sipri

RETAINING HUMAN RESPONSIBILITY IN THE DEVELOPMENT AND USE OF AUTONOMOUS WEAPON SYSTEMS

On Accountability for Violations of International Humanitarian Law Involving AWS

MARTA BO, LAURA BRUUN AND VINCENT BOULANIN

STOCKHOLM INTERNATIONAL
PEACE RESEARCH INSTITUTE

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

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

STOCKHOLM INTERNATIONAL
PEACE RESEARCH INSTITUTE

RETAINING HUMAN RESPONSIBILITY IN THE DEVELOPMENT AND USE OF AUTONOMOUS WEAPON SYSTEMS

On Accountability for Violations of International Humanitarian Law Involving AWS

MARTA BO, LAURA BRUUN AND VINCENT BOULANIN

sipri

STOCKHOLM INTERNATIONAL
PEACE RESEARCH INSTITUTE

October 2022

Contents

Acknowledgements

The report was produced in the context of a SIPRI project titled 'Limits on Autonomy in Weapons Systems—Retaining Human Responsibility in the Development and Use of Autonomous Weapon Systems: Exploring Key Challenges and Opportunities'. The authors would like to express their sincere gratitude to the Dutch Ministry of Foreign Affairs, the Swedish Ministry for Foreign Affairs and the Swiss Federal Department of Foreign Affairs for their generous financial support for this project.

The authors are indebted to all the experts who shared their knowledge and experience in the virtual discussion series SIPRI organized in February 2022: Tabitha Bonney, Christine Boshuijzen-van Burken, Berenice Boutin, Maya Brehm, Karl Chang, Rebecca Crootof, JJ Domingo, Ola Engdahl, Paola Gaeta, Dustin Lewis, Eve Massingham, Pamela Moraga, Lauren Sanders, Filippo Santoni de Sio, Michael Siegrist, Jeroen van den Boogaard and Binxin Zhang.

The authors wish to thank the experts who participated in background interviews: Mirco Anderegg, Ilana Gimpelsen, Hans Boddens Hosang, Dvir Saar, Marco Longobardo, Marco Sassolí and Michael Siegrist.

The authors are also grateful for the comments provided by Dustin Lewis, Maya Brehm, Rebecca Crootof, Paola Gaeta, Lauren Sanders, as well as SIPRI colleagues Giovanna Maletta, Netta Goussac and Janet Feenstra.

Finally, the authors would like to acknowledge the invaluable work of Linda Nix and the SIPRI Editorial Department.

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

Executive summary

It is undisputed that humans must retain responsibility for the development and use of autonomous weapon systems (AWS) because machines cannot be held accountable for violations of international humanitarian law (IHL). However, the critical question of how, in practice, humans would be held responsible for IHL violations involving AWS has not featured strongly in the policy debate on AWS. This report aims to support a deeper and more focused expert discussion on that very question. There are multiple legal frameworks through which human responsibility for IHL violations may be ensured. This report focuses on two central frameworks: the rules governing state responsibility for internationally wrongful acts and individual criminal responsibility for war crimes.

The rules governing state responsibility for internationally wrongful acts and individual criminal responsibility for war crimes are essential to upholding respect for IHL in the development and use of AWS, including by preventing an 'accountability gap' for IHL violations that involve AWS. These rules fulfil different, yet complementary, functions. The rules governing state responsibility provide a framework for collective responsibility. They aim to provide accountability for any act or omission that would constitute a breach of a state's international obligations, and they cover the conduct of any agents whose acts are attributable to the state. In the context of AWS, this means that the violation of any rule of IHL applicable in the development and use of AWS conducted by an agent of the state could, in theory, engage the responsibility of that state.

The rules governing individual criminal responsibility for war crimes are meant to ensure an individualized form of accountability for certain serious violations of IHL. They provide a framework to prosecute individuals who, for instance, commit or participate in violations of the rules governing the conduct of hostilities. Individuals who develop or use AWS with the intent to attack people or objects that are protected under IHL, or in the knowledge that the AWS will bring about civilian harm that is excessive in relation to the anticipated military advantage, could be held criminally responsible by a domestic or international criminal court.

However, tracing back—that is discerning, scrutinizing, and attributing—IHL violations in the development and use of AWS that would engage state responsibility or individual criminal responsibility (or both) remain, in some respects, challenging for at least four reasons.

First, how the rules of IHL should be interpreted and applied in the context of AWS is unsettled. While the rules governing state responsibility for internationally wrongful acts and individual criminal responsibility for war crimes are intrinsically linked to the content of the IHL rules, major interpretative questions persist as to what these rules require, permit or prohibit in the development and use of AWS, especially in terms of types and degrees of human–machine interaction. These conflicting interpretations constitute an obstacle to agreeing on the basis for establishing state responsibility or individual criminal responsibility.

Second, the unpredictability associated with AWS highlights—and potentially exacerbates—existing legal disputes around the interpretation and enforcement of state responsibility and individual criminal responsibility. In the context of state responsibility, challenges relate to establishing whether an unintended harmful incident amounts to a breach of IHL that engages the responsibility of the state. With regard to individuals, issues of unpredictability reignite debates around the criminalization of risk-taking behaviours, especially the question of whether being reckless about the effects of an attack would trigger criminal responsibility for war

crimes. The pathways through which AWS may put protected people and objects at risk also underlines the lack of an international definition of the war crime of indiscriminate attacks and the lack of criminalization of the duty to take precautions in attack.

