sipri

ICRC

LIMITS ON AUTONOMY IN WEAPON SYSTEMS

Identifying Practical Elements of Human Control

VINCENT BOULANIN, NEIL DAVISON,
NETTA GOUSSAC AND MOA PELDÁN CARLSSON

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.

The International Committee of the Red Cross

The ICRC is an impartial, neutral and independent organization whose exclusively humanitarian mission is to protect the lives and dignity of victims of armed conflict and other situations of violence and to provide them with assistance.

The ICRC also endeavours to prevent suffering by promoting and strengthening humanitarian law and universal humanitarian principles.

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

LIMITS ON AUTONOMY IN WEAPON SYSTEMS

Identifying Practical Elements of Human Control

VINCENT BOULANIN, NEIL DAVISON,
NETTA GOUSSAC AND MOA PELDÁN CARLSSON

June 2020

sipri
STOCKHOLM INTERNATIONAL
PEACE RESEARCH INSTITUTE

COMITE INTERNATIONAL GENEVE
ICRC

Contents

Preface

Since 2014, the challenges posed by autonomy in weapon systems have been the focus of an intergovernmental discussion under the framework of the Convention on Certain Conventional Weapons (CCW). The Stockholm International Peace Research Institute (SIPRI) and the International Committee of the Red Cross (ICRC) have followed the discussion closely from the outset. In line with their respective institutional mandates, SIPRI and the ICRC have contributed analyses to support an informed discussion at the CCW, focusing on furthering states' understanding of the legal, ethical, technical and humanitarian implications of autonomy in weapon systems.

Experts from SIPRI and the ICRC share a similar understanding of the challenges posed by autonomy in weapon systems and how these might be addressed. The work of SIPRI and the ICRC notably shares the viewpoint that, when exploring the need for limits on autonomous weapon systems (AWS), a fundamental issue to be addressed is that of human control. Autonomy in weapon systems transforms the way humans interact with those systems and ultimately make decisions on the use of force. Although autonomy will never completely displace humans from decision making, the concern is that it creates more distance in time, space and understanding between human decisions to use force and the consequences. This distancing, and the unpredictability in consequences it brings, in turn raises concerns about the application of international humanitarian law, ethical acceptability and operational effectiveness. The central question is: How can we ensure that humans continue to play their necessary role in decisions to use force in specific attacks in armed conflict, regardless of the sophistication of the technology, while meeting legal, ethical and operational requirements? That is, what type and degree of human control is required in practice?

Recognizing that human control could form the basis for establishing necessary international limits on autonomy in weapon systems, SIPRI and the ICRC collaborated on a project to generate insights into the practical elements of human control that will need to be at the heart of any policy response by states As part of this project, SIPRI and the ICRC co-hosted a two-day expert workshop in Stockholm in June 2019. Based on discussions at the workshop and additional research, this report presents the key findings and recommendations of this project.

The findings and recommendations are the views of the authors and they do not necessarily reflect the views—consensus or otherwise—of participants in the workshop. Moreover, they do not pre-judge the policy responses that states should choose to regulate AWS.

SIPRI and the ICRC commend this report primarily to government decision makers in the realms of international law, arms control, defence and foreign affairs. This report may also be of relevance for international organizations, non-governmental organizations, industry, researchers and students in the fields of international law, international relations, ethics, politics, and science and technology policy.

Dan Smith
Director, SIPRI
Stockholm, June 2020

Acknowledgements

The authors would like to express their sincere gratitude to the Ministry for Foreign Affairs of the Netherlands, the Ministry of Foreign Affairs of Sweden and the Swiss Federal Department of Foreign Affairs for their generous financial support of the project.

The authors are also indebted to all the speakers and participants who shared their knowledge and experience at the workshop held on 17–18 June 2019 on 'Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control': Peter Asaro, Subashi Banerjee, Sibylle Bauer, Patrick Bezombes, Adolf Brückler, Jenny Burke, Jim Burke, Raja Chatila, Anja Dahlmann, Merel Ekelhof, Ljupco Jivan Gjorgjinski, Martin Hagström, Peter Lee, Larry Lewis, Peng Li, Rain Liivoja, Ulf Lindell, Ian Macleod, David McNeish, Richard Moyes, Libin Niu, Merel Noorman, SeHoon Park, Ludovic Righetti, Heigo Sato, Michael Siegrist, Patrick Stensson, Pauline Warnotte and Albert Yefimov.

The authors wish to thank peer reviewers Rain Liivoja, Merel Noorman, Patrick Stensson, Kathleen Lawand and Maya Brehm for their comprehensive and constructive feedback. Finally, we would like to acknowledge the invaluable work of editor Linda Nix and the SIPRI Editorial Department.

The views and opinions in this report are solely those of the authors and do not represent the official views of SIPRI, the ICRC or the funders. Responsibility for the information set out in this report lies entirely with the authors.

Abbreviations

AI	Artificial intelligence
AWS	Autonomous weapon systems
CCW	1980 United Nations Convention on Certain Conventional Weapons
DCDC	Development, Concepts and Doctrine Centre, United Kingdom
GGE	Group of Governmental Experts
IHL	International humanitarian law
NGO	Non-governmental organization
OOTL	Out-of-the loop
R&D	Research and development
T&E	Testing and evaluation
UNIDIR	United Nations Institute for Disarmament Research
V&V	Validation and verification

Executive summary

The challenges posed by autonomous weapon systems (AWS) are the focus of an intergovernmental discussion under the framework of the United Nations Convention on Certain Conventional Weapons (CCW). Despite enduring disagreements on whether additional regulation is needed, and in what form, there is emerging consensus among states that autonomy in weapon systems cannot be unlimited: humans must 'retain' and 'exercise' responsibility for the use of weapon systems and the use of force in armed conflict. This report explores the difficult question of how that principle must be applied in practice. It offers an in-depth discussion on the type and degree of control that humans need to exercise over AWS, in light of legal requirements, ethical concerns, and operational considerations. It provides policymakers with practical guidance on how measures for human control should form the basis of internationally agreed limits on AWS—whether rules, standards or best practices.

The report is the result of a joint project of the Stockholm International Peace Research Institute (SIPRI) and the International Committee of the Red Cross (ICRC). Chapter 1 introduces the context and conceptual approach. Chapter 2 explores the legal, ethical and operational perspectives on human control. Chapter 3 provides practical guidance on the type, degree and combination of control measures needed for compliance with international humanitarian law (IHL) and to address ethical concerns, while taking into account military operational considerations. Chapter 4 presents the key findings and recommendations for policymakers.

A core problem with AWS is that they are triggered by the environment, meaning the user does not know, or choose, the specific target, timing and/or location of the resulting application of force. This process by which AWS function, and associated unpredictability in the consequences of their use can raises serious risks for civilians and challenges for compliance with IHL, as well as fundamental ethical concerns about the role of humans in life-and-death decisions, and challenges for military command and control.

A key question, therefore, is what limits are needed on AWS to address these challenges. An examination of the legal, ethical and operational requirements for human control indicates the need for a combination of three types of control measures:

1. *Controls on the weapon system's parameters of use*, including measures that restrict the type of target and the task the AWS is used for; place temporal and spatial limits on its operation; constrain the effects of the AWS; and allow for deactivation and fail-safe mechanisms.
2. *Controls on the environment*, namely, measures that control or structure the environment in which the AWS is used (e.g. using the AWS only in environments where civilians and civilian objects are not present, or excluding their presence for the duration of the operation).
3. *Controls through human–machine interaction*, such as measures that allow the user to supervise the AWS and to intervene in its operation where necessary.

These control measures can help reduce or at least compensate for the unpredictability inherent in the use of AWS and to mitigate the risks, in particular for civilians. From a legal perspective, a user must exercise sufficient control to have reasonable certainty about the effects of an AWS when used in an attack and to be able to limit them as required by IHL. Ethical considerations may demand additional constraints, especially given concerns with AWS designed or used against persons.

The report concludes with five recommendations. First, **states should focus their work on determining how measures needed for human control apply in practice**. Since these three types of control measures are not tied to specific technologies, they provide a robust normative basis applicable to the regulation of both current and future AWS.

Second, **measures for human control should inform any development of internationally agreed limits on AWS—whether new rules, standards or best practices**. This work must be guided by the legal, ethical and operational requirements for human control. Any normative development should also focus on human obligations and responsibilities, not on technological fixes, so as to remain relevant and practical, and adaptable to future technological developments.

Third, **states should clarify where IHL rules already set constraints on the development and use of AWS, and where new rules, standards and best practice guidance may be needed**.

Fourth, **any new rules, standards and best practices must build on existing limits on autonomy under IHL, and should draw on existing practice**. It is likely that new rules, standards and best practice guidance can be most effectively articulated in terms of limits on specific types of AWS, of the manner and circumstances of their use and on requirements for human supervision and intervention.

Fifth, **human control criteria should be considered in the study, research and development, and acquisition of new weapon systems**.

1. Introduction

The challenges posed by autonomous weapon systems (AWS) are currently the focus of an intergovernmental discussion under the framework of the United Nations (UN) Convention on Certain Conventional Weapons (CCW).[1] The expert discussion on 'emerging technologies in the area of lethal autonomous weapons systems', which is now in its seventh year (the past four as a Group of Governmental Experts, GGE), made some significant progress in 2018 and 2019.[2] Despite disagreements on whether additional regulation is needed, and in what form, there is an emerging consensus among states that autonomy in weapon systems cannot be unlimited: humans must 'retain' and 'exercise' 'responsibility for the use' of weapon systems, as reflected in the GGE's affirmation in 2019 of 11 guiding principles. Notably, the first four principles reflect agreement that: international humanitarian law (IHL) applies to all weapon systems, including AWS; 'Human responsibility for decisions on the use of weapons systems must be retained since accountability cannot be transferred to machines'; 'Human–machine interaction . . . should ensure that the potential use of AWS . . . is in compliance with applicable international law, in particular IHL'; and 'Accountability for developing, deploying and using any emerging weapons system in the framework of the CCW must be ensured in accordance with applicable international law, including through the operation of such systems within a responsible chain of human command and control'.[3]

The question of how humans should retain and exercise this responsibility in practice remains, however, a central issue of debate. The GGE in 2019 concluded: 'Although there is agreement on the importance of the human element . . ., [f]urther clarification is needed on the type and degree of human–machine interaction required, including elements of control and judgement'.[4] States, international organizations, non-governmental organizations (NGOs), and independent experts have already been made a number of proposals.[5] The proposals may use different terminology—for example, 'meaningful human control', 'appropriate levels of human judgement',

[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, opened for signature 10 Apr. 1981, entered into force 2 Dec. 1983.

[2] CCW GGE (Group of Governmental Experts of the High Contracting Parties to the Convention on the Prohibition or Restriction on the Use of Certain Conventional Weapons Which May Be Deemed Excessively Injurious or to Have Indiscriminate Effects), *Draft Report of the 2019 Session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems*, Geneva, 21 Aug. 2019, CCW/GGE.1/2019/CRP.1/Rev.2.

[3] CCW GGE (note 2), Annex IV, principles (a), (b), (c) and (d).

[4] CCW GGE, CCW/GGE.1/2019/CRP.1/Rev.2 (note 2), para. 22(a)–(b).

[5] ICRC, 'Statement of the ICRC on agenda item 5(a)', GGE, Geneva, 25–29 Mar. 2019; and ICRC, 'Statement of the ICRC on agenda item 5(b)', GGE, Geneva, 25–29 Mar. 2019; Ekelhof, M., 'Autonomous weapons: operationalizing meaningful human control', ICRC Humanitarian Law and Policy Blog, 25 Aug. 2018, Lewis, L., *Redefining Human Control: Lessons from the Battlefield for Autonomous Weapons*, CNA Occasional Paper (Center for Autonomy and AI, CNA: Arlington, VA, 2018); Roff, H. and Moyes, R., M*eaningful Human Control, Artificial intelligence and Autonomous Weapons*, Briefing Paper (Article 36: London, 2016); International Committee of the Red Cross (ICRC), *The Element of Human Control*, Working Paper submitted at the Meeting of High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, Geneva, 21–23 Nov. 2018, CCW/MSP/2018/WP.3, 19 Nov. 2018; International Panel on the Regulation of Autonomous Weapons (IPRAW), *Focus on Human Control*, 'Focus on' Report no. 5 (IPRAW: Berlin, Aug. 2019); Amoroso, D. and Taburrini, G., *What Makes Human Control Over Weapons 'Meaningful'?*, International Committee for Robots Arms Control (ICRAC) Working Paper no. 4 (ICRAC, Aug. 2019); United States Government, 'Human–machine interaction in the development and use of emerging technologies in the area of lethal autonomous weapon systems', Working Paper submitted at the 2018 Meeting of the Group of Governmental Experts of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, Geneva, 27–31 Aug. 2018, CCW/GGE.2/2018/WP.4, 28 Aug. 2018; British Government, 'Human machine touchpoints: the United Kingdom's perpective on human control over weapon development and targeting circles', Working Paper Submitted at the 2018 Meeting of the Group of Governmental Experts of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, Geneva, 27–31 Aug. 2018, CCW/GGE.2/2018/WP.1, 8 Aug. 2018; Australian Government, 'Australia's system of control and applications for autonomous weapon systems', Working Paper submitted at the

'sufficient level of human involvement'—but they all recognize that humans need to exert some form of control over weapons and the use of force in specific attacks in armed conflict. Where they diverge is on questions of how and when humans should exercise that control in operational contexts. As highlighted in the GGE, there are likely to be some control measures that may apply in all circumstances and others whose necessity depends on the context.[6]

Therefore, after plotting the background to the international debate on AWS and an explanation of 'human control' as a conceptual framework (chapter 2, section I), this report addresses the following questions:

1. Why is human control needed and what practical requirements can be derived from legal, ethical and operational perspectives? (chapter 2, sections II–IV).
2. What concrete requirements for human control can be derived from existing IHL rules, ethical considerations and operational practice? (chapter 3, section I).
3. When and how can these requirements for human control be operationalized, or implemented, in practical terms? (chapter 3, sections II–III).

Chapter 4 then presents the key findings of the report, along with recommendations for international discussions, including those taking place at the CCW GGE.

2019 CCW Meeting of Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapon Systems, 26 Mar. 2019, CCW/GGE.1/2019/WP.2, 20 Mar. 2019.

[6] CCW GGE, CCW/GGE.1/2019/CRP.1/Rev.2 (note 2), para 16.

2. The exercise of human control from legal, ethical and operational perspectives

I. A conceptual framework for addressing autonomous weapon systems

This report relies on the concept of 'human control' in order to analyse the challenges posed by AWS and formulate possible responses. Introduced early in the international discussions, this term shifted the focus of the debate from technology and definitions to the legal and ethical role played by humans in the use of force and what this means for autonomy in weapon systems.[7]

This conceptual approach was taken for several reasons. First, because the approach has gained widespread support among GGE participants and has become commonly used over time. A variety of qualifiers and terms have been used by states and in reports of the GGE—for example, appropriate/sufficient/adequate level of human judgement/responsibility/intervention/supervision/involvement/authority in human–machine interaction.[8] However, at some level these terms all express the same recognition of the challenges that AWS raise, and a desire to ensure that the 'human element' is retained in the use of weapons and the use of force.

The second reason for adopting this conceptual approach is because it is pragmatic, offering a way to avoid the pitfalls of a technology-centric definitional approach and to elaborate requirements for human control that will be future-proof and applicable to the ever-evolving technology applicable to AWS.

