and population, launches a series of attacks against military objectives which form part of the enemy's capabilities in a specific area. The destruction of strategically important enemy capabilities, including tanks, missiles and other weapons, and military equipment, is considered a key step to prevent, or at least delay, State B's offensive operations aiming to take further control of State A's territory. State B most likely possesses advanced military equipment, including long-range missiles and electronic jamming capabilities.
- *Environment*: Unpopulated combat zone. No civilians have been observed in the area since the outbreak of the conflict. Communication is possible but may be jammed.
- State A intends to use Alpha-100 loitering munitions to carry out the operation.

Variation 1: Alpha-500 is deployed in a rural area between two villages

- Instead of an unpopulated combat zone, enemy tanks are operating in a rural area located between two villages, which means that civilian presence cannot be excluded. The increased accuracy and the enhanced ability to distinguish military objects from civilian objects associated with the use of Alpha-500 make it a relatively favourable choice.

Variation 2: Alpha-500 is used to direct attacks against vehicles used for military purposes

- Intelligence shows that the adversary uses civilian pick-up trucks from which it launches attacks. To do so, turrets are, among other methods, installed on the pick-ups. The target area is centred around a highway connecting two of State B's military locations. Usually, the highway is used by civilians, but no civilians have been observed passing through since the outbreak of the conflict (based on data collected by State A's ISR drones). Alpha-500, trained to detect and attack tanks as well as civilian vehicles used for military purposes, is suggested to carry out the mission.

Scenario B: Anti-personnel loitering munition ('Beta-100')

Part 1: Is the system's design and intended use permissible?

<table>
<tr><td><b>Capabilities and intended use</b></td><td>Beta-100 is intended to be used to target persons who are visibly armed and wearing a military uniform. Specifically, it is designed to clear trenches in environments where it is not possible or feasible to deploy remote-controlled systems. Once activated, the system can, without further human intervention, loiter for up to 20 minutes within a maximum range of 15 km^{2} to identify, track, prioritize, select and strike targets on the ground. The system's maximum speed is 100 km p/h and the maximum payload is 1.5 kg.</td></tr>
<tr><td><b>How the system recognizes targets</b></td><td>The system's ATR applies deep-learning algorithms to identify members of the enemy forces. The two proxy indicators, upon which the system recognizes targets, relate to whether the person is (<i>a</i>) visibly armed and (<i>b</i>) in uniform. To do so, it relies on image-recognition software using sensory input from cameras and lidar systems. The system has been trained on an extensive dataset that involved a vast number of objects that are structured similarly to weapons, but which are not weapons. Examples of this from the dataset include objects like saws, shovels and video cameras. The aim is to strengthen the system's ability to identify false IDs. The system is trained to detect surrender motions (in the shape of arms held above the head to indicate surrender) and is programmed <b>not</b> to engage if it detects gestures indicating surrender.</td></tr>
<tr><td><b>Performance</b></td><td>The system has been tested in computer simulations and in mock operational situations. Tests shows an accuracy rate of 85% when deployed in uncluttered, observable environments and under clear weather conditions (in the sense of daylight and no rain).</td></tr>
<tr><td><b>Limitations</b></td><td>The system is able to identify only certain types of negative IDs (i.e. that must <b>not</b> be made the object of attack) with a particular level of accuracy (88%). The type of negative IDs that the system has been tested to recognize concern persons or structures marked with predefined medical, UN or press emblems. The system's accuracy rate (in terms of target recognition) decreases to 65% when deployed at night.</td></tr>
<tr><td><b>Human–machine interaction during use</b></td><td>The system can be supervised by an operator in real-time in situations where communication is possible. When operating in communication-denied environments, which is the primary use case, humans will not be able to intervene, suspend or cancel attacks.</td></tr>
<tr><td colspan="2"><i>Variation 1: Based on a library of human targets, the system is designed to search for and attack specific members of enemy forces</i></td></tr>
<tr><td colspan="2">Beta-100 instead uses biometric target identification and facial recognition software to recognize high-ranking members of the enemy forces as listed in the library. The system relies on incoming sensor data from cameras and lidar systems to determine whether the observed person matches the characteristics of one of the pre-identified targets in its database. It is trained to recognize targets via photos and videos of members of the enemy forces that have been collected from open sources as well as ISR missions. The training dataset included pictures and videos of the target in different operational conditions (e.g. different weather conditions or backgrounds with varying clarity). Intended use: With its small payload, the system is primarily intended to replace precision airstrikes against high-value targets in populated and communication-denied areas. It is thus intended to be used in environments that are too dangerous or impossible to access by remote control systems or ground troops.</td></tr>
</table>

Variation 2: The system is designed to recognize human targets based on behaviour
Beta-100 is now designed to identify, select and attack civilians directly participating in hostilities. The basis on which targets are recognized is whether a person demonstrates the intent to use force. This element (i.e. the intention to use force) is determined based on a combination of factors, notably whether the person is carrying a weapon and whether the person's body language indicates an 'offensive posture'. The system's ATR relies on three techniques: a weapon-detection algorithm (described above); pose estimation to identify the location and orientation of a human; and AI-enabled emotion recognition. Pose estimation includes describing the joints of the human body, such as the wrist, shoulder, knees, eyes, ears, ankles and arms, while the AI-enabled emotion recognition relies on deep-learning to analyse and understand human nonverbal signs such as facial expressions, body language, gestures and voice tones to assess their emotional state. All three models rely on image-recognition software using sensory input from cameras and lidar systems. Intended use: The targeting of civilians directly participating in hostilities who are operating from environments that are difficult or not safe for other systems or friendly combatants to enter.