Third, the development and use of AWS are likely to involve a large number of varied actors. This poses questions as to how responsibility to implement IHL obligations may be distributed among multiple people and in what circumstances responsibility is imposed on one person, the commander. While it is recognised that AWS must be developed and used within a responsible chain of command and control, it remains unclear what that looks like in the context of AWS. The lack of common understanding around how to implement IHL obligations that may be distributed among multiple actors has practical implications for the establishment of both state responsibility and individual criminal responsibility. For example, it may be difficult to determine when and where an agent of the state committed a particular breach of any of the IHL obligations incumbent on the state, and how one or more breaches by one or more agents may be linked to each other. Similarly, it may be difficult to discern the conditions under which a commander's responsibility for a war crime in the use of AWS could arise, but also the conditions under which the acts or omissions of other actors involved in the development and use of the AWS (such as developers and people involved in the design and acquisition of the AWS) could amount to participation in the commission of a war crime.

Fourth, AWS present some challenges, but perhaps also opportunities, with regard to how responsibility for IHL violations may be investigated and attributed to individuals or states. On the one hand, challenges associated with the black box of artificial intelligence (AI) and unpredictability complicate the ability to collect and assess information surrounding an incident involving AWS. On the other hand, investigations could potentially be enhanced by some of the technical features of AWS, such as digital logs, and by auditing mechanisms that could facilitate the task of tracing specific conduct back to one or more agents involved in the decision-making process. However, the implications of AWS on the practical ability to trace back conduct are not well understood.

Overall, the legal questions raised by AWS with regard to how responsibility for IHL violations may be attributed provide an opportunity for states to clarify what respect for IHL demands from states and their agents, and potentially to resolve old debates around how the rules governing state responsibility for wrongful acts and individual criminal responsibility for war crimes may be interpreted and applied.

In light of the above, the report makes three recommendations.

First, states and experts should further deliberate on how IHL is to be respected in the development and use of AWS, particularly in determining who should do what, when and where. Clarifying what respecting and ensuring respect for IHL mean in the context of AWS is a necessary first step towards discerning what and whose acts or omissions would not only engage state responsibility but also individual criminal responsibility. Such an exercise would also help determine how roles and responsibilities for the use of AWS may be distributed in the (human) chain of command and control, which would in turn help prevent, investigate and suppress potential IHL violations in the development and use of AWS.

Second, states should share information and exchange views about national practices that can foster respect for IHL and help trace back IHL violations in the development and use of AWS. One practical way to support further deliberations on how IHL norms should be respected is to share information and views on practices and procedures to implement IHL obligations at the national level. That would entail

sharing further information around practices and procedures for not only the legal review of new weapons, means and methods of warfare but also the provision of legal advice and training to the armed forces. It would also be useful if states elaborated on what investigatory mechanisms they have in place for investigating harmful incidents. States could also share information on how they currently ensure compliance with IHL at a systemic level, for instance by elaborating on the roles and responsibilities of the different actors involved in decisions to use an AWS.

Third, states should elaborate on concrete limits and requirements in the development and use of AWS that could help ensure human responsibility in practice. With regard to limits, states could seek to identify technical features or standards that would make an AWS indiscriminate by nature, or that could make compliance with the principles of distinction, proportionality, and precautions potentially difficult. States could also seek to recognize technical features that would have implications on the practical task of tracing responsibility for IHL violations back to humans. The identification of such technical features could help more clearly delineate the contours of a two-track regulation on AWS, one that, as suggested by a number states, would prohibit certain types of AWS on the one hand, and regulate the development and use of all others on the other hand.

With regard to requirements, states could seek to clarify the standards of intent, knowledge, behaviour and care that are demanded from the different actors involved in the development and use of AWS. That would require adopting a holistic approach to the issue of human–machine interaction and elaborating on what decisions may be taken at the critical junctures in the development and use of AWS, and how these decisions would need to interact with one another. Such an exercise would not only generate concrete recommendations for normative and operational frameworks governing AWS, but also facilitate the task of discerning, scrutinizing and attributing unlawful conduct in the development and use of AWS.

1. Introduction

The legal, ethical and security challenges posed by (lethal) autonomous weapon systems (AWS[1]) have since 2013 been subject to intergovernmental discussions within the framework of the 1980 Convention on Certain Conventional Weapons (CCW Convention) under the auspices of the United Nations.[2] The discussion, which since 2017 has been led by the open-ended Group of Governmental Experts (GGE) on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, has gained significant attention as it addresses fundamental questions about how humans and machines could—and, importantly, should—interact in decisions to use force. While debates persist around how, and to what extent, AWS should be subject to regulation, the GGE has been able to agree on some fundamental principles, including that international humanitarian law (IHL) 'continues to apply fully to all weapons systems' in armed conflicts, including AWS, and that certain types and degrees of human–machine interaction are needed to ensure compliance with IHL.[3] Moreover, the GGE has agreed, in its guiding principles (b) and (d), that humans must 'retain' responsibility for the use of weapon systems on the basis that machines cannot be held accountable for violations of IHL.[4] However, how 'human responsibility' should be retained in practice remains a relatively underexplored, yet critical, question.

This question is often answered differently depending on the perspective from which it is approached, with the main domains being ethical, legal and operational. In the legal domain, the notion of human responsibility can be approached from two perspectives. A forward-looking perspective (prescription) focuses on the norms of IHL that states and individuals have to comply with in the development and use of AWS, prescribing what humans need to do in their future actions to behave responsibly. A backwards-looking or accountability perspective (ascription) focuses on the rules under which states and individuals would be held responsible for IHL violations and the legal consequences of past actions. This report addresses the question of human responsibility in the legal context from both perspectives.