Third, the approach is sound because it accurately places the focus on the central concern about AWS: loss of control over the use of force. AWS raise the prospect of unpredictable consequences, including for civilians, since they select and attack targets without human intervention based on technical indicators such as interaction with the environment. This means that the users, at the point of activation, do not know the specific target, and the timing and location of the force application(s) that will result.

Despite emerging consensus on the importance of the 'human element' indicated by the guiding principles, debate continues on whether, and how, to capture this principle as a new norm.[9]

This report also uses the expression 'elements of human control', which should be understood as the types of control measures necessary to ensure legal compliance, ethical acceptability and operational utility. They are best understood as any kind of measure that acts as a necessary constraint on AWS and that could form the basis of internationally agreed standards or rules.

[7] Moyes, R. 'Meaningful human control', ed. R. Geiss, *Lethal Autonomous Weapons Systems: Technology, Definition, Ethics, Law and Security* (German Federal Foreign Office: Berlin, 2016); Boulanin, V., *Mapping the Debate on LAWS at the CCW: Taking Stock and Moving Forward*, EU Non-proliferation Paper no. 49 (SIPRI: Stockholm, Mar. 2016). For earlier discussions of the regulation of AWS, see Anthony, I. and Holland, C., 'The governance of autonomous weapon systems', *SIPRI Yearbook 2014: Armaments, Disarmament and International Security* (Oxford University Press: Oxford, 2014), pp. 423–31; Davis, I. 'Humanitarian arms control regimes: Key development in 2016', *SIPRI Yearbook 2017: Armaments, Disarmament and International Security* (Oxford University Press: Oxford, 2017), pp. 559–61; Davis, I and Verbruggen, M., 'The Convention on Certain Conventional Weapons', *SIPRI Yearbook 2018: Armaments, Disarmament and International Security* (Oxford University Press: Oxford, 2018), pp. 383–86; and Boulanin, V., Davis, I and Verbruggen, M, 'The Convention on Certain Conventional Weapons and lethal autonomous weapon systems', *SIPRI Yearbook 2019: Armaments, Disarmament and International Security* (Oxford University Press: Oxford, 2019), pp. 452–57.

[8] See, e.g., CCW GGE, 'Chair's summary of the discussion of the 2019 Group of Governmental Experts on emerging technologies in the area of lethal autonomous weapons systems', Addendum, CCW/GGE.1/2019/3/Add.1, 8 Nov. 2019, paras 17–20.

[9] CCW GGE, CCW/GGE.1/2019/CRP.1/Rev.2 (note 2), Annex IV.

Box 2.1. Key international humanitarian law rules governing the use of means and methods of warfare

Distinction

Parties to armed conflicts must at all times distinguish between civilians and combatants, and between civilian objects and military objectives. Attacks may only be directed against combatants and military objectives, never against civilians or civilian objects.[a] Lawful targets include combatants and civilians directly participating in hostilities, and objects that constitute military objectives.[b]

Prohibition of indiscriminate attacks

Indiscriminate attacks are attacks of a nature to strike military objectives and civilians and civilian objects without distinction, either because: the attacks are not directed at a specific military objective, they employ a method or means of combat that cannot be directed at a specific military objective, or they employ a method or means of combat the effects of which cannot be limited as required by international humanitarian law (IHL).[c]

Prohibition of disproportionate attacks

The rule of proportionality prohibits attacks which, although directed at a military objective, are expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, that would be excessive in relation to the concrete and direct military advantage anticipated.[d]

Precautions in attack

In the conduct of hostilities, IHL requires parties to armed conflicts to take constant care to spare the civilian population, civilians and civilian objects. The obligation to take precautions in attack requires persons who plan, decide on and carry out attacks to:

- do everything feasible to verify that the objectives to be attacked are neither civilians nor civilian objects, and are not subject to special protection but are military objectives
- take all feasible precautions in the choice of means and methods of attack with a view to avoiding, and in any event minimising, incidental loss of civilian life, injury to civilians and damage to civilian objects
- refrain from deciding to launch an attack if it may be expected to cause disproportionate civilian harm, and
- cancel or suspend an attack if it becomes apparent that the objective is not a military one, or is subject to special protection, or that the attack may be expected to cause disproportionate civilian harm.[e]

[a] Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I), adopted 8 June 1977, 1125 UNTS 3, entered into force 7 Dec. 1978, Articles 48, 51(2) and 52(1); ICRC, Customary IHL Study, vol. 1, (Customary IHL), Rules 1 and 7.
[b] Protocol I, Articles 48 API, 51(2) and (3), 52(1) and (2); Customary IHL, Rules 1 and 7.
[c] Protocol I, Article 51(4); Customary IHL, Rules 11–13. The prohibition of indiscriminate attacks is distinct from the prohibition of the use of weapons that are indiscriminate by nature, referred to in Customary IHL, Rule 71.
[d] Protocol I, Article 51(5)(b); Customary IHL, Rule 14 ICRC Customary IHL Study.
[e] Protocol I, Article 57(2)(a) and (b); Customary IHL, Rules 15–19.

II. The legal perspective on requirements for human control

The use of any weapon, including an AWS, as a means of warfare during an armed conflict is governed by IHL rules on the conduct of hostilities, notably, the rules of distinction, proportionality and precautions in attack (see box 2.1).

It is undisputed that AWS must be used, and must be capable of being used, in accordance with IHL. It is the humans subject to IHL who are responsible for applying the law and who can be held accountable for violations, not the weapon itself.[10] The legal requirements under rules governing attacks must be fulfilled by those persons who plan, decide on and carry out attacks, in other words the *users* of an AWS.[11]

How does autonomy in the critical functions of weapon systems impact the ability of humans to apply key rules on the conduct of hostilities? The first part of this section

[10] See ICRC, *International Humanitarian Law and the Challenges of Contemporary Armed Conflicts*, (ICRC: Geneva, Oct. 2019), pp. 29–31; Davison, N, 'Autonomous weapon systems under international humanitarian law', in *Perspectives on Lethal Autonomous Weapon Systems*, United Nations Office for Disarmament Affairs (UNODA) Occasional Papers no. 30 (UNODA: New York, Nov. 2017), pp. 5–18.

[11] The term 'user' in this section of the report refers to a person or group of persons who plan, decide on or carry out attacks. Other writings on the topic of AWS and IHL variously refer to these persons as 'operators' or 'commanders'.

identifies inter-related challenges to the application of IHL posed by the use of AWS. These are grouped into three conceptual categories to facilitate an analysis of how 'human control' (and judgement) is required by IHL. The second part of this section examines divergent views on whether the identified challenges can be resolved through technology, and what concrete and practical measures of human control may be derived from IHL rules, regardless of where one stands on technological capabilities.

Challenges posed by AWS to the application of IHL

The 'numbers' challenge

Applying the IHL rules of distinction, proportionality and precautions in attack require a party to conflict to identify which targets are lawful (i.e. combatants, civilians participating directly in hostilities, and military objectives) and which persons and objects must be protected from attack (e.g. civilians). Such characterizations by their nature involve qualitative, or evaluative, judgements; that is, they require judgements made on the basis of values and interpretation of the particular situation rather than numbers or technical indicators.[12]

The rule of proportionality illustrates this challenge. Neither the incidental civilian harm—loss of civilian life, injury to civilians and damage to civilian objects—expected from an attack nor the anticipated military advantage can be easily quantified. Nor can the relationship between these dissimilar values be fixed in numerical terms. The prohibition on causing excessive civilian harm requires persons to assign values and make judgements that are not purely calculations; in other words, it requires uniquely human judgement. Such value judgements, which also reflect ethical considerations (see chapter 2, section III), are part of the training of armed forces, and are made in a particular context.[13]

The challenge arises because AWS operate and act on the basis of technical indicators, namely pre-programmed target profiles, information about their environment received through sensors, and computer-generated analysis of the data collected from these sensors and applied to the profiles. Whereas persons today may rely on statistical tools to support their decision making, many experts agree that such processes do not in themselves constitute the proportionality assessment and cannot replace the decisions required of persons under the rule of proportionality.[14]

Given that the user of an AWS does not know the exact target, nor its location, nor the timing and surroundings of the application of force against the target, including the presence of civilians or civilian objects at risk of being harmed incidentally, what controls over the weapon and its environment are required to safeguard the user's ability to make the necessary value judgements and avoid undue reliance on technical indicators?

[12] Switzerland refers to this as 'the application of evaluative decisions and value judgements': CCW GGE, 'A "compliance-based" approach to autonomous weapon systems', Working Paper submitted by Switzerland, CCW/GGE.1/2017/WP.9, 10 Nov. 2017, para. 13.

[13] For more detail, see ICRC, *The Principle of Proportionality in the Rules Governing the Conduct of Hostilities under International Humanitarian Law*, Report of an International Expert Meeting, 22–23 June 2016, Quebec (ICRC: Geneva, 2018).

[14] See, e.g., ICRC, *The Principle of Proportionality* (note 13), p. 65. See also ICRC, 'Commentary of 1987: precautions in attack', para. 2208; Sassòli, M., 'Autonomous weapons and international humanitarian law: advantages, open technical questions and legal issues to be clarified', *International Law Studies*, vol. 90 (2014), p. 331; Schmitt, M. N. and Thurnher, J. '"Out of the loop": autonomous weapon systems and the law of armed conflict', *Harvard National Security Journal*, vol. 4, no. 2 (2013), p. 256.

Box 2.2. Determining the context for assessments under international humanitarian law

International humanitarian law (IHL) rules on the conduct of hostilities provide protection against dangers arising from military operations generally, with several of these IHL rules, notably distinction, proportionality and precautions, specifically applying to 'attacks'. These are defined in Article 49(1) of Protocol I to the Geneva Conventions as 'acts of violence against the adversary whether in offence or defence'.[a]

An attack thus provides a relevant context for assessing compliance with these IHL rules. Whether those persons who plan, decide on or carry out an attack can fulfil their obligations under IHL when using an autonomous weapon system (AWS) will depend on the scope of an attack, including its duration. It is therefore critical to determine when an attack in the IHL sense commences and ends. Identifying the point at which an attack using an AWS commences, and its duration, will determine the frame of reference for assessing whether those who plan or decide on the attack have complied with their obligations under IHL.

Divergent views exist on how the term 'attack' should be interpreted in relation to traditional weapons. These come to the fore in the debate on AWS. According to one interpretation, an attack commences whenever an AWS detects or selects a person or object as a target. One objection to this view is that the consequences of an act of 'detecting' and 'selecting' a target need not materialize in order for that act to qualify as an attack.[b]

This report considers that an attack starts when an AWS is activated,[c] because it is at this point that any person or object that fits within the weapon's target profile and is in its area of operations will be 'directly endangered',[d] even if they have not (yet) been detected by its sensors or selected by its software.

[a] See, e.g., Article 51 and Article 57 of Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I), adopted 8 June 1977, 1125 UNTS 3, entered into force 7 Dec. 1978.

[b] Boothby, W., *The Law of Targeting* (Oxford University Press: Oxford, 2012), p. 81; Sassòli, M. *International Humanitarian Law: Rules, Controversies and Solutions to Problems Arising in Warfare* (Edward Elgar: Cheltenham, 2019), para. 8.295; Geiss, R., 'The legal regulation of cyber attacks in times of armed conflict', *Technological Challenges for the Humanitarian Legal Framework: Proceedings of the 11th Bruges Colloquium, 21–22 October 2010*, Collegium no. 41 (College of Europe and the ICRC: Bruges, 2011), p. 52.

[c] See Boothby, W. (note b), p. 282; Sparrow, R., 'Twenty seconds to comply: autonomous weapon systems and the recognition of surrender', *International Law Studies*, vol. 91 (2015), p. 725.

[d] The term is from a 1982 survey conducted by the International Society of Military Law and the Law of War, which noted, when considering the use of mines, members' 'general feeling . . . that there is an attack whenever a person is directly endangered by a mine laid'; cited in fn. 5 of ICRC, 'Commentary of 1987: definition of attacks and scope of application', para. 1881 (on Article 49 of Protocol I).

The context challenge

In order to make the judgements required by IHL rules, users of AWS must make context-dependent and time-bound decisions. The requirement to distinguish between civilian objects and military objectives provides an example. Attacks may only be directed against military objectives; that is, objects which by their nature, location, purpose or use make an effective contribution to military action and whose partial or total destruction, capture or neutralization, in the circumstances ruling at the time, offers a definite military advantage.[15] Adherence to this rule requires an assessment based on knowledge of context (see box 2.2) and users must be able to adapt to changing circumstances. Attacking an object whose destruction no longer offers a definite military advantage would not be permissible. Moreover, in accordance with the obligation to take precautions in attack, users must do everything feasible to cancel or suspend an attack if it becomes apparent that a target is not or is no longer a military objective.

This context dependency presents a challenge in light of the unique characteristics of AWS, where users will not know the specific context (timing, location and surroundings) of the resulting force application. Existing AWS are designed with specific technical indicators that are programmed to determine what objects will be targeted and under what conditions. These programming decisions are made in advance of the commencement of an attack, based on the intended tasks, type of target and anticipated operating environment, and method and circumstances of use.

[15] Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I), adopted 8 June 1977, 1125 UNTS 3, entered into force 7 Dec. 1978, Article 52(2); ICRC, Customary IHL Study, vol. 1, (Customary IHL), Rule 8.

However, in order to comply with IHL, users must take into account factors that vary over time, including between the programming and activation of an AWS, as well as between the activation of an AWS and the eventual selecting and application of force to a target.

The continuing validity of the user's assumptions about the context of the target and its environment when initiating the attack is fundamental to its lawfulness. Given that changes may occur after the commencement of an attack, what controls on the context may be required for users to reasonably rely on their planning assumptions?

The predictability challenge

To comply with IHL rules, notably the prohibition on indiscriminate attacks, the prohibition on indiscriminate weapons, the principle of proportionality and the requirement to take precautions in attack, persons must be capable of limiting the effects of the weapons they use. They can only do this, however, if they can reasonably foresee how a weapon will function in any given circumstances of use and the effects that will result. Predictability is therefore key to ensuring that a weapon's use complies with IHL.

All AWS raise some concerns about unpredictability. Even the functioning and effects of so-called rules-based systems (sometimes called 'deterministic' systems, where the potential inputs and resulting outputs of the system are determined by specific rules and fixed at the point of design; e.g. 'if x happens, do y') will be challenging to predict. This is because the consequences of any output will vary depending on the circumstances in the environment at the time of the attack.[16] In the case of an AWS, it will apply force at a specific time and place unknown to the user when they activated the AWS. Moreover, the environment in which the AWS is operating may vary over time. Particular difficulties arise where the status and surroundings of the target may change swiftly or frequently (e.g. when a house ceases to be a military objective once it is no longer used by enemy armed forces), or where sudden changes may occur around a target (e.g. when civilians have moved into its immediate vicinity).

A user's ability to predict and limit the effects of a weapon also depends on the weapon's design. For example, AWS that rely on machine learning to adapt their functioning—specifically the selection and application of force to targets after activation, based on interaction with the operational environment—have been called 'unpredictable by design'. Could the user of such a weapon be reasonably certain about the manner in which it will function in the circumstances of use; that is, will they be able to practicably limit the weapon's effects of as required by IHL? In light of the predictability challenges raised by AWS, what controls are needed on the context (environment) of use, the way in which AWS are used and on their design?

Application of IHL in practice: what requirements for human control?

As previously mentioned, the key feature of weapons that autonomously select and apply force is that the user will not know the exact target that will be struck, nor its location and surroundings, nor the timing and circumstances of the application of force. There are consequently significant difficulties in using AWS in a manner that retains the user's ability to reasonably foresee (predict) the effects of the weapon in the circumstances of use and to make the context-specific value-based judgments required by IHL rules.