Part 2: What restrictions are needed to ensure that the use of Beta-100 in the described situation is permissible?

Scenario description

- *Mission objective*: Clear trenches.
- *Context*: During the same international armed conflict as described in scenario A, the battle between States A and B is intensifying along the contested border area. In an attempt to push back State B, State A is launching an offensive operation to clear trenches within territory occupied by State B. According to satellite images, State B forces are operating from trenches along the border area in the mountains, both to seek shelter and to launch attacks.
- *Environment*: Unpopulated combat zone in the mountains. The exact movement within the specific target area is not completely observable to State A but considered relatively predictable. No civilians have been observed passing through the area since the outbreak of the conflict. The presence of medical personnel and members of the press, however, cannot be excluded.
- State A intends to use Beta-100 loitering munitions to carry out the operation.

Variation 1: The system is used to only attack high-profile members of enemy forces

- State A now seeks to eliminate high-ranking and influential members of State B's armed forces. The specific members are well-known to the public as playing a leading role in the development and execution of State B's military operations. Because these members are known to operate from cities, State A intents to use Beta-100 (programmed to target specific members of the enemy forces based on a library of targets) to carry out the operation.

Assessment templates

Assessment template: Part 1

What is your assessment of the system based on its technical characteristics, capabilities and/or intended effects in normal and intended use?
(please register your assessment by adding one or more 'X' in the table below)

The system is entirely permissible	
The system is permissible but certain limits are needed	
Due to legal considerations	
Due to ethical considerations	
Due to operational considerations	
Due to policy considerations	
Due to other reasons	
The system is off-limits	
Due to legal considerations	
Due to ethical considerations	
Due to operational considerations	
Due to policy considerations	
Due to other reasons	
I need to know more about the system	

Elaborate your answers:

1. Are there aspects of the system's characteristics and/or intended uses that are (or may be) problematic from a legal, ethical, operational[a] and or policy perspective?[b]
2. Are there aspects of the system's characteristics and/or intended uses that are—or may be—problematic for other reasons than those listed above?

[a] Operational standpoint refers to whether the system is considered military beneficial and whether it can be used in line with applicable rules of engagement and military doctrine, including within a responsible chain of command.

[b] For example, from the perspective of the Guiding Principles as adopted by the CCW or policy commitments made at a national level.

Assessment template: Part 2

What control measures are needed to ensure that the use of the system in the described situation is lawful and satisfies ethical, policy, security and operational considerations?

Examples of control measures	**Y/N**	**Such as?**
Control over the system		
Limits on target types		
Limits on the spatial scope of operation		
Limits on the temporal scope of operation		
Limits on weapon's effects		
Fail-safe requirements		
Other . . .		
Control over the environment		
Civilians and civilian objects are not present		
Maintain situational awareness		
Exclusion zone, physical barriers, warnings		
Other . . .		
Control via human–machine interaction		
Ensure human supervision		
Ensure ability to intervene and deactivate		
Ensure specific training of users		
Other . . .		

Elaborate your answers...

1. Why are the suggested control measures needed?
2. How, when, where and by whom can the suggested control measures be implemented?

About the author

Laura Bruun (Denmark) is a Researcher in the SIPRI Governance of Artificial Intelligence Programme. Her focus is on how emerging military technologies, notably autonomous weapon systems (AWS) and military Artificial Intelligence (AI), affect compliance with—and interpretation of—international humanitarian law (IHL). Laura has a background in Middle Eastern Studies (University of Copenhagen) and International Security and Law (University of Southern Denmark). Before joining SIPRI, Laura worked at Airwars in London, where she monitored and assessed civilian casualty reports from US and Russian airstrikes in Syria and Iraq. Laura has previously lived in both Egypt and Jordan, working with human rights issues in the MENA-region.

Laura's recent publications include 'Integrating Gender Perspectives into International Humanitarian Law', SIPRI Report (2023, co-author) and 'Compliance with International Humanitarian Law in the Development and Use of Autonomous Weapon Systems: What does IHL Permit, Prohibit and Require?', SIPRI Report (2023, lead author).

sipri

STOCKHOLM INTERNATIONAL
PEACE RESEARCH INSTITUTE
Signalistgatan 9
SE-169 72 Solna, Sweden
Telephone: +46 8 655 97 00
Email: sipri@sipri.org
Internet: www.sipri.org