To date, the GGE has mainly approached the issue of legal responsibility from the forward-looking perspective of prescription. It has focused on the norms of IHL with which states and individuals have to comply when developing and using AWS. States and experts have debated extensively—without coming to a definitive answer—on, for example, what the cardinal principles of distinction, proportionality and precautions in attack demand from humans and their interaction with technology: who needs to do what, when, where and how. Less attention has been cast on legal responsibility from a backwards-looking perspective, although it can provide useful insights for the policy process on the regulation of development and use of AWS, particularly with

[1] This report refers to autonomous weapons systems (AWS) rather than lethal autonomous weapon systems (LAWS), which is the term the GGE has been using. AWS is preferred because the view shared by the International Committee of the Red Cross (ICRC) and a number of states is that lethality is a superfluous qualifier. For more explanation see box 1.1.

[2] Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons which may be Deemed to be Excessively Injurious or to have Indiscriminate Effects (CCW Convention, or 'Inhumane Weapons' Convention), opened for signature with protocols I, II and III on 10 Apr. 1983, entered into force on 2 Dec. 1983; amended protocol II entered into force on 3 Dec. 1998; protocol IV entered into force on 30 July 1998; protocol V entered into force on 12 Nov. 2006.

[3] CCW Convention, Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems (GGE), 'Report of the 2019 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems', CCW/GGE.1/2019/3, 25 Sep. 2019, annex IV, 'Guiding principles'.

[4] CCW Convention, GGE, 'Guiding principles' (note 3): '(b) Human responsibility for decisions on the use of weapons systems must be retained since accountability cannot be transferred to machines. This should be considered across the entire life cycle of the weapons system'; and '(d) 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'.

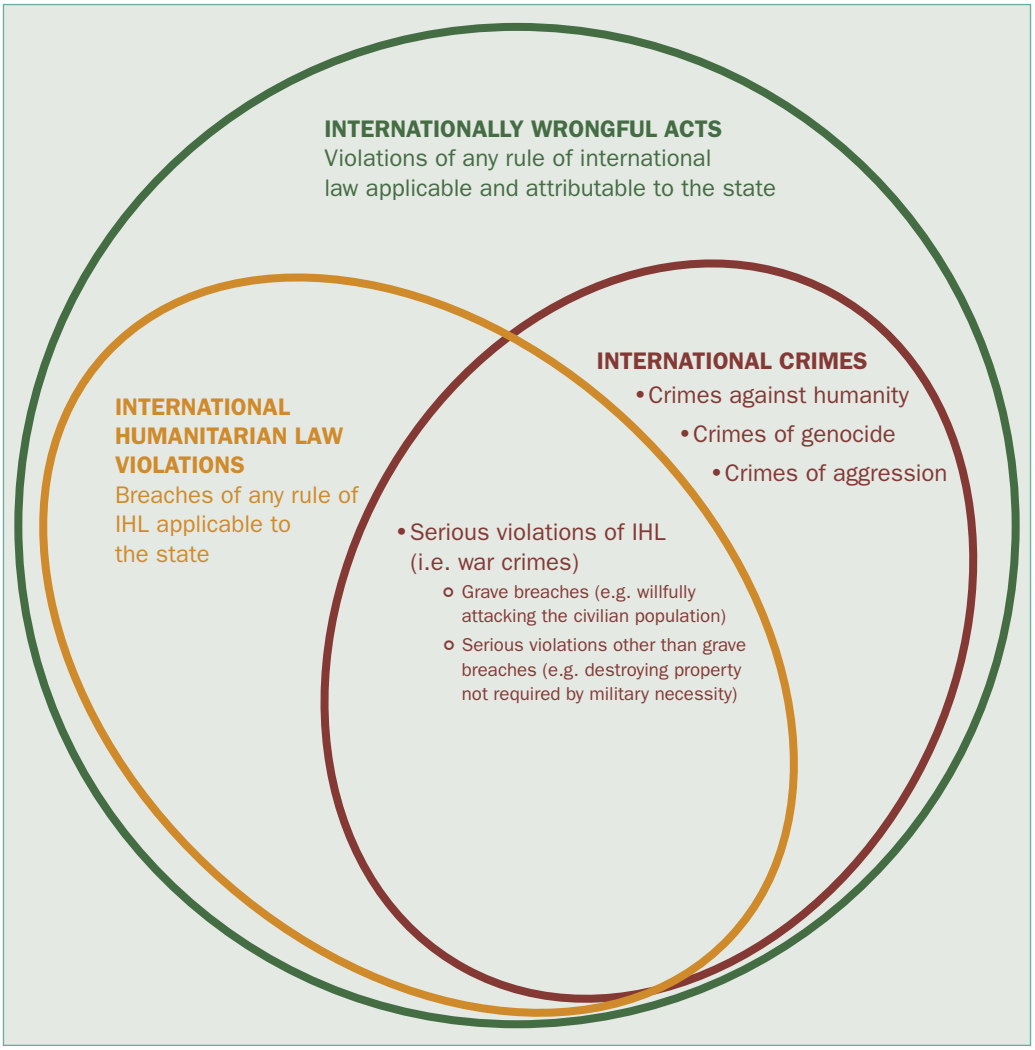

Figure 1.1. The relationship between internationally wrongful acts, international humanitarian law (IHL) violations and war crimes

regard to what is demanded from states and individuals. Both approaches are critical to preventing an accountability gap in the event of a breach of IHL arising from the use of an AWS.

As the issues in the CCW debate reach a deeper level of granularity, the question of ensuring that states and individuals are held responsible for their actions now warrants further elaboration and clarification. This report responds to that call by seeking to provide a basis for a more informed and focused debate on how, both in theory and in practice, humans can and should be held responsible for IHL violations involving AWS; and ultimately to help states elaborate and express their views on how the normative and operational governance frameworks may need to be clarified and developed further. Since establishing that an internationally wrongful act (giving rise to state responsibility) or a war crime (giving rise to individual criminal responsibility) has occurred depends on the normative standards set by primary IHL rules, approaching these rules from an accountability perspective provides an opportunity to clarify and substantiate what they demand from humans and permit from machines.