[16] Boulanin, V. and Verbruggen, M., *Mapping the Development of Autonomy in Weapon Systems* (SIPRI: Stockholm, 2017), pp. 9–11; Righetti, L., 'Emerging technology and future autonomous systems: speakers summary', *Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons*, Report of the Expert Meeting, Versoix, Switzerland, 15–16 Mar. 2016 (ICRC: Geneva, 2016), p. 39.

However, there are different views on whether the above-mentioned three interrelated challenges—numbers, context and unpredictability—can be overcome through technological fixes, or whether they are markers of the limits of autonomy under IHL? In any case the question remains: what type and degree of human control is required under IHL?

Contributions made by experts on these questions during the June 2019 workshop can be grouped into two broad schools of thought. Those who adopt a technology-orientated approach argue that AWS could one day be designed and programmed in ways that overcome these challenges.[17] To proponents of this approach, the only limit to autonomy is technology: the more 'sophisticated' the weapon, the more tasks it can be assigned, and the less user control will be needed during an attack. Having decided to use an AWS in a specific attack in accordance with the IHL requirement to take all feasible precautions in the choice of means and methods of warfare, users need do no more than take all feasible precautions to avoid or minimize incidental civilian casualties and damage to civilian objects.

Those who view these three challenges as indicators of the limits of autonomy permissible under IHL argue that human control is not antithetical to autonomy in weapon systems: the two co-exist.[18] According to this approach, no matter the technical characteristics of the AWS, IHL rules on the conduct of hostilities demand contextual and value-based judgements by persons who plan, decide on and carry out attacks. To compensate for the unpredictability introduced when using AWS, users would need to apply strict control measures so as to ensure that the attack complies with IHL.

In short, for those who see technological fixes, practical measures are a response to technological limitations that may in future be overcome. For those who see implied limits on autonomy under IHL, practical measures are necessary for ensuring that users comply with their obligations under IHL and for preserving the role of humans in making context-specific value judgements, no matter the technical characteristics of the weapon systems they use.

These two different approaches to the role of the human (and, concomitantly, of the technology) in the application of IHL rules are not wholly incompatible. In practice, the proponents of the two approaches agree on three practical measures that, at a minimum, would be expected of users in relation to an attack, namely control: (*a*) over the parameters of use of the AWS; (*b*) over the environment of use of the AWS; and (*c*) through human–machine interaction. The subsections below give some indication of the types of control measures that are needed, but further study may be needed.

Control over the parameters of use of the autonomous weapon system

As stated above, IHL requires the users of weapons to direct attacks only against combatants and military objectives, never against protected persons and objects. It prohibits indiscriminate and disproportionate attacks, and requires all feasible precautions to avoid or in any event minimize incidental civilian harm. In practice, controls on the parameters of use of an AWS that could practically facilitate a user's application of these rules include constraints on the type of targets an AWS may select, as well as, potentially, constraints aimed at preventing strikes on or in the vicinity of certain protected objects (e.g. ambulances or hospitals).

In certain circumstances, it may be reasonable for a user to assume, at the moment of activating an AWS, that all objects that conform to the target profiles and other

[17] See, e.g., Schmitt and Thurnher (note 14).

[18] See, e.g., Brehm, M., *Defending the Boundary: Constraints and Requirements on the Use of Autonomous Weapon Systems under International Humanitarian and Human Rights Law*, Briefing no. 9 (Geneva Academy: Geneva, 2017).

technical indicators of the AWS will indeed be military objectives and that the attack would not otherwise fall foul of IHL in the context of use. This could be the case where the target parameters are limited with the intent to only capture objects that are by their nature military objectives (e.g. military naval vessels or incoming missiles) and that there is reasonable certainty that the nature of the target will not change for the duration of the weapon's operation. However, in relation to targets that are not military objectives by nature (e.g. a bridge), this assumption could not remain valid throughout the weapon's operation without concomitant real-time controls on the environment of use.

Given the dynamic and complex nature of most combat situations today, it is difficult to see how controls on weapon parameters alone will suffice to ensure compliance with IHL. Such control measures would not be a substitute for the contextual value-based judgements that IHL requires of persons; at best, they may supplement or support human judgement. Consequently, limits on the types of environments in which the AWS can operate, and real-time human supervision of the weapon's operation, will be needed to ensure compliance with IHL.[19]

Control over the environment of use

The challenge of predicting when, where and to what an AWS may apply force means the user of an AWS faces particular difficulties in making the context-dependent value-based judgements required under IHL. For example, their proportionality assessment is based on planning assumptions and information known about the weapon's environment of use. Spatial and temporal limitations on the operation of the AWS may help the user to exert control over the environment of use. The duration and area of operation permissible under IHL will, however, depend on a range of factors, including the nature of the AWS and its location. In practice, limiting AWS use to locations where there are no civilians present could facilitate the user's ability to reasonably rely on planning assumptions and trust in their continued validity during the weapon's operation. By contrast, where targets may be located in a concentration of civilians or civilian objects, where circumstances may change quickly, it would not be possible for the user to foresee the effects of the attack using AWS by relying only on the planning assumptions. This loss of predictability would need to be compensated for by other control measures such as spatial and temporal limits and real-time supervision for the duration of the weapon's operation.

Control through human–machine interaction

In situations where there is a clear risk that circumstances on the ground could change after the activation of an AWS and invalidate the planning assumptions—for example, the risk that the status of the target or that the environment around it changes—users would need to retain the ability to supervise and intervene in—including to deactivate—the operation of an AWS during the course of an attack to ensure compliance with IHL. Human supervision would therefore seem to be required in all but the most static and predictable targeting and conflict environments.

There remain many questions regarding what additional concrete measures of human–machine interaction derive from IHL, and how they would be implemented in practice. Given the technical challenges in exercising effective human control in AWS through human–machine interaction (described in section IV below), presently

[19] For example, the ICRC has stated that 'Human control at the development stage alone—control in design—will not be sufficient to ensure compliance with IHL in the use of an autonomous weapon system for attacks in armed conflict given the inherently variable and unpredictable nature of real-world operational environments.' ICRC, 'Statement of the ICRC on agenda item 5(a)' (note 5).

it is clear that the speed of a weapon's functioning, as well as its duration and area of operation, need to have limitations.

It is against the background of the challenges discussed in this section that the ICRC has raised serious doubts about the ability of AWS to be used in compliance with IHL in most circumstances, particularly in dynamic and unpredictable environments, over large geographic ranges, or for long periods.[20]

III. The ethical perspective on requirements for human control

Ethical considerations are central to the debate about the use of AWS because it is 'precisely anxiety about the loss of human control over weapon systems and the use of force' that broadens the issue from simply compliance with the law to wider questions of acceptability to ethical standards and social values.[21] Many states have raised ethical concerns about AWS, including during meetings of the CCW since 2014.[22] Others who have raised such concerns include the UN Secretary-General, who said 'machines with the power and discretion to take lives without human involvement are politically unacceptable, morally repugnant and should be prohibited by international law'; a UN Special Rapporteur at the Human Rights Council, who stated that 'there is widespread concern that allowing LARs [lethal autonomous robots] to kill people may denigrate the value of life itself'; Human Rights Watch (part of the Campaign to Stop Killer Robots), which notes 'for many the thought of machines making life-and-death decisions previously in the hands of humans shocks the conscience'; leaders in the scientific and technical communities; the United Nations Institute for Disarmament Research (UNIDIR); and the ICRC, which stated that 'there is a sense of deep discomfort with the idea of any weapon system that places the use of force beyond human control'.[23]

Ethical considerations have often preceded and motivated the development of new international legal constraints on means and methods of warfare, including constraints on weapons that pose unacceptable risks for civilians.[24] This link between ethics and IHL is also found in the Martens Clause, a provision that first appeared in the Hague Conventions of 1899 and 1907, was later incorporated in the 1977 Additional Protocols to the Geneva Conventions, and is considered customary law.

[20] ICRC, *The Element of Human Control* (note 5); Davison (note 10); ICRC, *International Humanitarian Law and the Challenges of Contemporary Armed Conflicts* (note 11), p. 45.

[21] ICRC, *Ethics and Autonomous Weapon Systems: An Ethical Basis for Human Control?* Report (ICRC: Geneva, Apr. 2018), Executive Summary, para. 5. This section is largely based on that report.

[22] These states include Algeria, Argentina, Austria, Belarus, Brazil, Cambodia, Costa Rica, Cuba, Ecuador, Egypt, France, Germany, Ghana, the Holy See, India, Kazakhstan, Mexico, Morocco, Nicaragua, Norway, Pakistan, Panama, Peru, Republic of Korea, Sierra Leone, South Africa, Sri Lanka, Sweden, Switzerland, Turkey, Venezuela, Zambia and Zimbabwe. For a record of discussions, including interventions by states, see CCW, 'Discussions on emerging technologies in the area of LAWS'.

[23] UN Secretary-General, 'Secretary-General's message to Meeting of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems', 25 Mar. 2019, para. 2; Human Rights Council, *Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions, Christof Heyns*, A/HRC/23/47, 9 Apr. 2013, para. 109; Human Rights Watch, Losing Humanity: The Case against Killer Robots, Report, 19 Nov. 2012, Conclusion; ICRC, 'Autonomous weapon systems: is it morally acceptable for a machine to make life and death decisions?', ICRC statement, 13 Apr. 2015; United Nations Institute for Disarmament Research (UNIDIR), The Weaponization of Increasingly Autonomous Technologies: Considering Ethics and Social Values, UNIDIR Resources no. 3 (UNIDIR: Geneva, 30 Mar. 2015). For examples of leaders in the scientific and technical communities see, e.g, Future of Life Institute, 'Autonomous weapons: an open letter from AI & robotics researchers', 28 July 2015; and Future of Life Institute, 'An open letter to the United Nations Convention on Certain Conventional Weapons', 21 Aug. 2017.

[24] ICRC, *Ethics and Autonomous Weapon Systems* (note 21), p. 5.

The Martens Clause provides:

> In cases not covered by this Protocol or by any other international agreements, civilians and combatants remain under the protection and authority of the principles of international law derived from established custom, from the principles of humanity and from dictates of public conscience.[25]

The ICRC notes that the clause 'prevents the assumption that anything that is not explicitly prohibited by relevant treaties is therefore permitted'.[26]

Ethical approaches to human control

Ethical considerations are a strong driver for limiting autonomy in weapon systems and can provide guidance on the necessary character of human control. While ethical arguments can reflect different emphases on the results of an action (consequentialist ethics) as opposed to the action or process itself (deontological ethics), both of these approaches provide useful insights.[27]

Results-driven approaches

Consequentialist, results-driven approaches to human control tend to focus on the likely consequences of the use of AWS, in particular on whether such use will result in greater or fewer risks for civilians.[28] An overall results-driven concern with AWS stems from the unpredictability in the consequences of their use, which presents potential risks to civilians and civilian objects that are present. However, additional considerations of human control may be scenario-driven; that is, the determination of how to maintain human control is based on an assessment of whether that control will increase or decrease the likelihood of the desired consequences in a given circumstance.

The desired consequences or results—the ethical drivers—of a given action, such as the use of an AWS for a specific attack, may be different for different actors, such as the military forces carrying out the attack and those affected by armed conflict. Military forces might see the decision to use an AWS as an ethical obligation to reduce the risks to their forces while also giving them an advantage in a battle. But the primary ethical driver for those affected—such as civilians and humanitarian organizations—will be to avoid risks for civilians. The same actors may also have different ethical drivers that can be difficult to reconcile with one another, such as military forces seeking a tactical advantage and minimizing risks for themselves while also minimizing risks for civilians. In such cases, the approach taken to human control may depend on which ethical driver of consequences is allocated the most value in a specific situation.

Action- or process-driven approaches

Deontological approaches, based on the ethics of an action or process, place emphasis on the duties governing the human role in the use of force and the rights of those against whom force is used, rather than solely on the consequences of the use of an AWS in a specific circumstance. Here the central concern with AWS is about delegating 'life-and-death' decisions, and ethical considerations centre around three interrelated duties and rights for decision making on the use of force: human agency, moral responsibility and human dignity.

[25] Protocol I (note 15), Article 1(2); Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of Non-International Armed Conflicts (Protocol II), adopted 8 June 1977, 1125 UNTS 609, entered into force 7 Dec. 1978, Preamble.

[26] ICRC, *Ethics and Autonomous Weapon Systems* (note 21), p. 6.

[27] ICRC, *Ethics and Autonomous Weapon Systems* (note 21), pp. 7–9.

[28] ICRC, *Ethics and Autonomous Weapon Systems* (note 21), pp. 8–9.

Under this approach, the process of making a decision on the use of force must preserve the causal link between the intention of the human using a weapon system and the consequences of their decision to use it in a specific situation. Human agency in that decision-making process is required to uphold moral responsibility for that decision in recognition of the human dignity of those affected, whether combatants or civilians. Upholding moral responsibility entails being answerable to others about the decision to use force and its consequences, being able to interrogate and articulate the reasons for that decision, and for others to be able to contest those reasons. It requires active engagement with the decision-making process.[29]

Applying this approach to human control, the overall ethical driver is to ensure that the process of making life-and-death decisions reflects human intentions and to avoid effectively delegating these decisions to machines or algorithms. The core ethical argument is that humans must remain the decision makers since only humans can engage in moral reasoning. Machines—including AWS—remain tools: inanimate objects with no ethical, moral or legal responsibility.

Unpredictability and distancing

AWS raise a serious ethical challenge for decision making on the use of force, from both a results-driven and an action- or process-driven perspective, for two main reasons: the consequences of the use of AWS are always unpredictable to a degree; and the connection between intention and consequences—that is, human agency in the specific decision to use force—is diluted or removed.[30] The use of AWS puts a degree of distance—in time, space and understanding—between the human decision to use force and the consequences.

An AWS that applies force to a target based on a target profile, without human intervention, introduces unpredictability at the point of the decision to use or activate the AWS, about who or what will be attacked, where and when the use of force will take place, and the specific consequences that will result. This unpredictability is a type of distancing in understanding, what might be called 'cognitive distancing', which raises ethical concerns from both a results-driven and action- or process-driven perspective. It is fundamentally different from the use of non-autonomous weapons, where the weapon user chooses the specific target, and the time and location of the application of force, although there may still be unpredictability stemming from the 'fog of war'.

Cognitive distancing arises from, among other sources, unpredictability about the consequences of using an AWS in the period between its activation and its eventual application of force—which could be hours, days, weeks, even months later (temporal distancing)—and uncertainty about the location within an area at which the force will be applied (spatial distancing). Generally, greater temporal and spatial distancing is likely to lead to greater unpredictability in consequences, although this will depend on the type of AWS and the type of environment of use.

For AWS that have control software incorporating artificial intelligence (AI), and especially machine learning, there is also a risk of unpredictability by design, where the process by which the software functions is neither predictable nor explainable, in effect a 'black box'.[31] This additional layer of unpredictability at the design level is a further, and fundamentally problematic, form of cognitive distancing.

Cognitive distancing—driven by temporal and spatial distancing and, in the case of AI and machine learning-based AWS, design-level explainability—sits alongside physical

[29] ICRC, *Ethics and Autonomous Weapon Systems* (note 21), pp. 9–13.

[30] ICRC, *Ethics and Autonomous Weapon Systems* (note 21), pp. 11–13.

[31] ICRC, *Ethics and Autonomous Weapon Systems* (note 21), pp. 15–16; ICRC, *Autonomy, Artificial Intelligence and Robotics: Technical Aspects of Human Control* (ICRC: Geneva, Aug. 2019), pp. 13–19.

distancing, which has also been raised by some as an ethical concern in relation to long-range weapons and remotely controlled armed drones. Here the ethical concern is that the decision maker is physically very remote from the situation of the attack; this physical remoteness may also be relevant to the use of AWS. Collectively, these dimensions of distancing may also lead the user of an AWS to feel less responsible because the link between their actions and the consequences are less clear, a concern that can manifest in the 'moral buffer'—where the user transfers responsibility for the consequences to the machine.[32]

While results-driven ethical concerns about unpredictable consequences apply to all AWS to varying degrees, action- or process-driven concerns about removing or diluting human agency in decisions to use force—and interrelated issues of moral responsibility and human dignity—tend to be most acute with AWS used to attack humans. As then UN Special Rapporteur on extrajudicial, summary or arbitrary executions, Christof Heyns, stated in an intervention at a meeting of the CCW:

> to allow machines to determine when and where to use force against humans is to reduce those humans to objects; they are treated as mere targets. They become zeros and ones in the digital scopes of weapons which are programmed in advance to release force without the ability to consider whether there is no other way out, without a sufficient level of deliberate human choice about the matter.[33]

The concern expressed here has two important components. One is the fact of encoding human beings as targets, of equating human life to mere data points, which would be an inherent property of anti-personnel AWS.[34] For many governments and civil society organizations raising ethical concerns, this is unacceptable from a process-driven perspective. The lack of human agency in the specific decision to kill or injure means there is no recognition of the target's humanity, undermining the human dignity of both those making the decision and those who are targeted, and so contributes to the dehumanization of warfare. Similar ethical concerns could also extend to the use of AWS to attack objects, not human beings, in that they would pose risks to people inside or in the vicinity of target objects (such as buildings, vehicles or aircraft).[35]

The second component is the removal of the possibility for the person using the AWS to exercise discretion, and ultimately flexibility or adaptability in decisions to use force. This uniquely moral agency enables those persons to exercise restraint or mercy, even when that course of action (restraint/mercy) is not strictly required for them to comply with the law. An AWS applies force when the data received as input from its sensors matches the parameters of the target profile. This, in effect, removes the user's discretion to not shoot, unless there is a person directly supervising the AWS who has the time available and a mechanism with which to intervene and override or veto a force application. Even if an AWS, theoretically and for example, was programmed not to fire if it detected the digital representation of a white flag, human flexibility and ability to adapt decisions on the use of force to the specific circumstances would still be lost.

[32] Cummings, C., 'Automation and accountability in decision support system interface design', *Journal of Technology Studies*, vol. 32, no. 1 (2006), p. 29: 'decision support systems that integrate higher levels of automation can possibly allow users to perceive the computer as a legitimate authority, diminish moral agency, and shift accountability to the computer, thus creating a moral buffering effect'.

[33] Heyns, C., 'Autonomous weapon systems: human rights and ethical issues', Presentation to the Meeting of High Contracting Parties to the Convention on Certain Conventional Weapons, Geneva, 14 Apr. 2016.

[34] Article 36, *Target Profiles*, Discussion Paper (Article 36: London, Aug. 2019), p. 6; Brehm, M., 'Targeting people: key issues in regulation of autonomous weapons systems', Policy Note (Article 36: London, Nov. 2019). This characteristic can also be seen as a property of certain prohibited weapons with autonomous functions, such as anti-personnel mines. These concerns might also arise in relation to non-autonomous weapons if they are used where there is overreliance on data points, such as a mobile phone signal, in a decision about whether to attack a person.

[35] ICRC, *Ethics and Autonomous Weapon Systems* (note 21), p. 22.

Ethics in practice

So what does human control that reduce or compensate unpredictability while ensuring human agency in decisions to use force look like in practice?

A strict implementation of the action- or process-driven approach would require a human to engage in affirmative reasoning about every use of force, which would generally rule out AWS altogether, at least where their use poses dangers to human beings and their property (perhaps with the exception of uninhabited targets that are purely military in nature, such as missiles, rockets or drones).[36] A related ethical question remains about whether human approval or veto of a machine-generated course of action to apply force would constitute sufficient engagement in the decision, and under what circumstances.

Nevertheless, fundamental ethical concerns do appear to be heightened in situations where AWS are used to target humans, and in situations where there are incidental risks for civilians (though such concerns could also be raised in relation to inhabited military targets, such as military aircraft, vehicles and buildings).

Results-driven ethical demands to reduce unpredictability in the consequences of using AWS through specific measures for human control could also satisfy some process-driven concerns, though the extent would depend on the composition of the measures. Such measures might include: controls on the parameters of use of the AWS, such as limits on targets (e.g. exclusion of human targets), limits on system design (e.g. excluding systems based on AI or machine learning, which are unpredictable or unexplainable by design), and limits on the time and space of operation; controls on the environment of use, such as avoiding or excluding the presence of civilians or civilian objects; and controls on human–machine interaction, such as requiring human supervision of the AWS with the ability to intervene and deactivate the AWS during an attack.[37]

A related question is how such ethics-driven measures for human control could be applied to a system of decision making by multiple individuals leading up to the use of an AWS (see chapter 3, section I). From an ethical perspective, such distributed control may exacerbate core ethical concerns about the dilution or erosion of moral responsibility in decision making on the use of force.[38]

There is of course a parallel between ethical and legal drivers for human control. IHL rules on the conduct of hostilities reflect both results-driven and process-driven approaches, since these rules aim to minimize adverse consequences while requiring combatants to take an active role in applying those rules in a context-specific manner and to ensure accountability (see section II in this chapter, especially box 2.2). At their core, IHL rules that protect civilians and persons *hors de combat* during armed conflict reflect a recognition of human dignity, and of ethical considerations in decisions to use force.

IV. The operational perspective on requirements for human control

The need to exercise human control does not only derive from legal and ethical considerations. This section explores how the exercise of human control is also necessary for ensuring the safety and efficiency of military operations. It also sets out a number of lessons to be learned from current and past uses of AWS in military operations with regard to the types and degrees of human control that may be deemed

[36] See, e.g., Sharkey, N., 'Staying in the loop: human supervisory control of weapons', in N. Bhuta et al. (eds), *Autonomous Weapons Systems: Law, Ethics, Policy* (Cambridge University Press: Cambridge, UK, 2016).

[37] ICRC, *Ethics and Autonomous Weapon Systems* (note 21), pp. 17–19.

[38] ICRC, *Ethics and Autonomous Weapon Systems* (note 21), pp. 11–13.

appropriate to address these specific objectives. The focus here is not compliance with IHL, although that is a key requirement for military efficiency; rather, the aim is to understand how other considerations—such as safety of friendly personnel, operational viability and success of operation—guide the requirement for human control.

Military reasons for maintaining human control: the military ethos, safety and efficiency

Military ethos and desire to maintain human control

Control over the use of a weapon is a fundamental element of the military ethos. The historical constant has been that military institutions have leveraged technology to exert greater, not less, control over the use of force. In recent times, technology has allowed higher military command to get closer and closer to the battlefield.[39] A number of military experts have pointed out during GGE discussions that, from an operational standpoint, the military does not believe in the utility of a so-called 'fully AWS': users would always want to exercise some form of control over its functioning and actions.[40]

According to that narrative, the purpose of pursuing advances in autonomy is not to reduce the role of the human but to change it by creating new models of human–machine collaboration in which the capabilities of both can complement each other more effectively, to allow the conduct of military operations in compliance with the law, but also in accordance with the other essential military objectives of safety and military efficiency.[41]

Safety and reliability as military requirements for human control

Ensuring operational safety and reliability is an important reason for the military's keenness to maintain human control over AWS, notably through some form of human supervision. To date, technical challenges have presented an obstacle to achieving the levels of safety and reliability demanded in military operations.

Autonomous technologies *in general* have made great strides in recent years, but the state of the art, while impressive, still trails by a wide margin the popular perception of what advanced autonomous systems ought to be able to do in a military context; namely, operate safely and reliably in complex and uncertain adversarial environments.[42] The 'perception' and 'decision-making' capabilities of autonomous systems remain, in many regards, too limited.[43] Autonomous systems may 'outperform' humans at simple perceptual tasks, such as recognizing people or objects, but only in narrow circumstances—for example, when going through an online database of pictures.[44] They still struggle with higher cognitive tasks such as recognizing intent or context (see box 2.3).[45] With regard to decision making, autonomous systems can carry out calculations much faster than humans, but they still have major difficulties

[39] Singer, P. 'Tactical generals: leaders, technology and peril', Brookings, 7 July 2009.

[40] French Government, 'Human–machine interaction in the development and use of emerging technologies in the area of lethal autonomous weapon systems', Working Paper submitted at the 2018 CCW GGE meeting, Geneva, 27–31 Aug. 2018, CCW/GGE.2/2018/WP.3, 28 Aug. 2018.

[41] United States Air Force (USAF), Office of the Chief Scientist, *Human–Autonomy Teaming*, vol. 1 of *Autonomous Horizons: Systems Autonomy in the Air Force—A Path to the Future*, vol. 1: Human–autonomy teaming (USAF: Washington, DC, 2015).

[42] Boulanin and Verbruggen (note 16), pp. 12–16, 65–66.

[43] In this subsection, terms such as *autonomy*, *autonomous systems*, *perception*, *decision making* and *understanding* refer to technical characteristics of machines and not human cognition. For example, autonomy in robotics has nothing to do with free will, but relates to a machine's ability to accomplish complex tasks without human intervention, perception refers to a machine's ability to match patterns in sensory input with pre-recorded data, while decision making refers to a machine's ability to select a course of action based on pre-defined objectives and selection criteria.

[44] Boulanin and Vergruggen (note 16), pp. 15, 24–25.

[45] Boulanin and Verbruggen (note 16), p. 15.

Box 2.3. Understanding the limitations of computer vision technology

For Andrey Karparthy, Senior Director of Artificial Intelligence at Tesla and deep-learning and computer vision expert, the above photo provides a striking illustration of the limitations of computer vision. For a human, it takes less than a few seconds to process the 'huge amount' of details in the picture and understand what is happening: former US President Barack Obama is playing a prank and the people watching find it funny. State of the art computer vision technology would not be able to grasp the full sense of the picture. Current technology can recognize that people in the picture are smiling, and potentially that some of them are reflected in the mirror (i.e. they are not two different people), but it cannot understand the reasons for the smiles. This is because full understanding requires a lot of implicit knowledge (e.g. people are self-conscious about their weight) that is really hard to code. Enabling computers to infer meaning from the interactions of people and objects remains a fundamental research problem.

Source: Andrey Karparthy, 'The state of computer vision and AI: we are really far away', Andrey Karparthy Blog, 22 Oct. 2012.

inferring general rules from single real-life cases.[46] In practice these limitations mean that autonomous systems could be easily tricked and defeated by an intelligent adversary, and are unable to adapt to novel situations that the programmer did not foresee and plan for at the design stage.

These limitations of the underlying technology fundamentally restrict the number of situations in which AWS without direct human supervision could currently be used safely and reliably, from an operational standpoint. These restricted situations are environments that are known in advance, highly predictable or controllable, and combat scenarios that are unambiguous (ambiguous scenarios include people or objects entitled to protection under IHL, but also friendly forces that could be mistaken for legitimate military targets) or do not allow an adversary to deploy decoy and deception tactics such as spoofing.

Technological advances in the field of AI and machine learning hold some promise that 'perceptual' and 'decision-making' capabilities of AWS might be improved. However, as discussed in section II of this chapter, the unpredictability introduced by autonomy is likely to remain a challenge to achieving the levels of safety and reliability demanded in military operations, no matter the technical capabilities of the system. This is because predictability—particularly with respect to the consequences of use—is a key component of safe and reliable operations, allowing operators to plan for and

[46] Boulanin and Verbruggen (note 16), p. 15.

mitigate risks to military personnel. In most foreseeable combat scenarios, a human would need to remain, at a minimum, in a supervisory role to intervene and handle situations that the AWS is not able to understand or was not programmed to address.

Military efficiency and the need for human control

Human control is not only needed to overcome the limitations of AWS technology and ensure safety in unexpected circumstances; it is also needed to make sure the actions of the individual AWS operate in line with the predominant military strategy and context (and of course legal obligations).[47] Humans possess certain cognitive capabilities that autonomous technologies do not have—and arguably may never acquire—that allow humans to understand political contexts, changing tactics, and overarching goals and strategies (see box 2.3). Humans can make qualitative judgements, apply knowledge-based reasoning, and think reflectively about the consequences of their actions and how they might be perceived or anticipated by an adversary.[48]

This is the reason for much modern military insistence, in strategic documents related to AI and autonomous systems, on the concept of 'human–machine teaming'. For example, the British Ministry of Defence think tank, the Development, Concepts and Doctrine Centre (DCDC), explained that the best way to leverage military advantage with AI is to combine its strengths with those of humans: 'Developing the right blend of human–machine teams—the effective integration of humans and machines into our war fighting systems—is the key'.[49] This teaming can take multiple forms depending on the types of systems and tasks, but the general idea is to combine computing capability to carry out complex mathematical computations with human ability to exercise qualitative judgements necessary for legal compliance and adapting to complex and potentially new situations.[50]

Exercising human control: lessons from existing autonomous systems

The narrative on human–machine teaming indicates that the military is keen to ensure humans will retain some agency and exert control over autonomous systems. The question then is how that control would work in practice.

Autonomy in existing weapon systems and human control

As a starting point, it is useful to remember that modern militaries have been using AWS for highly constrained tasks in narrow circumstances for decades, and so the question of maintaining human control over AWS is not a new challenge for the military. All existing AWS, both defensive and offensive, operate under some form of human control.[51] From a military standpoint, the unpredictability introduced by AWS are reduced or compensated through one or more of three methods: strict control

[47] Cooke, N. J. and Chadwck, R. A., 'Lessons learned from human–robotic interaction on the ground and in the air', eds M. Barnes and F. Jentsch, *Human–Robot Interactions in Future Military Operations* (CRC Press: Boca Raton, FL, 2010), p. 358; Barnes, M. J. and Evans III, A. W., 'Soldier–robot teams in future battlefields: an overview', eds Barnes and Jentsch (above), p. 11; Hawley, J. K., *Patriot Wars: Automation and the Patriot Air and Missile Defense System*, Ethical Autonomy Series (Center for a New American Security: Washington, DC, 25 Jan. 2017), pp. 9–10; Reason, J., *Human Error* (Cambridge University Press: Cambridge, UK, 1990), p. 182.

[48] Reason, J., 'Safety paradoxes and safety culture', *Injury Control and Safety Promotion*, vol. 7, no. 1 (2000), p. 176; Barnes and Evans III (note 47), p. 10.

[49] British Ministry of Defence, Development, Concepts and Doctrine Centre (DCDC), *Human–Machine Teaming*, Joint Concept Note no. 1/18, p. 16.

[50] DCDC (note 49), p. 42.

[51] The following discussion is based on the following comprehensive studies of existing autonomous weapon systems by SIPRI and the ICRC. See Boulanin and Verbruggen (note 16), pp. 36–55; and ICRC, *Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons*, Report of the Expert Meeting, Versoix, Switzerland, 15–16 Mar. 2016 (ICRC: Geneva, 2016).

measures on the weapon system itself, controlling its environment of use, and human–machine interaction.

All existing AWS are designed to allow a human to restrict the parameters of the weapon's operation. A human determines where, when and for how long the system can operate; the types of targets; and under what conditions the weapon can attack a target autonomously. These parameters are strictly defined for all existing AWS. Typically, an AWS can only apply force to predefined material targets in situations where the time-frame for the 'targeting cycle' (i.e. from target identification to application of force) is deemed too short for human capabilities, such as in defence against incoming missiles or other projectiles. Some AWS also include fail-safe mechanisms that guarantee they will deactivate themselves, cancel their mission or self-destruct when they end up in situations beyond that allowed by their programming.