I. Background and aim of the report

This report is the result of a one-year research project that involved desk research; an online expert workshop with legal experts from academia, international organizations

and states; and a series of background interviews with state representatives and legal scholars.[5] The report is primarily targeted at the governmental and non-governmental stakeholders that contribute to the international policy debate on AWS, including within the framework of the CCW GGE on AWS. It is particularly designed for legal advisers, diplomats and non-governmental experts who seek to: (*a*) deepen their understanding of rules structuring the ascription of responsibility for IHL violations; (*b*) identify issues that would make IHL violations involving AWS development and use potentially difficult to discern, scrutinize and attribute, and consequently difficult to prevent, intercept, investigate, adjudicate, penalize and remedy; (*c*) identify practical measures concerning the required types and degrees of human–machine interaction that would address those issues and operationalize the CCW's guiding principles (b) and (d) on human responsibility and accountability, respectively; and (*d*) identify whether and how the framework of state responsibility and individual criminal responsibility and the related IHL primary legal norms need to be further clarified, developed and observed, in order to both uphold respect for IHL and reduce challenges to holding actors legally responsible.

II. Report scope, research questions and structure

There are multiple legal frameworks through which human accountability for IHL violations that involve AWS may be ensured. This report focuses on two central frameworks, namely the rules governing state responsibility for internationally wrongful acts and individual criminal responsibility for war crimes. Other legal frameworks, such as domestic laws for administrative offences and corporate responsibility, while also critical to ensuring human accountability for IHL violations that involve AWS, are not addressed in this report, but they would deserve dedicated attention elsewhere.

The rules governing state responsibility for internationally wrongful acts and individual criminal responsibility for war crimes are two distinct (yet related) frameworks that serve a common aim: to hold actors accountable for their behaviour and its consequences, even in war (see figure 1.1). Both frameworks are essential to the promotion of justice and the rule of law, and they also serve an important preventative function. Under the framework governing state responsibility for internationally wrongful acts, a state bears responsibility for its acts or omissions that constitute violations of IHL. Notably, the violation of *any rule* of IHL applying to both the development and use of AWS could give rise to state responsibility. Under the framework governing individual criminal responsibility, an individual bears responsibility for certain serious breaches of IHL that amount to war crimes, which the individual commits, contributes to, orders or fails to prevent. In this way, the framework governing individual criminal responsibility complements the framework governing state responsibility—even for the same act or omission. For example, an unlawful attack involving the use of AWS could, under certain conditions, give rise to both state responsibility and individual criminal responsibility.

However, the unique characteristics of AWS (see box 1.1) raise a number of questions as to how existing rules governing state responsibility and individual criminal responsibility should be interpreted and applied. These include the fact that AWS are preprogrammed weapons that are triggered by their software programming interacting with the environment, rather than by direct user input.[6] The parameters for target identification, selection and engagement are also determined in advance, from

[5] The workshop was held on 8–10 Feb. 2022. The background interviews were conducted between Dec. 2021 and Apr. 2022. The workshop and interviews were all conducted under the Chatham House Rule.

[6] The decision-making process that leads to a use of force in military action, such as an attack with an AWS, involves different actors. This may mean that more than one person is considered the 'user' of an AWS. In this report, the term

Box 1.1. A working definition of autonomous weapon systems

There is no internationally agreed definition of 'autonomous weapon systems' (AWS). This report defines them as weapons that, once activated, can identify, select and apply force to targets, without human intervention. The term AWS is preferred to that of lethal autonomous weapon systems (LAWS)—although the latter is used in the mandate of the open-ended Group of Governmental Experts (GGE) on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems—because the concept of 'lethality' pertains to how the weapon system is used and its effects rather than the way it is designed. Moreover, AWS are capable of causing harm in the form of material damage or injury, irrespective of whether death was the intended or actual result.

AWS come in many shapes and forms, but at the core, they share several distinctive sociotechnical features that are essential for the legal analysis. First, AWS function based on preprogrammed target profiles and technical indicators that AWS can recognize through their sensors and software. Second, since AWS are triggered to apply force partly by their environment of use (rather than a user's input), aspects of a decision to apply force can be made further in advance than with traditional weapons, based on assumptions about the circumstances that will prevail at the time of the attack. A human operator may supervise and retain the possibility of overriding the system, but the system's default functioning is that human input is not required to identify and select targets, nor to apply force against them. These features mean that certain AWS can be operated in 'communications denied' environments and permit fast reaction time in decisions to use force. However, these features also mean that those who configure and deploy an AWS will not necessarily know the exact targets, location, timing or circumstances of the resulting use of force.

Sources: Moyes, R., 'Target profiles', Article 36 Discussion Paper, Aug. 2019; and Boulanin, V. and Verbruggen, M., *Mapping the Development of Autonomy in Weapon Systems* (SIPRI: Stockholm, Nov. 2017).

the design stage of the weapon to the activation of the system. Third, the parameterizing process involves multiple people along the command-and-control chain, not just the end user. Finally, AWS are programmed to attack targets based on generalized target profiles rather than specific people or objects. This means that, at the moment of activation, the user may not know *what* or *who*, *specifically*, the system will target, nor *where* and *when*, *precisely*, it will strike.

These unique features raise both old and new questions about what conduct amounts to a breach of IHL and how that unlawful conduct may be attributed to specific states and individuals. Answering these questions demands consideration of how existing legal frameworks address responsibility for unintentional, but not unlawful, harm and unforeseen incidents, as well as situations of distributed responsibility.