The second method through which the military has historically exerted human control over existing AWS is through the application of control measures on the environment of use. In the case of landmines—arguably one of the oldest and crudest forms of autonomous weapons—control measures in the form of human supervision are impossible, so control measures designed to minimize the risk to civilians and friendly forces include warning signs and physical landmarks to indicate the presence of landmines, and fencing off mined areas to create de facto no-go zones. Ultimately these measures have proved ineffective at reducing the risks for civilians, and anti-personnel mines were prohibited in 1997 by the Ottawa Treaty.[52] For more technologically advanced weapons, the military has relied on similar approaches to controlling the environment. For example, air-defence systems, radar and radio-communications are used to indicate to civilian aircraft and ships that they are excluded from the area of operation of an AWS for the duration of its activation.

The third method the military uses for exercising control is through human–machine interaction. All existing AWS are operated under direct human supervision. Existing air-defence systems that can identify, select and apply force to targets autonomously allow human supervisors to exert different degrees of control depending on the operational circumstances. Active protection systems, designed to protect vehicles against incoming anti-tank missiles or rockets, function autonomously but allow the human operator to deactivate the system if the environment or circumstances make its use problematic. Robotic 'sentry' weapons (e.g. gun turrets that can autonomously detect, track and potentially apply force to a target) are operated under direct human control. Currently, once the 'sentry' system detects a target, it passes control to the human command and control centre for a human decision on whether to fire. Most loitering munitions, the only real form of offensive AWS deployed today, are also intended to operate under human supervision either through the use of onboard sensors and communication devices that allow the user to monitor and intervene, or through the use of external surveillance systems that monitor the operation systems and the environment of use.

Human–machine interaction: lessons from autonomous systems in general

One of the key lessons from past use of autonomous systems in both military and civilian contexts is that exercising effective human control, through some form of direct human–machine interaction, can be challenging.[53]

All human control or no human control is not common when it comes to human–machine interaction: there is usually *some* level of control by humans. Typically,

[52] Convention on the Prohibition of the Use, Stockpiling, Production and Transfer of Anti-Personnel Mines and on their Destruction (APM Convention), opened for signature 3 Dec. 1997, entered into force 1 Mar. 1999.

[53] DCDC (note 49), p. 42.

autonomy has not removed the role of human controller but changed it from operators of the machine to supervisors of the machine's operations. However, it is now well documented that humans are by default bad supervisors of autonomous systems for the simple reason that human attention is neither constant nor consistent, and that autonomy is a factor that contributes to reduced situational awareness.[54] Researchers in human–machine interaction talk of a 'dangerous middle ground of automation'.

There are three major risks associated with human supervision of autonomous systems: (*a*) automation bias or over-trust, (*b*) under-trust, and (*c*) out-of-the loop (OOTL) control problems.[55] An additional problem is the 'moral buffer', where the user transfers responsibility for the consequences to the machine (see chapter 2, section III).

Automation bias or over-trust (sometimes called *automation complacency*), is the propensity for humans to over-rely on an autonomous system and assume the information provided by the system is correct. Research has shown that the more reliable human operators perceive the system to be, the more their cognitive resources are mobilized elsewhere and the less likely they are to monitor it properly. This risk is high when operators are expected to multitask or control multiple systems at once. When autonomy is used for life-critical tasks like weapon systems and the use of force, the consequences of automation complacency can be severe to catastrophic.

Under-trust is the opposite: it is the propensity for human operators to place insufficient reliance on an autonomous system. This generally happens when operators deal with systems that are known for producing false-positives or errors, or when the user interface is prone to inducing misinterpretations. A typical consequence of under-trust is that the human operator ignores relevant information provided by the system or overrides its action without justification. A number of incidents with autonomous air-defence systems have been caused this way. A famous example was the destruction of a commercial aircraft—Iran Air Flight 655—on 3 July 1988 by an Aegis Combat System stationed on the *USS Vincennes*, a US Navy warship.

OOTL control problems happen when an emergency or critical situation occurs, and the human operator is unable to regain sufficient situational awareness to react appropriately and in time.[56] As the DCDC report on human–machine teaming noted:

> Simply monitoring systems holds people's attention poorly. It is often very difficult for a previously unengaged person to be able to ramp up their mental alertness at a point of crisis, or orient themselves sufficiently quickly to the key variables and context in time to act.[57]

The OOTL control problem is well known in the commercial aircraft industry, and many aviation accidents have occurred because of a sudden transfer of control from the autopilot to the human pilot (as in the 2009 crash of the Air France Flight 447).[58] It is also a common problem associated with autonomous air and missile defence systems.[59] The USA's Patriot autonomous air defence system, for instance, was involved in two fratricide incidents during Operation Iraqi Freedom in 2003 because of that problem. According to John Hawley, a human–machine interaction specialist from the US Army Research Laboratory, one of the first reactions of the commander in charge

[54] Reason, 'Safety paradoxes and safety culture' (note 48); Hawley (note 47).

[55] Parasuraman, R., Molloy, R. and Singh, I. L., 'Performance consequences of automation-induced complacency', *International Journal of Aviation Psychology*, vol 3, no. 1 (1993); Murphy, R. and Burke, J. 'The safe human–robot ratio', eds Barnes and Jentsch (note 47); Sharkey (note 36), p. 36; ICRC, *Autonomy, Artificial Intelligence and Robotics: Technical Aspects of Human Control* (ICRC: Geneva, Aug. 2019), pp. 9–10.

[56] Murphy and Burke (note 55), p. 45; United States Air Force, Office of the Chief Scientist (note 41), p. 6.

[57] DCDC (note 49), p. 42.

[58] For a detailed account of how problems of human–machine interactions led to the crash, see Langewieshe, W., 'The human factor', *Vanity Fair*, 17 Sep. 2014; see also Mindell, D. A., *Our Robots, Ourselves: Robotics and the Myths of Autonomy* (Viking: New York, 2015).

[59] Hawley (note 47), p. 6.

of the Patriot systems at the time was to ask, 'How do you establish vigilance at the proper time? 23 hours and 59 minutes of boredom followed by one minute of panic'.[60]

This statement sums up one of the most fundamental problems posed by autonomy in weapon systems: how to calibrate the interaction between the human user and the AWS to ensure that it remains adequate and effective? Providing a technical solution to this problem is a significant challenge:

> On a commonly used scale of levels of autonomy, level one is fully manual control and level 10 is full autonomy. . . . History and experience show, however, that the most difficult, challenging and worthwhile problem is not full autonomy but *the perfect five*—a mix of human and machine and the optimal amount of automation to offer trusted, transparent collaboration, situated within human environments. . . . It takes more sophisticated technology to keep the humans in the loop than it does to automate them out.[61]

Moreover, what constitutes a 'perfect five' is context-based and depends on a number of variables including: (*a*) the type of tasks to be carried out; (*b*) the complexity of the environment; (*c*) the sophistication (or complexity) of the systems; and (*d*) the cognitive abilities and workload of the human supervisor.[62]

The complexity of weapon systems is, in this regard, a particularly challenging variable. The constant in the recent history of military technology has been that as technology systems have grown in complexity, it has become increasingly difficulty for human users to understand how they work. This trend is poised to be compounded by the potential integration of the latest advances of AI and machine learning in the design of these systems, given that software based on machine learning is typically far more opaque in its functioning than traditional rule-based software.[63] The incorporation of machine learning algorithms in the design of AWS could make their behaviour unpredictable and less understandable to the user.[64] But understanding and predicting the behaviour of an AWS is essential for achieving safety, military efficiency and most importantly legal compliance. In order for an AWS to be approved for use by both a system safety and legal review, the functioning of the system must be understandable, predictable and explainable.

Technically speaking, users would not need deep understanding of the inner workings of the system to operate it. Their understanding needs to be sufficient for predicting the operation of the system and any foreseeable consequences in the specific circumstances of use, and for trusting it will operated as intended. However, when a system fails—as all complex systems do—a lack of deep system understanding can lead to what industrial engineers call 'automation surprise'; it can hinder the user's ability to maintain or regain situational awareness and ensure the safety of operations. Again, this is a well-known problem in the commercial aviation industry, where automation surprise has been a factor in numerous plane crashes.[65] Increased complexity of weapon systems is, in other words, a growing challenge as far as human control is concerned. It places greater demands on the user's side: the user of an AWS increasingly needs to have an appropriate level of technical knowledge or be supported by technical experts to be able to operate the system safely and reliably.

[60] Hawley (note 47) p. 7; also cited in DCDC (note 49), p. 32.

[61] Mindell, D., 'Driverless cars and the myths of autonomy', *Huffpost*, 14 Oct. 2015.

[62] Boulanin and Verbruggen (note 16), p. 68.

[63] Boulanin, V., 'Artificial intelligence: a primer', ed. V. Boulanin, *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk*, vol. 1, *Euro-Atlantic Perspectives* (SIPRI: Stockholm, May 2019).

[64] ICRC, Autonomy, Artificial Intelligence and Robotics (note 55), p. 19.

[65] Reportedly, one the of the more common questions asked by commercial pilots in cockpits today is: 'what is it doing?'; see Langewieshe (note 58); see also Mindell, *Our Robots, Ourselves* (note 58).

Exercising human control: drawing from existing practice

Research on past and current use AWS, and autonomous life-critical systems more generally, provides useful lessons with regard to how human control can and should be exercised.

First, as discussed above, human control can be exercised in three ways: through controls on the weapon parameters (e.g. type of target, time and space constraints), controls on the environment; and controls on human–machine interaction during use.

Second, human supervision has been and is likely to remain a necessary form of control measure to ensure safety, reliability and efficiency of military operation—in addition to the legal obligations discussed in section II of this chapter and taking into account ethical considerations discussed in section III of this chapter. However, human supervision as a form of control is not a panacea. In most situations, having a human supervisor cannot be the only safeguard. In addition to the above methods of human control, there should be further safeguards, which can be either technical in nature, such as a fail-safe mechanism, or organizational, such as a procedure that distributes the decision-making process across multiple individuals to ensure the system's operator or supervisor is not the sole decision maker and that other humans can act as checks and balances. Research on past use of AWS and unarmed robots in military operations provide some useful insights in that regard. For example, Robin Murphy and Jennifer Burke demonstrated that an AWS needs to be under the oversight of not one but multiple humans.[66] One human operator alone cannot effectively monitor the critical functions of the AWS, the evolution of the environment of use, and the functioning of the systems. Multiple individuals need to be present, each with a clear and well-defined role, such as monitoring targets, monitoring the environment, or monitoring the functioning of the systems. To determine the number of humans that should be responsible for a machine, to best ensure its safety and efficiency—what they call a 'safe human–machine ratio', Murphy and Burke came up with the following formula:

$$N_h = N_v + N_p + 1$$

where

N_h is the number of humans needed
N_v is the number of vehicles
N_p is the number of payloads on those vehicles
+ 1 is an additional safety officer.[67]

Murphy and Burke showed that finding the right ratio is critical. If the human–machine ratio is too low, the result is increased risk because anomalies and errors might go unnoticed, and if it is too high, the overall coordination becomes inefficient, while operators are also exposed to unnecessary risk.[68]

Third, humans need to be cognitively involved to exercise effective control. Research on human–machine interaction in both military and civilian domains has also shown that human supervisors perform better when they feel that are actively involved in the process. Humans are 'far more mentally engaged' when they are 'searching for an already understood and defined object' or 'exploring for things of interest' like boundaries, anomalies and undefined targets.[69] This has practical implications for the design of human–machine interfaces, which need to be optimized to maximize

[66] Murphy and Burke (note 55), p. 35.
[67] Murphy and Burke (note 55), pp. 31, 47.
[68] Murphy and Burke (note 55), p. 35.
[69] DCDC (note 49), p. 32.

Box 2.4. Key recommendations from the technical literature on human–robot interaction for design of human–machine interfaces

There is an entire body of empirical research on human–robot interaction and supervisory control that demonstrates that to be good supervisors of autonomous systems, humans need to be highly cognitively involved in the operation of the systems. The human operator needs to pay attention but that attention is a finite resource and human reaction time is also limited.[a] This research has provided a number of useful recommendations on how human cognitive involvement can be supported via the design of the human–machine interface.

- Information displays and information presentation formats (e.g. text or graphics) should be adapted to the environment in which the operator works.[b]
- Interfaces should be adaptable to the individual differences in skills and experience levels of operators.[c]
- Interfaces should be designed to help operators achieve a greater understanding of the overall context.[d]
- Interfaces should be designed so that the operator can focus on their primary task rather than controlling the movement of the machine.[e]
- Interfaces could indicate the estimated level of accuracy of the information provided to the operator. Colours can also be used to make that information easier to grasp for the human operator.[f]

[a] Reason, J., 'Safety paradoxes and safety culture', *Injury Control and Safety Promotion*, vol. 7, no. 1 (2000); Reason, J., *Human Error* (Cambridge University Press: Cambridge, UK, 1990); Mindell, D. A., *Our Robots, Ourselves: Robotics and the Myths of Autonomy* (Viking: New York, 2015), p. 118.

[b] Cosenzo, K., Parasuraman, R. and deVisser, E., 'Automation strategies for facilitating human interaction with military unmanned vehicles', eds M. Barnes and F. Jentsch, *Human–Robot Interactions in Future Military Operations* (CRC Press: Boca Raton, 2010), p. 118.

[c] Riley, J. M. et al., 'Situation awareness in human-robot interaction: challenges and used interface requirements', eds Barnes and Jentsch (note b), p. 188; Cosenzo, Parasuraman and de Visser (note b), p. 118.

[d] Riley et al. (note c), p. 188.

[e] Riley et al. (note c), p. 190.

[f] Jenny Burke, conversation with authors 17 June 2019.

the user's cognitive participation (see box 2.4).[70] Perhaps more critically, the need for cognitive involvement indicates that as far as targeting is concerned, it might be preferable to keep humans in direct (remote) control of the targeting functions—though that may not be possible in some narrow operational situations where human reaction speed is a limiting factor, as illustrated by current use of AWS to defend against incoming missiles.

Fourth, training of operators is fundamental. Without proper training, the operator becomes 'a warm body at the system's control station'.[71] According to Hawley, training should equip operators of an AWS with the following key elements: (*a*) thorough knowledge of how the AWS works and will interact with the intended environment of use; (*b*) an understanding of the limitations of the AWS, and the ability to learn 'by experience when a machine can be trusted and when additional scrutiny and oversight are necessary'; (*c*) preparedness to handle risks related to use of the AWS on the battlefield; and (*d*) familiarity with 'edge cases' and 'corner cases' and how to handle the unexpected mindfully.[72] Training also needs to be provided on an ongoing basis. Human skills that go unpractised atrophy over time. Continuous training is the only way for operators to maintain their skill set.

[70] Reason, 'Safety paradoxes and safety culture' (note 48); Mindell, *Our Robots, Ourselves* (note 58), p. 118

[71] Hawley (note 47), p. 10.

[72] Hawley (note 47), pp. 9–10.

3. Operationalizing human control

This chapter explores how elements of human control could be applied to AWS in practice; that is, operationalized or implemented. The fundamental issue, in that regard, is determining exactly what falls under the concepts of 'human' and 'control'. This can generally be divided into four parameters:

1. ***Who*** *is/are the human/s in control?* The operator directly using the AWS, the commander in charge of the attack, the people in the broader command and control structure, or a combination of these?
2. ***What*** *is under human control?* The AWS as a whole, the critical functions of selecting and applying force to targets, or the consequences of the specific attack?
3. ***When*** *should human control be exercised?* Up to the point of activation of the AWS only, or during its operation as well?
4. ***How*** *should human control be exercised?* What form should the control take?

This chapter addresses these questions from the perspectives of the legal, ethical and operational requirements for human control presented in chapter 2. Section I starts with discussion of the parameters *who* and *what*, section II discusses *when*, and section III focuses on *how*.

I. *Who* and *what*: users of an autonomous weapon system and the consequences of its use

The baseline for answering the questions 'who?' and 'what?' is the relationship between *the users of the AWS and the consequences, or effects, of its use in a specific attack*. From an IHL perspective, the users of an AWS are the humans who plan, decide on and carry out attacks involving the AWS on behalf of a party to conflict, because they are responsible for complying with IHL. From an ethical perspective, a user is any human involved in the decision to use, and the actual use of, an AWS. The user's ability to exercise influence over the use of an AWS is the basis for their moral obligation. From an operational standpoint, the users are part of a wider human command and control chain, in which targeting decisions are made.[73] Control over the use of force in an attack may be distributed across a range of individuals who operate in various phases of a 'targeting cycle' that can involve many steps, including steps that belong to the deployment phase such as mission planning.[74] In practice, each of these users will be able to exercise some degree of control over an AWS and the consequences of its use, at different points in time. However, the overarching requirement applicable to all users is the need to maintain a certain level of human control over the use of an AWS and its effects.

The 'what' is effectively the use of the AWS in a specific attack and the consequences of that attack. While autonomy in these critical functions is the defining property of an AWS, and at the heart of the problem of ensuring control, the nature of the consequences will generally be determined by the interaction between the AWS and its operating environment during use. Put another way, what is at stake is the ability

[73] Ekelhof, M., 'Autonomous weapons: operationalizing meaningful human control', ICRC Humanitarian Law & Policy Blog, 15 Aug. 2018.

[74] Ekelhof, M., 'Moving beyond semantics on autonomous weapons: meaningful human control in operation', *Global Policy* vol. 10, no. 3 (2019); Ekelhof, 'Autonomous weapons' (note 73).

of the AWS users to reduce or compensate, through measures for human control, the inherent unpredictability that autonomy in the critical functions of an AWS introduces to the consequences an attack.

In summary, the questions of who is in control and what they are in control of calls for a holistic approach. While the focus must be on the relationship between the users of the AWS in a specific attack, there will need to be multiple safeguards distributed across the command and control chain. However, as discussed in section III of this chapter, what is critical—especially when applying IHL—is that human involvement in earlier stages of the decision-making process cannot be a substitute for control measures implemented at the point of use to enable context-specific judgements by the users in relation to a specific attack.

II. *When*: focus on the use of an AWS

The idea that human control is a conceptual approach that has implications from use all the way back to the design of an AWS has already been highlighted. In 2018, the GGE report identified five key phases in the development, deployment and use of a weapon system where measures for human control could be implemented:

- Phase 0. *The study or pre-research and development phase* is when technical requirements for the AWS are identified.
- Phase 1. *The research and development phase* is when the technical characteristics of the AWS are determined.
- Phase 2. *The acquisition phase* is when the AWS goes through testing and evaluation, validation and verification, and final legal review.
- Phase 3. *The deployment phase* ranges from the time of implementing user training and standard operating procedures for the AWS, to the time of mission planning (i.e. defining objectives, target profiles and rules of engagement) for its use.
- Phase 4. ***The use phase* starts with activation of the AWS until the attack is completed, aborted or terminated.**
- Phase 5. *The post-use phase* involves assessment of the effects of the AWS, to generate lessons that can be applied in future.[75]

There has been a tendency during international discussions to consider these phases as a linear process, with equal importance given to human control requirements implemented at each of these phases. However, as the analysis in chapter 2 illustrates, it is the consideration of the use of an AWS in a specific attack which is critical to assessing control measures necessary from a legal, ethical and operational standpoint. It is autonomy in the critical functions that makes the use of AWS unique, as compared to other weapon systems. And it is the analysis of the use of an AWS that will determine the necessary measures for human control—or limits on autonomy—to be implemented and reflected in all other aspects of the development, testing and deployment processes, including in the process of legal review before deployment of the AWS. This is further explored in section III below.

[75] CCW GGE, *Report of the 2018 Session of the Group of Governmental Experts on Emerging Technologies in the Areas of Lethal Autonomous Weapon Systems*, CCW/GGE.1/2018/3, 23 Oct.2018, p. 14.

III. *How*: determining the type and degree of human control required in practice

Objectives to be balanced: what the human needs to be able to do

The starting point for this analysis is that, in any given situation of use, the users must balance three objectives: (*a*) compliance with applicable international law; (*b*) the ability to retain and exercise human agency and moral responsibility over the use force and its consequences; and (*c*) ensuring military effectiveness while mitigating risks to friendly forces (see figure 3.1). In general, a key purpose of measures for human control is to reduce or compensate the inherent unpredictability in the consequences of an attack introduced by autonomy in the critical functions of a weapon system.

While there are many elements that need to be taken into consideration to determine the appropriate type and degree of human control in a given situation, as explored in the next subsection, there is one implicit question that is invariably at stake. Can the user have sufficient understanding of the AWS and its environment of use, and be able to take sufficient precautions to ensure that its use and consequences in a specific attack comply with IHL, are ethically acceptable and are operationally effective and safe? In that regard, if an overall guiding requirement for human control had to be formulated, it would have two components:

1. The user **must have reasonable certainty about the effects** of the AWS in the specific environment of use. That means the human must have sufficient understanding of both the environment of use and the AWS to be able to make the judgements required from legal, ethical and operational perspectives.
2. The users **must exercise judgement and intent**—or agency—in the use of force in specific attacks, both to ensure compliance with IHL and to uphold moral responsibility and ensure accountability for the consequences. This means the human user must exert effective influence over the functioning of the AWS in a specific attack.

A typology of measures

The *type* of measures for human control that apply to the use of an AWS can be parsed into the three general categories identified in chapter 2, section II: control over the weapon system's parameters of use, control over the environment of use, and control through human–machine interaction (see figure 3.2).

Control over the weapon system's parameters of use

One way to address legal, ethical and operational concerns posed by an AWS is to place controls over the parameters of its use. This can be done by formulating restrictions and requirements during the design and programming of an AWS. These could take the form of limitations on parameters such as: (*a*) target type and profile, (*b*) spatial and temporal scope of operation, (*c*) the weapon's effects, and (*d*) requirements for fail-safe rules and mechanisms.

Target type and profile. Most existing AWS are designed with constraints on the type of target to apply force to and the conditions under which to apply force. This kind of control measure has notable legal, ethical and operational significance. By controlling the type of target, those who carry out attacks using an AWS can reduce ethical concerns and the risk that the AWS will be triggered to use force against people or objects protected under IHL. For example, the use of an AWS to target only typical

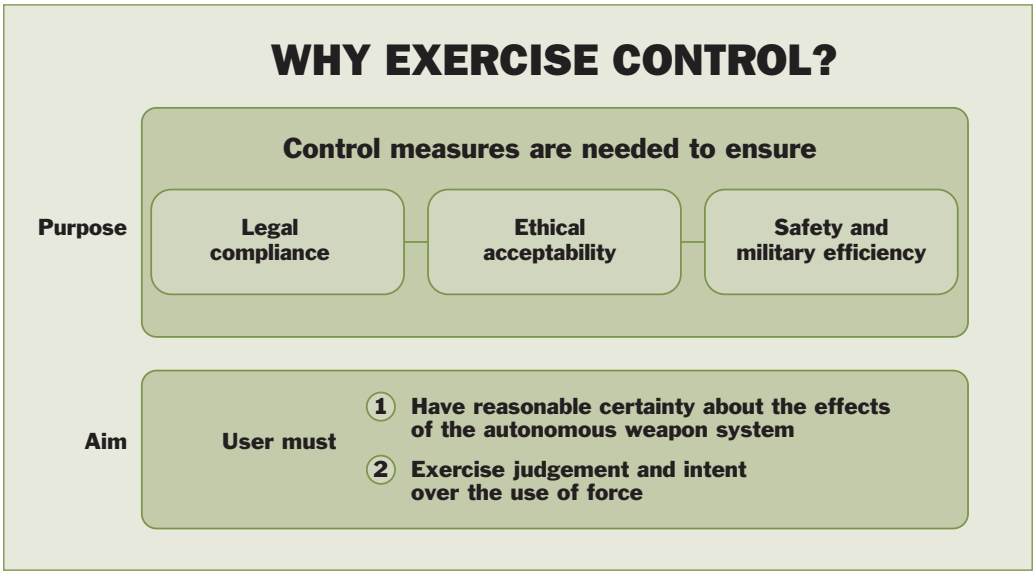

Figure 3.1 Why exercise control?
Source: Authors' own conceptualization

military objects, such as tanks or munitions, would appear to raise less acute concerns compared to the use of an AWS against a wider range of vehicles or objects that might also be used by civilians, or—ethically, particularly problematic—against people.

The specificity with which targets are encoded in an AWS—in other words, the narrowness of the weapon's target profile—is also a key control measure, closely linked to the type of target. For example, a target profile may incorporate criteria based on heat, weight, density, velocity, electromagnetic signatures and even image recognition. However, because protection from attack afforded to people and objects under IHL is not fixed, but can change swiftly once a person's behaviour or an object's use alters, target profiles built on technical characteristics are not an effective measure on their own to control against the risk of IHL violations. Target profiles that potentially encompass objects used by both enemy combatants and civilians—such as certain types of vehicles—raise particular concerns unless all civilian objects of this type were reliably excluded from the AWS's area of operation. As a general rule, profiles meant to capture targets that are difficult to specify in technical terms, or whose protection under IHL is highly variable, present a high risk—from a legal, ethical or operational perspective— of force being applied to objects that are not intended and lawful targets.

It may also be possible to program additional types of constraints, such as adding certain objects or areas to a no-strike list based on a prior understanding and mapping of the environment. These might include marking buildings that are protected (e.g. schools, hospitals) or marking areas where the presence of civilians is likely.

Spatial and temporal scope of operation. Constraints on the mobility of an AWS and its duration of autonomous operation, or 'loiter time', are important control factors, which are closely linked to controls on the environment discussed below. Greater mobility and duration of operation generally increases unpredictability, because the area of operation becomes larger and the environment changes over time. Therefore, important control measures are restricting mobility to a specific area in which the AWS can apply force to a specific type of target; and limiting the duration for which the AWS is activated. These types of spatial and temporal constraints may be programmed into the system or otherwise part of its design—or they may be specified directly; for example, where the users directly control the area of operation (e.g. by fixing the AWS in place or operating its movement by remote control) or the time-

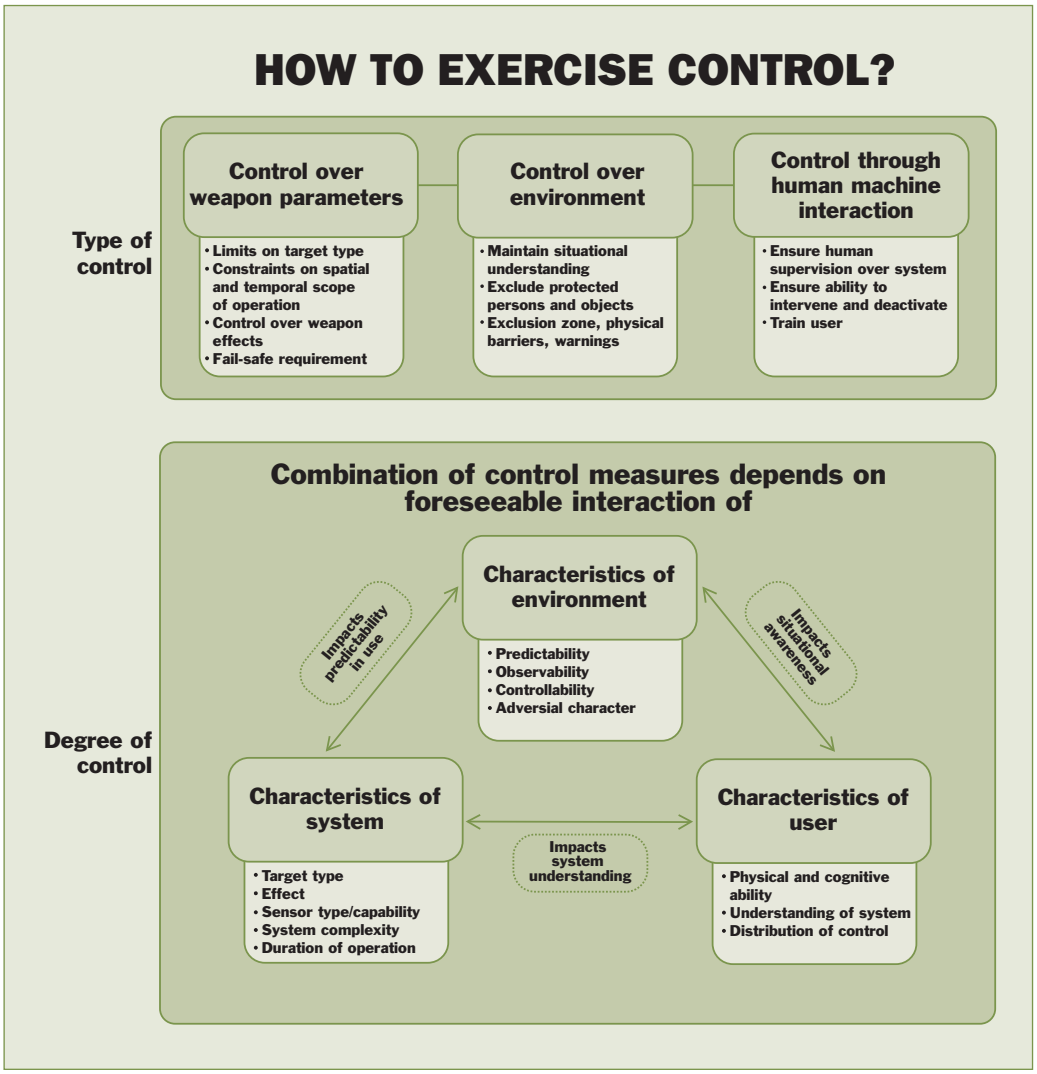

Figure 3.2 How to exercise control?

Credit: This figure is inspired by a model created by Patrik Stensson.

frame of activation (e.g. by activating the AWS for a limited time before deactivating it remotely).

Weapon effects. Understanding and limiting the effects that an AWS can produce in a given operational environment is needed in order to assess the risk of incidental harm to civilians and civilian objects under IHL rules, and of other effects that may weigh on ethical or operational acceptability. To exert adequate control over the weapon's effects, a user would, for instance, be expected to know the type and number of munitions that can be delivered in an attack.

Fail-safe rules and mechanisms. Fail-safe rules or mechanisms that are integrated in the design of the AWS can then be calibrated by the users before the activation and launch of the weapons. These could take the form of a deactivation function (e.g. stop, self-destruct or return to a specific area) that triggers when the system operates outside defined temporal and spatial boundaries, or when it malfunctions.