This report explores these issues in detail by reviewing the conditions necessary to attribute state responsibility for internationally wrongful acts and to impose individual criminal responsibility for war crimes. Underlying the discussion are the following four questions:

1. *What act or omission* in the development and use of AWS would give rise to state responsibility for an internationally wrongful act or individual criminal responsibility for a war crime (or both)?
2. *Whose conduct* in the development and use of AWS would give rise to state responsibility for an internationally wrongful act or individual criminal responsibility for a war crime (or both)?
3. *What standards of intent, knowledge, behaviour and care* on the part of agents of the state and individuals involved in the development and use of AWS—including developers, decision makers, planners, commanders and operators—would give rise to state responsibility for a breach of IHL or individual criminal responsibility for a war crime (or both)?
4. How in practice would IHL violations in the development and use of AWS be traced back to a particular state and particular individuals? That is, *how are IHL violations to be discerned, scrutinized and attributed*?

'user' refers to the person or group of persons who plans, decides on or carries out military action involving an AWS, and encompasses 'operators' and 'commanders'.

This report is structured as follows. Chapters 2 and 3 discuss the respective conditions necessary to assign state responsibility for internationally wrongful acts and to impose individual criminal responsibility for IHL violations amounting to war crimes, with a focus on how these elements would apply in relation to AWS. Chapter 4 discusses the practical processes through which IHL violations in the development and use of AWS could be discerned, scrutinized and attributed. Chapter 5 summarizes the project's key findings and presents recommendations for the international policy conversation on the governance of AWS.

2. State responsibility for internationally wrongful acts in the development or use of AWS

State responsibility is the baseline responsibility framework for violations of IHL. The legal framework governing state responsibility is often understood through the lens of primary and secondary rules. The primary rules are the norms, principles, rules and standards of international law, the violation of which gives rise to the responsibility of the state, and whose content is found in the international obligations applicable to the state, such as IHL. Secondary rules establish the conditions necessary for the international responsibility of the state to arise and its consequences. The secondary rules are enshrined in the International Law Commission's articles on 'Responsibility of States for Internationally Wrongful Acts' (ARSIWA),[7] which although not binding are widely considered to reflect customary norms and are therefore applicable to all states (see box 2.1).[8]

State responsibility covers breaches of any binding IHL norm.[9] As such, the state responsibility framework is meant to have a crucial preventative function in the context of AWS, since it can be triggered by a broad range of breaches, including those applicable at the AWS development stage. Moreover, state responsibility is a collective form of responsibility that takes into account the fact that obligations under IHL are often implemented through a web of agents of the state at multiple stages, rather than by one specific person at one particular moment.

Within policy and academic discussions, state responsibility is considered one of the possible applicable frameworks for violations of IHL in the development and use of AWS.[10] While it is uncontested that state responsibility applies, the contours and specificities of how this framework applies to violations of IHL in the context of AWS are relatively underexplored. This chapter looks at the conditions necessary to establish state responsibility for IHL violations involving AWS and, in light of those conditions, addresses two critical questions. What constitutes an act or omission contrary to an international law obligation in the development and use of AWS (section I)? And whose conduct in the development and use of AWS is attributable to the state and could thereby engage state responsibility (section II)? The chapter concludes with a summary of this chapter's main findings (section III).

[7] International Law Commission, 'Responsibility of states for internationally wrongful acts', Draft articles, Text adopted by the Commission at its 53rd session, 23 Apr. to 1 June and 2 July to 10 Aug. 2001, subsequently adopted by the United Nations General Assembly through Resolution No. 56/83 of 12 Dec. 2001 (ARSIWA). It must be acknowledged that, according to Art. 55, the ARSIWA does 'not apply where and to the extent that the conditions for the existence of an internationally wrongful act or the content or implementation of the international responsibility of a State are governed by special rules of international law'. On the overlaps or partial discrepancies between ARSIWA rules and IHL specific implementing rules and mechanisms, see Sassoli, M., 'State responsibility for violations of international humanitarian law', *International Review of the Red Cross*, vol. 84 (2012).

[8] Gaeta, P., Viñuales, J. and Zappalá, S., *Cassese's International Law* (Oxford University Press: Oxford, 2020), p. 248.

[9] This is in contrast to the framework of individual criminal responsibility, which is triggered by violations of IHL in the use of AWS (see chapter 3).

[10] For academic discussion see e.g. Boutin, B., 'State responsibility in relation to military applications of artificial intelligence', *Leiden Journal of International Law* (forthcoming 2022); Crootof, R., 'War torts', *New York University Law Review*, vol. 97 (2022); Geiss, R., 'State control over the use of autonomous weapon systems: Risk management and state responsibility', eds R. Bartels et al., *Military Operations and the Notion of Control Under International Law* (T.M.C. Asser Press: The Hague, 2021); and McFarland, T., 'Accountability', *Autonomous Weapon Systems and the Law of Armed Conflict* (Cambridge University Press: Cambridge, 2020). For policy discussion by the CCW Convention, GGE, see 'Draft report of the 2021 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems', CCW/GGE.1/2021/CRP.1, annex III, paras 18–21, 8 Dec. 2021; 'Switzerland's food for thought as requested by the Chair of the Group of Governmental Experts (GGE) on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems (LAWS) within the Convention on Certain Conventional Weapons (CCW)', Submission of Switzerland (2021); 'Elements for a future normative framework conducive to a legally binding instrument to address the ethical humanitarian and legal concerns posed by emerging technologies in the area of (lethal) autonomous weapons (LAWS)', Submission of Brazil, Chile and Mexico (2021); and 'US proposals on aspects of the normative and operational framework', Working paper submitted by the USA, CCW/GGE.1/2021/WP.3, 27 Sep. 2021.