Control over the environment of use

Controlling or structuring the environment in which an AWS is intended to be used is a critical way of mitigating unpredictable effects and addressing some of the associated legal, ethical and operational risks. For example, limiting the use of AWS to areas where there are no civilians, and taking measures to exclude civilian objects

that may fall within the AWS's target paramaters from that area, would considerably reduce the risk of civilian harm. In maritime and air environments, such measures could involve declaring and enforcing an exclusion zone for civilian ships and aircraft, while on land measures could involve setting up physical barriers with warning signs or voice warnings to prevent civilians from entering a specific area. Such controls are already in place to restrict the use of sea mines, anti-vehicle mines and autonomous air defence systems (see chapter 2, section IV). Similar control measures could be applied to address ethical and operational risks, by excluding classes of persons and objects (beyond those protected by IHL) whose death/injury/destruction/damage would raise ethical or operational concerns, such as friendly forces. However, in some circumstances, particularly in populated areas and other places frequented by civilians—it will be challenging to exert controls on the environment that effectively and sufficiently exclude the risk of IHL violations or accidents that have unacceptable consequences. Additional measures for human control will be necessary. Critically, measures to control the environment do not discharge users of an AWS from their duty to make judgements of distinction, proportionality and precaution in attack under IHL.

Control through human–machine interaction

The third type of control is through human–machine interaction; that is, controlling how users of an AWS interact with that system in its environment of use. This category of control measures can be parsed into two sub-categories human-*in*-the-loop measures and human-*on*-the-loop measures, each reflecting different *degrees of human involvement* in the operation of the system.

Human-in-the-loop measures enable the users of a weapon system to directly and actively control the critical functions of selecting and applying force to targets—that is, the human has remote control of the weapon, and it is therefore not an AWS. In contrast, *human-on-the-loop* measures allow the user to supervise the functioning of the AWS in the environment while retaining the power to intervene by authorizing, vetoing or overriding the targeting functions, aborting a task or deactivating the system—that is, the human has the power to take remote control of the AWS.

These measures rely on giving human operators remote control and telepresence via technical components such as: a communication link sufficient to transfer raw sensor data; sensors and a human–machine interface allowing the operator to have direct visual contact and so obtain some operational situational awareness; and physical controllers (e.g. joysticks, keyboards) allowing the operator to control the actuators of the AWS, from the navigation system to the fire control system. The design of these technical components, particularly the human–machine interface and the physical controllers, must enable the necessary degree of human supervisory control. Notably, they will affect the manner and extent to which the users (*a*) perceive the targets and the environment of use; (*b*) are able to make deliberate choices about a target before initiating the application of force to it; (*c*) are in 'direct' control of the AWS; and, more generally, (*d*) are able to predict and understand the behaviour of the systems.

Perception of the target and the environment of use. The level of situational awareness and cognitive involvement of the AWS users will depend on whether they have direct visual contact with a target and its surroundings (e.g. through being present at the location or via a video link) or perceive the target through an interface (e.g. a radar screen) that registers the presence, location and movement of a target type in abstract form.

Deliberate choices about a target. Building on the classification levels proposed by Daniele Amoroso and colleagues, there are four types of deliberation involving a user's

choice of target for a weapons system to attack.[76] *Type 1:* users deliberate about a target before initiating any and every application of force. *Type 2:* the weapon system provides a list of targets and the users choose which to apply force to. *Type 3:* the weapon system selects the target and the users must approve the application of force. *Type 4*: the AWS selects and applies force to the target but the users retain the option to override and veto or cancel the application of force and/or attack. Under this typology 'Type 1' would be a remote-controlled weapon and not an AWS. 'Type 2' and 'Type 3' also represent systems that could be considered remote-controlled provided the human deliberation about the specific target is adequate and meaningful. However, should this human intervention only amount to a rubber-stamping of the machine calculation about a target, then 'Type 2' and 'Type 3' could effectively be AWS masquerading as remote-controlled systems.

Direct control of the weapon. The nature of the communication link and the volume and type of data fed to the user will determine the latency time—that is, the length of the delay between the moment the user decides and the moment the AWS actually executes that decision. Latency times can range from milliseconds to minutes. The longer the latency, the less direct control the user has over the AWS.

Predictable and transparent functioning. A necessary precondition for humans to be effective users of an AWS is being able to understand how the system works to the extent that it allows them to be able to predict its behaviour and effects (see chapter 2, section IV). From a technical standpoint, this creates the requirement for predictability and transparency in the functioning of the AWS. The system needs to be designed to enable the user to understand how it works, and tests of the system need to demonstrate how it behaves in different scenarios to ensure a certain level of predictability about its functioning. (Complete predictability is, however, impossible to achieve from a technical standpoint: engineers cannot foresee the behaviour in all possible situations.) These issues are particularly heightened should AI and machine-learning systems be used to control the critical functions of selecting and attacking targets, because of their problems of unpredictability and lack of explainability, especially of machine-learning systems.

The decision to implement one or more types of human control depends on the extent to which a user needs—be it from a legal, ethical or operational standpoint—to maintain situational awareness and agency in the operation of the weapon in a given set of circumstances. This will be discussed in greater detail in the next subsection. In any case, effective human–machine interaction for supervisory control requires both constant situational awareness and time to intervene (i.e. override a task, or deactivate or take back control of the AWS).[77] However, the demands on human users in a supervisory role are different from those on active controllers. Any supervisory control measures need to take into account the physical and cognitive limitations of humans in this respect. As discussed in chapter 2, section IV, the challenge for humans in maintaining the same level of attention and concentration for an extended period of time is well documented. Even personnel who are trained to understand the behaviour of the AWS and its limitations may encounter human–machine interaction problems such as automation bias, lack of operator situational awareness and the moral buffer. Users generally perform better when they are in an active role rather than a supervisory role.

[76] Amoroso, D. et al., *Autonomy in Weapon Systems: The Military Application of Artificial Intelligence as a Litmus Test for Germany's New Foreign and Security Policy*, Heinrich Böll Foundation (HBF) Democracy series, vol. 49 (HBF: Berlin, 2019), pp. 42–45.

[77] ICRC, *Autonomy, Artificial Intelligence and Robotics* (note 55), pp. 8–10.

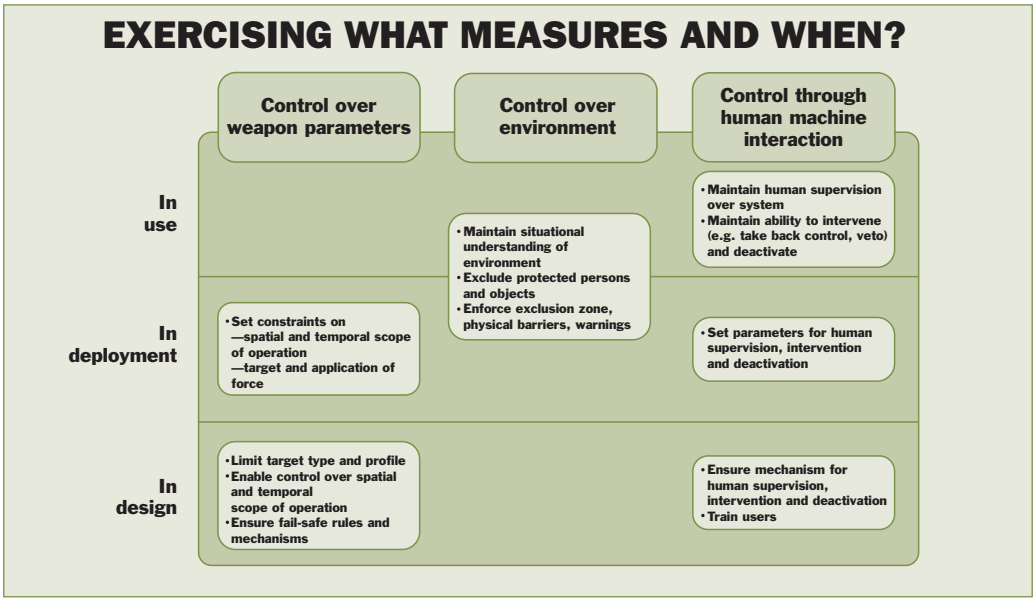

Figure 3.3. Exercising what measures and when?
Source: Authors' own conceptualization

Complementarity of and variation among the types of control

The above three types of control measures provide different ways to reduce unpredictability in the use of AWS and mitigate the risk of wrongful and unintended attacks. They each have different focus and operate on different time-frames (see figure 3.3). However, the measures are in no way mutually exclusive. In fact, in order to address legal, ethical and operational issues raised by AWS, these three types of control measures would need to be combined in all foreseeable scenarios, with the specific combination to a certain extent dependent on the context.

Determining the necessary combination of control measures applicable to autonomous weapon systems

To explore how context plays a role in the determination of type and degree of human control needed in practice, the workshop convened by SIPRI and the ICRC included fictional but realistic AWS use-scenarios in land, sea (surface, underwater) and air domains. Variables considered and varied in the exercises included but were not limited to: (*a*) the environment of use, including known or possible presence of protected persons and objects, such as civilians, civilian objects or combatants *hors de combat*; (*b*) the target profile; (*c*) other characteristics of the AWS, including the degree to which it relied on AI and machine learning, the types and capacities of the sensors, communication links, weapon effects and loitering capacity. This exercise was helpful in putting legal, ethical and operational standpoints into practice, and identifying the degree to which different measures of control will need to be adopted and combined depending on the type of AWS and the context of its use.

The exercise showed that there is likely no one-size-fits-all combination of control measures for every situation. That said, it is possible to identify a general model for determining, on a case-by-case basis, the necessary combination of measures. As the visual model in figure 3.2 illustrates, determining the best combination of measures depends on the interaction between variables in the characteristics of (*a*) the weapon system; (*b*) the environment of use (operating environment); and (*c*) the users.

Interaction with the characteristics of the weapon system

When considering the characteristics of an AWS, the key parameters to be taken into consideration are (*a*) type of target; (*b*) type of effect; (*c*) mobility; (*d*) types and capabilities of sensors; (*e*) system complexity; and (*f*) duration of autonomous operation.

Type of target. Anti-personnel AWS and anti-materiel AWS may demand different mixes of control measures. In the case of anti-personnel AWS, greater control in the form of human–machine interaction, whether human-on-the-loop or even human-in-the-loop measures (the latter meaning the system would be remote controlled and not an AWS), are likely to be needed in all circumstances. Since the status of whether a person is a lawful target can change from moment to moment (e.g. a combatant becoming *hors de combat* or surrendering, or a civilian directly participating in hostilities), controls over the environment of use and controls over the system's parameters are not likely to be sufficient to exclude the risk of IHL violations, not to mention ethical concerns. In the case of anti-materiel AWS, human-on-the-loop control measures might not be needed in some narrow circumstances; for instance, where control over the environment of use and control over the system's parameters provide sufficient certainty to the user that protected persons and objects are not put at risk by the use of the AWS.

Type of effect. The effect of a weapon is a relevant variable for determining the risk to civilians and civilian objects. Controlling these effects may be more or less difficult depending on the type of effect (e.g. narrow versus wide area) and the environment of use (e.g. unpopulated versus populated area).

Mobility. In the case of a mobile AWS, spatial and temporal constraints will be of critical importance. However, these might not be sufficient alone and will need to be supplemented by other forms of control measures.

Types and capabilities of sensors. The types, accuracy and level of 'perceptual capabilities' (i.e. ability to detect persons, objects or events) of sensors the AWS uses to perceive its environment and to identify targets is essential to determining the role of users in assessing situational awareness and predicting the consequences. At the same time the characteristics of the sensors also affect the user's ability to supervise effectively, which in turn can increase the need to exert control in other ways.

System complexity. More complexity in a system, especially one incorporating AI and machine learning, means it may not function in a predictable and transparent manner, making it even harder to exercise effective supervisory control. Complexity can notably increase the risk of the human–machine interaction problems of automation bias, under-trust, out-of-the-loop control and moral buffer (see chapter 2, section IV).

Duration of autonomous operation. Time is a critical factor in unpredictability. The longer the AWS operates, the greater the risk that the situation on the battlefield might change—for example, people or objects protected under IHL could enter the environment of use, military personnel might surrender. Strict constraints are needed, therefore, on the duration of operation. Systems that operate for long periods would need to be constrained even further by narrow target profiles, spatial restrictions, fail-safe mechanisms, and controls over the environment. Given that the human attention span decreases over time, human–machine interaction would also need to be optimized to allow users to maintain the necessary level of situational awareness that ensures acceptable use of the AWS.

Interaction with the characteristics of the environment

When considering the characteristics of the operating environment, several parameters need to be taken into consideration. The key parameters are the *predictability*,

observability, *controllability* and *adversarial character* of the intended location of use. These will not only affect the degree of human control necessary from operational, ethical and legal standpoints, but also the type of human control that is feasible.

Predictability of the environment depends on multiple variables, including: (*a*) the scope of the geographical area of use; (*b*) the presence of civilians and civilian objects; (*c*) the complexity of the environment and how it might change during the time-frame of the attack.

Observability depends not only on the extent of advance assessment of the geographical area of use, but also on how the users of the AWS will be able to observe the environment and update their assessment during its operation, through sensors and data from the system or other surveillance systems. The extent to which the capacity to observe over time might be hindered by loss of communication links is another factor.

Controllability refers to the extent to which the users of the AWS could effectively deploy measures—including physical controls or warnings—to control the environment to ensure no civilians or civilian objects are present but also that combatants *hors de combat*, including combatants whose status changes to protected during an attack, are not at risk.

Adversarial character determines the extent to which the users will be able to deploy measures to control the environment and the systems without interference during the operation.

The less predictable, observable and controllable the environment is to the user of an AWS, the more difficult it will be for the user to employ effective controls over the environment. This difficulty may necessitate a greater emphasis on restrictions on the weapon parameters as well as on the need for human–machine interaction.[78] The adversarial character of the environment is an important variable with regard to human–machine interaction as it affects the viability of human supervision.[79] In a non-adversarial environment, the users could afford to shift between different forms of human–machine interaction. For example, the phase during which the AWS is travelling to its target areas may require less supervisory control than the phase in which the weapon is in the target area, armed and searching for targets. In an adversarial environment, the users would have to plan for the adversary being able to force variation in the mode of human–machine interaction, for instance by jamming the communication link to make human–in- and human-on–the-loop control impossible. The possibility of such a scenario increases the need to employ other forms of control such as restrictions on the parameters of the weapon (e.g. the type of target, and a fail-safe mechanism).

Interaction with the characteristics of the user

Reviewing the environmental and technical parameters in the use of an AWS in a specific attack is only the first step in determining the type and degree of human control that is needed to ensure that its use in general, and in any specific attack, is lawful and ethically acceptable. The second step is to consider the capacity of individual operators and of the military as an organization to exercise that control. Three parameters that need to be taken into consideration are: (*a*) the physical and cognitive abilities of humans; (*b*) the user's ability to understand the system; and (*c*) the distribution of human control.

Physical and cognitive ability of humans to apply control. Depending on the operational situation, human physical and cognitive abilities can be a strength or weakness

[78] ICRC, *Autonomy, Artificial Intelligence and Robotics* (note 55), pp. 8–10.
[79] Boulanin and Verbruggen (note 16).

as far as legal compliance and operational effectiveness are concerned. Some operational situations allow humans to fully deploy their ability to make qualitative judgements and navigate uncertainty in rational and prudent ways (e.g. air strike against a target in non-adversarial airspace); situations that require the identification, selection and attack of targets at machine speed do not (e.g. ship defending itself against a large number of incoming missiles). In the latter case, the limitations of the human user's supervisory capabilities need to be compensated for by placing stricter controls over the environment of use and the AWS parameters.