I. Identifying the breach of an international law obligation: What acts and omissions in the development and use of AWS could trigger state responsibility?

For state responsibility to be triggered, there has to be a breach of one of the international obligations incumbent on the state. While several sets of rules, including international human rights and the rules governing the legality of the use of force by states (*jus ad bellum*), may separately or additionally apply to the development and use of AWS, the rules of IHL are considered central in the international policy debate and are therefore the focus of this chapter. IHL is a legal framework grounded in, among other treaties, the Geneva Conventions (GC I, GC II, GC III, GC IV) and the two additional protocols to the Geneva Conventions (AP I and AP II).[11] However, key rules and norms of IHL are recognized as being customary IHL, which means that they apply to all states, beyond those party to the Geneva Conventions and additional protocols (see table 2.1).

Determining what constitutes a breach of IHL depends on how the nature, content and scope of primary rules of IHL are interpreted. As laid down in Common Article 1 of the Geneva Conventions, states are obliged to not only 'respect' but also to 'ensure respect' for IHL. Clarifying what both commitments entail in relation to AWS is a critical first step to discerning what acts or omissions would engage state responsibility for an internationally wrongful act.

What do respecting and ensuring respect for IHL entail?

States have the duties to respect and ensure respect for IHL in all circumstances.[12] The duty to respect IHL is a reaffirmation of the obligatory force of IHL rules and the principle known as *pacta sunt servanda* (literally, agreements are to be kept), according to which states must perform all obligations arising from a treaty to which they are a party. The IHL obligations that states have to comply with—that is, perform or abstain from violating—can be divided into two categories: fundamental and facilitative (see table 2.1). The first category comprises the duty to respect the fundamental rules of IHL, which impose prohibitions and restrictions on weapons, means and methods of warfare. The second category comprises a set of obligations aimed at facilitating and securing respect for the fundamental rules of IHL. In other words, compliance with fundamental IHL rules partly or wholly depends on the fulfilment of the second category of IHL obligations. While both categories apply during an armed conflict, the latter applies in peacetime too. These facilitative obligations are thus instrumental to securing compliance with and respect for IHL fundamental rules and preventing violations.

The scope of the duty to 'ensure respect' is more contested than that of 'respect'. There is general agreement that it requires a state to ensure that IHL is implemented and applied at the national level—the so-called internal component of the duty. A narrow interpretation of the duty suggests that states must ensure respect for IHL in relation to actors whose conduct is already attributable to the state, such as its armed

[11] The four Geneva Conventions (GCs) and two additional protocols (APs) are: Geneva Convention (I) for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field (GC I); Convention (II) for the Amelioration of the Condition of the Wounded, Sick and Shipwrecked Members of Armed Forces at Sea (GC II); Convention (III) Relative to the Treatment of Prisoners of War (GC III); Convention (IV) Relative to the Protection of Civilian Persons in Time of War (GC IV), opened for signature 12 Aug. 1949, entered into force 21 Oct. 1950; Protocol I Additional to the 1949 Geneva Conventions, and relating to the Protection of Victims of International Armed Conflicts (AP I); and Protocol II Additional to the 1949 Geneva Conventions, and relating to the Protection of Victims of Non-International Armed Conflicts (AP II), opened for signature 12 Dec. 1977, entered into force 7 Dec. 1978.

[12] GCs (note 11), Common Art. 1; see also International Committee of the Red Cross (ICRC), Customary IHL Database, [n.d.], 'Rule 139. Respect for international humanitarian law'.

Box 2.1. Conditions necessary to establish state responsibility for internationally wrongful acts

The International Law Commission's articles on 'Responsibility of States for Internationally Wrongful Acts' (ARSIWA) recognize several principles for establishing state responsibility.[a] Underlying the ARSIWA as a whole is the principle that states are the main bearers of obligations under international law, and that a state is responsible for the conduct of persons or entities acting on its behalf or with its authorization or endorsement. The ARSIWA set out several conditions for establishing state responsibility for internationally wrongful acts. The following are relevant to the context of autonomous weapons systems:

- Any international wrongful act of the state triggers its international responsibility (art. 1).
- An internationally wrongful act comprises two elements: (*a*) there must be conduct, i.e. action or omission, not in conformity with the international obligations of the state (a *breach*); and (*b*) the breach is *attributable* to the state (art. 2).

In addition to these two basic requirements, there are five elements that are relevant to consider for the establishment of an internationally wrongful act by a state and for the invocation of state responsibility:

1. The rules on state responsibility *do not require 'fault'* on the part of the state agent as an element of a breach.[b] Whether this element is required depends on the primary obligation.
2. The rules on state responsibility *do not require damage* for a breach to occur; whether this condition is required depends on the primary obligation. However, the existence of injury, harm or damage is relevant in terms of the invocation of responsibility (see item 5 below).
3. Certain circumstances may preclude the wrongfulness of an act of a state not in conformity with one of its international obligations (arts 20, 21, 23, 24 and 25).
4. Certain *legal consequences* flow from the commission of an internationally wrongful act, including the obligations to cease the act (if it is continuing), to offer appropriate assurances and guarantees of non-repetition (if circumstances so require), and to make full reparation for the injury caused (arts 30–31).
5. State responsibility can be invoked by:
    - an injured state under three circumstances: (*a*) where the obligation is owed to the state individually (e.g. under a bilateral treaty); (*b*) in cases of multilateral obligations—including an obligation to the international community as a whole—in circumstances where the breach of the obligation '[s]pecially affects that State'; and (*c*) in cases such as disarmament treaties where the obligation 'is of such a character as radically to change the position of all the other States to which the obligation is owed with respect to the further performance of the obligation' (art. 42)
    - a non-injured state if 'the obligation breached is owed to a group of States including that State, and is established for the protection of a collective interest of the group' or the obligation breached 'is owed to the international community as a whole' (art. 48).

[a] International Law Commission, 'Responsibility of states for internationally wrongful acts', Draft articles, Text adopted by the Commission at its 53rd session, 23 Apr. to 1 June and 2 July to 10 Aug. 2001, subsequently adopted by the United Nations General Assembly through Resolution No. 56/83 of 12 Dec. 2001.

[b] For a discussion of whether some fundamental rules of international humanitarian law require fault, see section II in this chapter.

forces; private individuals and entities within the territory of the state and in any other place where the state exercises sufficient control and authority; and other persons or groups acting on behalf of the state.[13] The internal duty to ensure respect could entail a 'broad range of preventive, supervisory, and punitive measures', including domestic legislation and regulation, dissemination of IHL through education and training, and legal advice.[14]

The less settled dimension pertains to the so-called external component of the duty to ensure respect for IHL, that is, whether states have to ensure respect for IHL by actors other than those mentioned above. A broad interpretation suggests states must ensure respect in relation to other international actors, including other states.[15] It is

[13] See ICRC, 'Commentary [to GC I (note 11)] of 2016, Article 1: Respect for the Convention', 2016, paras 150–52; and Dörmann, K. and Serralvo, J., 'Common Article 1 to the Geneva Conventions and the obligation to prevent international humanitarian law violations', *International Review of the Red Cross*, vol. 96, no. 895/896 (2014), p. 709.

[14] Melzer, N., *International Humanitarian Law: A Comprehensive Introduction* (ICRC: Geneva, 2016), p. 268.

[15] Seixas-Nunes, A., 'Autonomous weapons systems and the procedural accountability gap', *Brooklyn Journal of International Law*, vol. 46, no. 2 (2021), p. 461; Wiesener, C. and Kjeldgaard-Pedersen, A., *State Responsibility for the Misconduct of Partners in International Military Operations: General and Specific Rules of International Law* (Djøf Publishing in cooperation with the Centre for Military Studies: Copenhagen, 2021), p. 73; and interpretation offered by a legal scholar, Interview with the authors, Online, 21 Dec. 2021.

also debated whether the duty to ensure respect concerns the Geneva Conventions only or the whole body of IHL.[16]

The obligation to ensure respect must be carried out, according to several authorities, with due diligence.[17] Due diligence can be defined as a standard of care or parameter that is used to assess states' implementation and compliance with obligations of conduct.[18] Due diligence is a central notion concerning state responsibility, but what it entails and requires from states is disputed in certain respects.[19] The relevance (and implications) of due diligence obligations in relation to AWS are addressed in section II of this chapter.

Having outlined the different range of obligations that states have to uphold, this report next considers what acts or omissions in the development and use of AWS would constitute a breach of IHL obligations.

What acts or omissions amount to a breach of 'fundamental' IHL rules?

The fundamental IHL rules that states have to *respect* can be divided into: (*a*) specific and general rules prohibiting or restricting specific weapons, means and methods of warfare ('weapons laws'); and (*b*) general prohibitions and restrictions on the conduct of hostilities ('targeting rules'). While the first category can be said to relate to whether a weapon, means or method of warfare is unlawful per se, the second category regulates how weapons, means and methods can be lawfully used.

In the context of a state developing or using AWS, determining what particular acts or omissions amount to a breach of these fundamental rules can be difficult for three main reasons. First, what the targeting rules prohibit or require is debatable, at least in certain respects. Second, the precise standards for assessing whether an AWS is indiscriminate, by its nature or through its use, are not necessarily settled. Third, AWS make it more challenging to identify whether a harmful incident is the result of an accident or a breach of an IHL obligation. This subsection explores these three sets of issues in turn.

What the targeting rules prohibit and require

The targeting rules oblige parties to armed conflict to comply with the principles of distinction, proportionality and precautions in attack. The precise acts or omissions that these rules prohibit and require, and the implications for the development and use of AWS, are contentious.

The most critical targeting rule in relation to AWS is the principle of distinction.[20] It prohibits making the civilian population (and other protected individuals and objects) the object of attacks and conducting indiscriminate attacks (see next subsection). As the primary principle of targeting, it also determines the ability of an attack to comply with the principle of proportionality in that proportionality cannot be assessed without first distinguishing the object of attack.[21]

[16] Hathaway, O. A. et al., 'Ensuring responsibility: Common Article 1 and state responsibility for non-state actors', *Texas Law Review*, vol. 95, no. 3 (2017), p. 566.

[17] View expressed by legal scholars, Interviews with the authors, Online, Dec. 2021 to Feb. 2022; Seixas-Nunes (note 15), p. 456; and Longobardo, M., 'The relevance of the concept of due diligence for international humanitarian law', *Wisconsin International Law Journal*, vol. 37, no. 1 (2019), pp. 183–84.

[18] Longobardo (note 17), pp. 183–84.

[19] Longobardo (note 17); Zhang, B., 'Accountability and responsibility for AI-enabled conduct', eds R. Geiss and H. Lahmann, *Handbook on Warfare and Artificial Intelligence* (Edward Elgar, forthcoming); Seixas-Nunes (note 15), p. 457; and Geiss (note 10).

[20] View expressed by state representatives consulted by the authors, Experts workshop, Online, 8 and 10 Feb. 2022.

[21] Van den Boogaard, J., 'Proportionality and autonomous weapons systems', *Journal of International Humanitarian Legal Studies*, vol. 6, no. 2 (2015), p. 261.