User's ability to understand the system. The familiarity that the user has with the AWS, particularly in understanding the consequences of its use in a specific environment, is also an important variable. Some systems, particularly those incorporating AI and machine learning, may also risk introducing unpredictability by design. The more complex the system, the more training its users need to operate the weapon safely and reliably. The training should ensure that users have sufficient understanding of the AWS to predict its likely behaviour and effect in an environment of use. The extent to which the users have a detailed technical understanding of the system's functioning may be a consideration in the calibration between the different type of control measures.

Distribution of human control. In most foreseeable scenarios, an AWS does not interact with one single user but several users who share different responsibilities for supervision and control of the AWS—for example, one user for navigation systems and another for targeting functions. The question then is how many users need to be involved in the supervision of the AWS, and what would be their respective roles? The difficulty in that regard will be to ensure that the distribution of roles reinforces, rather than dilutes, the power to control the AWS.

Balancing the variables

In summary, there are many variables that need to be considered to determine the best combination of control measures needed to limit autonomy of a weapon system in a way that ensures compliance with IHL, ethical acceptability, and operational effectiveness. What is clear is that in most foreseeable circumstances all three types of measures for human control will be needed: control over the weapon system's parameters of use, control over the environment and control through human–machine interaction. The importance of each type may vary according to context, such as characteristics of the AWS and the environment of use. Where one type of control is inadequate or challenging to implement, others may rise in prominence. The critical issue from a legal and ethical perspective is that the combination of control measures must allow the user to compensate for the unpredictability introduced by the AWS while retaining human agency in decisions to use force.

Implications for the study, research and development, and acquisition phases

The above analysis drew conclusions for how to operationalize or implement measures for human control of an AWS when it is used in a specific attack. These conclusions have implications for how, where and at what point to implement those measures before the AWS is deployed: be it in the study, research and development (R&D), and acquisition phases (see chapter 3, section II). This is especially relevant for the conduct of legal reviews, as required by Article 36 of Protocol I, to determine whether the use of a new weapon, means or method of warfare—including any new AWS—would in some or

all circumstances be prohibited by applicable international law.[80] Such reviews will make an initial determination of the types and degrees of human control necessary for compliance with IHL.

Study phase

Human control should be a central consideration at the study phase of a new AWS when the technical requirements for the R&D process are formulated. There are a number of measures that can be implemented at the policy and organizational levels to ensure that human control is taken into serious consideration at that stage. At the broad policy level, a national policy document, in the form of a directive, strategy or roadmap, can provide political direction in terms of how autonomous technologies should be developed and used. One example of such a document is the United States Department of Defense Directive no. 3000.9, which spells out concrete requirements for the design, testing and use of autonomous and semi-autonomous weapon systems.[81] At the organizational level, there some practical measures that could help defence procurement agencies and personnel in charge of conducting legal reviews to understand and consider the requirements for human control when studying a new weapon or method of warfare that involves an AWS. An example would be to include legal and ethical considerations for human control in handbooks on systems safety that some countries, like Sweden, produce to provide guidance and instructions for all procurement, modification, renovation and decommissioning of military material.[82]

Research and development phase

The R&D phase is the point at which technical parameters for the AWS are designed and set. What is critical for human control at this stage is that the designer must think ahead to where and how the technology is intended to be used. Reviewing the characteristics of the potential environment of use and the nature of the mission will be informative as to the types and degrees of human–machine interaction needed, but it will also inform the types of programmable constraints to be built into the design of the AWS. People in charge of conducting legal reviews can here play an essential role in helping technical personnel on both demand and supply sides to understand the legal and ethical requirements for human control and translate them into engineering decisions. The above analysis clearly showed that the most basic requirement is that an AWS should be designed so that the user can place restrictions on the parameters of use. If the legal review shows that the AWS requires a user to be in position to supervise and intervene during its operation, particular attention should be given to the design of the human–machine interface. The interface must allow the operator to understand the behaviour of the AWS sufficiently to be able to intervene successfully if something goes wrong.

Acquisition phase

In the acquisition phase, three key junctures provide complementary opportunities to exercise and improve the possibility of retaining human control over the AWS: the certification process, user trials and the legal review.

[80] Boulanin, V. and Verbruggen, M., *Article 36 Reviews: Dealing with the Challenges Posed by Emerging Technologies* (SIPRI: Stockholm, 2017), pp. 17–25; Goussac, N., 'Safety net or tangled web: legal reviews of AI in weapons and war-fighting', ICRC Humanitarian Law & Policy blog, 18 Apr. 2019; Lewis, D., 'Legal reviews of weapons, means and methods of warfare involving artificial intelligence: 16 elements to consider', ICRC Humanitarian Law & Policy blog, 21 Mar. 2019; ICRC, *International Humanitarian Law and the Challenges of Contemporary Armed Conflicts* (note 11), pp. 34–35.

[81] United States Government, Department of Defense, Directive no. 3000.09 on autonomy in weapon systems, 21 Nov. 2012 (updated 8 May 2017).

[82] Swedish Armed Forces, *Armed Forces Handbook on System Safety 2011*.

In order to be certified for use, an AWS, like any life-critical system, needs to go through testing and evaluation (T&E) as well as validation and verification (V&V) of the system as a whole and of its individual components. These processes are essential for determining the level of predictability and reliability of the systems. They can also inform the type and degree of control necessary.

User trial is a process intended to verify that a system's design actually responds to the needs of the end user. This stage is essential for exploring how the users of an AWS will interact with it in practice and for determining whether the types and degrees of human control envisioned at the design stage will be effective or need to be changed prior to deployment. These trials can reveal the extent to which the user will be able to be cognitively involved in the operation of the AWS in the intended context of use, as well as the extent to which the user will be able to understand and trust how the systems works.

4. Key findings and recommendations

This report presents the findings of the ICRC/SIPRI joint project entitled 'Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control'. This project, which encompassed a workshop convened in Stockholm in June 2019, was designed to support international discussions on AWS, including those in process at the UN CCW GGE. The work comprised a mapping exercise to identify the practical elements of human control and assess how these can be operationalized or implemented, based on legal, ethical and operational requirements and considerations in armed conflict. This final chapter summarizes the key findings and recommendations.

I. Key findings

There are legal, ethical and operational imperatives for human control over the use of force and, therefore, over autonomous weapon systems

Humans must retain control (and judgement) over the use of force in specific attacks in armed conflict, and therefore over the use and effects of weapons, for legal, ethical and operational reasons. However, AWS—weapons that select and attack targets without human intervention—pose fundamental difficulties for humans in exerting such control, which raises associated legal, ethical and operational challenges.

From a legal perspective, the use of AWS raise important, interrelated challenges for the interpretation and application of IHL. The user's ability to make the value-based judgements required by IHL rules can be compromised by undue reliance on sensor data and programming (the 'numbers' challenge). AWS also challenge the user's ability to make context-based decisions in light of the circumstances ruling at the time of an attack (the context challenge). Finally, the unpredictability introduced by AWS can hinder the user from properly anticipating and limiting the effects of the weapon as required by IHL (the predictability challenge).

From an ethical perspective, the use of AWS presents a serious challenge for ensuring human agency and upholding moral responsibility in decisions on the use of force, as well as maintaining respect for human dignity and avoiding further dehumanization of warfare. This challenge arises because the connection between intention of the user and the consequences that result is diluted or removed; there is always a degree of 'cognitive distance'—in time, space and understanding—between the human decision to use an AWS in an attack and the eventual consequences of that attack.

Operationally speaking, while AWS may offer certain military advantages—speed of action and continued operation where remote control is not viable—unpredictability in the consequences of their use also brings challenges for ensuring safety and efficiency in military operations.

Measures for human control need to be a combination of controls on weapon system parameters, on the environment and through human–machine interaction

Measures for human control must ensure that users have reasonable certainty about the effects of the AWS in the specific environment of use—in other words, that the consequences will be sufficiently predictable. This requires sufficient situational understanding of the environment of use, understanding of the technical functioning of the AWS, and foreseeability of the interactions between the two. Further, users

must exercise judgement and intent in decisions to use force, which requires effective influence over the functioning of the AWS in a specific attack.

Measures for human control can be grouped into three main categories:

1. **Controls on the weapon system's parameters of use** include measures that restrict the type of target and the task the AWS is used for; place temporal and spatial limits on its operation; constrain the effects of the AWS; and allow for deactivation and fail-safe mechanisms.
2. **Controls on the environment** are measures that control or structure the environment in which the AWS is used, and which overlap with the weapon system's parameters above. Examples include using the AWS only in environments where civilians and civilian objects are not present, or excluding their presence for the duration of the operation, such as through strict temporal and spatial constraints on the use of the AWS, exclusion zones, physical barriers and warnings.
3. **Controls through human–machine interaction** include measures that allow the user to supervise the AWS and to intervene in its operation where necessary, through direct active control, vetoing or overriding AWS functions, aborting a task or mission, or deactivating the AWS.

Each of these three categories provides different ways to reduce or compensate for the unpredictability in the use of an AWS and to mitigate the risk of wrongful and unintended consequences especially risks for civilians. In order to address legal, ethical and operational considerations, all three types of control measures likely need to be combined in any scenario. However, the particular combination of these measures will vary according to the specific context. Where one type of control measure is inadequate, insufficient or challenging to implement, the other types may rise in prominence. Overall, what is important when applying IHL is that the combination of control measures must sufficiently reduce or compensate the unpredictability in the use of AWS such that users have reasonable certainty about the effects of using the weapon in an attack. Ethical considerations may demand additional constraints, such as on the types of target, given the particular concerns with AWS designed or used to apply force against persons.

II. Recommendations

The key findings of this report can be summarized in five recommendations that are intended to inform international efforts to agree on limits on AWS, whether in new legally binding rules, non-binding standards or best practice guidance.

1. States should focus their work on determining how measures needed for human control apply in practice. Preserving human control over the use of force—whether characterized as human judgement, human involvement or the 'human element'—as well as human responsibility and accountability for the use of force, should remain the focus of states' work. Control measures—on weapon system parameters, on the environment, and through human–machine interaction—provide a practical framework for determining necessary limits on autonomy in weapon systems, which are in turn based on the legal, ethical and operational obligations and responsibilities of parties to armed conflict and individual combatants who must implement these obligations and responsibilities. Since these control measures are not tied to specific technologies, they provide a robust normative basis applicable to the regulation of

both current and future AWS. States should now focus on determining, and agreeing on, how these measures should be applied in practice.

2. Measures for human control should inform any development of internationally agreed limits on AWS—whether new rules, standards or best practices. The three types of control measures—on weapon system parameters, on the environment, and through human–machine interaction—indicate what limits may be needed to reduce or compensate for the unpredictability of AWS and associated risks for persons and objects protected under IHL, as well as to address ethical concerns.

Controls on weapon system parameters of use could inform the development of limits or restrictions on the types of targets against which AWS may be used (especially in view of ethical concerns over human targets), constraints on a weapon's freedom of movement or duration of autonomous operation, and requirements for deactivation and fail-safe mechanisms.

Controls on the environment could inform the development of limits aimed at avoiding (or at least minimizing) the risk of harm to civilians and civilian objects, for example, by precluding the use of AWS in civilian areas or otherwise excluding civilians from a weapon's area of operation.

Controls through human–machine interaction could inform the development of limits addressing the need for human supervision of the AWS and the ability to intervene in its operation and deactivate it ('human-on-the-loop' or 'human supervisory control').

3. States should clarify where rules under international humanitarian law rules already set constraints on the development and use of autonomous weapon systems, and where new rules, standards and best practice guidance may be needed. While it is beyond dispute that IHL applies to the development and use of AWS in armed conflicts, and rules of IHL set constraints on their lawful use, different views remain on the extent to which the practical measures for human control identified in this report are derived from existing IHL rules. Further expert discussions on states' interpretation and application of IHL with respect to AWS is important. These discussions could be structured around the relevant IHL obligations and, in view of the unique characteristics of AWS, an assessment of any challenges of interpretation and practical implications for their application. Such discussions could facilitate common understanding on the constraints imposed by existing rules while also identifying potential gaps, where additional understandings or new rules, standards and best practice guidance may be needed. Such discussions should also take into account the potential for ethical considerations to be a driver of the development of new rules, standards and best practice guidance.

4. New rules, standards and best practices must build on existing limits on autonomy under international humanitarian law, and should draw on existing practice. Should states decide to further develop the normative framework applicable to AWS, whether in the form of new legally binding rules, non-binding standards or best practice guidance, such developments should be closely tied to legal, ethical and operational obligations for human control. Any normative development should also focus on human obligations and responsibilities, not on technological fixes, so as to remain relevant and practical, notwithstanding any future technological developments. Such rules, standards and best practices must build on the limits on autonomy under existing IHL rules and draw on existing practices for using AWS. It may be difficult, and perhaps counterproductive, to try to encapsulate human control criteria in technical detail. Rather, it is likely that new rules, standards and best practice guidance can be more effectively articulated in terms of limits on specific

types of AWS, of the manner and circumstances of their use and on requirements for human supervision and intervention.

5. Human control measures should be considered in the study, research and development, and acquisition of new weapon systems. Preserving human control—through practical control measures on the weapon system parameters, on the environment, and on human-machine-interaction—should be a central consideration in the study, R&D and acquisition phases of new weapon systems. In the study phase, state and organizational policy and procedural documents that determine the development direction of a new weapon should consider human control requirements applicable to its intended use. Technical parameters for human control should also be designed and set in the R&D phase. Acquisition phase processes, such as testing and evaluation, certification, user trials and legal review, should assess the weapon against human control criteria.

About the authors

Dr Vincent Boulanin is a senior researcher at SIPRI, where his work focuses on the challenges posed by the advances of autonomy in weapon systems and the military applications of artificial intelligence (AI) more broadly. Before joining SIPRI in 2014, he completed a doctorate in political science at the École des Hautes Études en Sciences Sociales, Paris. His recent publications include *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk: Vol. I Euro-Atlantic Perspectives* (SIPRI, 2019, editor); *Cyber-incident Management: Identifying and Dealing with the Risk of Escalation* (SIPRI, 2020, co-author); *Bio Plus X: Arms Control and the Convergence of Biology and Emerging Technologies* (SIPRI, 2019, co-author), *Mapping the Development of Autonomy in Weapon Systems* (SIPRI, 2017, co-author).

Neil Davison is a senior adviser in the Department of International Law and Policy at the ICRC's headquarters in Geneva. For the past nine years he has represented the ICRC on weapons and disarmament issues, including throughout the United Nation's discussions on autonomous weapon systems. Prior to joining the ICRC, he led policy initiatives on international security and diplomacy at the Royal Society, the United Kingdom's national academy of science (2007–11), and he was a researcher on arms control at the University of Bradford, UK (2002–07). A biologist by initial training, he holds a PhD in Peace Studies.

Netta Goussac has worked as an international lawyer for over a decade, including for the ICRC (2014–20) and the Australian Government's Office of International Law (2007–14), and as a lecturer at the Australian National University. In 2020, Netta joined SIPRI as an Associate Senior Research within the Armament and Disarmament research area. Netta has particular expertise in legal frameworks related to the development, acquisition and transfer of weapons. She has provided legal and policy advice related to new technologies of warfare, including autonomous weapons, military applications of AI, and cyber and space security. Since 2017, Netta has participated in the United Nations' Group of Governmental Experts on lethal autonomous weapon systems.

Moa Peldán Carlsson was a research assistant at SIPRI between 2019 and 2020. Her research focused on emerging military and security technologies, including AI, cybersecurity and arms control. Her other research interests include the relationship between gender and terrorism. She wrote her bachelor's thesis on alternative paths for female empowerment in militarized societies. Prior to joining SIPRI, Moa was an intern at the Swedish national committee of UN Women and managed her own graphic design and communications company.

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

ICRC