Table 2.1. Primary obligations of international humanitarian law applicable to states

Obligation	Source
A. Fundamental obligations	
Under IHL, any new weapon, means or method of warfare would be deemed inherently unlawful if it has one or more of the following characteristics:	
The weapon is of a nature to cause superfluous injury or unnecessary suffering	AP I, Art. 35(2); CIHL, Rule 70
The weapon is by nature indiscriminate (that is, weapons that cannot be directed at a specific military objective or its effects cannot be limited as required by IHL	AP I, Art. 51(4)(b)–(c); CIHL, Rule 71
The weapon is intended, or may be expected, to cause widespread, long-term and severe damage to the natural environment	AP I, Arts 35(3) and 55; CIHL, Rule 45
The weapon (or its injury mechanism) is already prohibited by a specific treaty	See relevant protocols[a] and CIHL, Rules 72–74 and 86
The weapon contradicts the principles of the law of nations as they result from the usages of international law, from the laws of humanity, and the dictates of public conscience (the Martens Clause)	GC I, Art. 63; GC II, Art. 62; GC III, Art. 142; GC IV, Art. 158; AP I, Art. 1(2)
In addition, IHL includes general prohibitions and restrictions on the conduct of hostilities:	
The **principle of distinction**, which obliges parties to an armed conflict to distinguish between the civilian population and combatants, between militarily active combatants and those *hors de combat*, and between civilian objects and military objectives, and accordingly to direct their operations only against military objectives. The principle of distinction prohibits making a civilian population, as well as individual civilians, the object of attack.	AP I, Arts 48, 51(2), 51(4), and 51(5); CIHL, Rules 1, 7 and 13
The **prohibition against indiscriminate attacks**, which prohibits attacks that are of a nature to strike military objectives and civilians or civilian objects without distinction, because such an attack: (*a*) is not directed at a specific military objective, (*b*) employs a method or means of combat which cannot be directed at a specific military objective, or (*c*) employs a method or means of combat the effects of which cannot be limited as required by IHL.	AP I, Arts 51(4) and 51(5)(a); CIHL, Rules 1 and 7
The **principle of proportionality**, which prohibits the conduct of an attack that may be expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, that is excessive in relation to the concrete and direct military advantage anticipated.	AP I, Art. 51(5)(b); CIHL, Rule 14
The **principle of precautions**: (*i*) *In the conduct of military operations*, parties must take constant care to spare the civilian population, civilians and civilian objects. (*ii*) *During an attack*, parties must (among other things): (*a*) take all feasible steps to verify that the objectives to be attacked are neither civilians nor civilian objects, nor subject to special protection, but are military objectives (*b*) take all feasible precautions in the choice of means and methods of attack to avoid, and in any event minimize, incidental loss of civilian life, injury to civilians and damage to civilian objects (*c*) do everything feasible to assess whether the effects of the attack may be expected to violate the principle of proportionality (*d*) 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 violate the principle of proportionality.	(*i*) AP I, Art. 57(1); CIHL, Rule 17 (*ii*) AP I, Art. 57(1); CIHL, Rules 15, 16, 18 and 19
Special protections should be ensured to (among others) medical units, religious personnel, cultural property, the natural environment, persons *hors de combat*.	GC I, Arts 19 and 24; GC II, Art. 36; GC IV, Arts 18 and 33; AP I, Arts 12, 15, 35(3), 41(1) and 55; CIHL, Rules 27, 28, 38, 44, 45, 47 and 14

Obligation	Source
B. Facilitative obligations	
To ensure compliance with the fundamental rules, states must comply with a number of facilitative rules, which include, but are not limited to, the following obligations:	
To conduct a legal review of new weapons, means and methods of warfare[b]	AP I, Art. 36
To provide legal advisers to armed forces	AP I, Art. 82; CIHL, Rule 141
To disseminate IHL to the wider public, including education and training in IHL to the armed forces	GC I, Art. 47; GC II, Art. 48; GC III, Art. 127; GC IV, Art. 44; AP I, Art. 83; CIHL, Rule 142
To repress grave breaches of IHL and suppress all other breaches	GC I, Art. 49(3); GC IV, Art. 146; AP I, Arts 85 and 86(1)

AP I = Additional protocol I to the Geneva Conventions (note 11); CIHL = customary IHL; GC I, II, III, IV = Geneva Conventions I, II, III and IV; IHL = international humanitarian law.

[a] See e.g. Convention on the Prohibition of the Development, Production and Stockpiling of Bacteriological (Biological) and Toxin Weapons and on their Destruction (Biological and Toxin Weapons Convention, BWC), opened for signature 10 Apr. 1972, entered into force 26 Mar. 1975; Convention on the Prohibition of the Development, Production, Stockpiling and Use of Chemical Weapons and on their Destruction, opened for signature 13 Jan. 1993, entered into force 29 Apr. 1997; Protocol for the Prohibition of the Use of Asphyxiating, Poisonous or Other Gases, and of Bacteriological Methods of Warfare, Geneva, signed 17 June 1925, entered into force 8 Feb. 1928; and Protocol on Blinding Laser Weapons (Protocol IV) to the CCW Convention, issued 13 Oct. 1995, entered into force 30 July 1998.

[b] This is the only obligation in this table not recognized as customary international law. States not party to AP I are, therefore, not bound by it.

Sources: Geneva Conventions and additional protocols (note 11); and International Committee of the Red Cross, Rules, Customary IHL Database, [n.d.].

An important interpretative question in the use of AWS is how to determine when the principle of distinction has been violated. This question may depend on whether the principle of distinction covers only *direct and deliberate* attacks or whether it also covers attacks conducted with *negligence* (i.e. 'a demonstrable and inexcusable failure to take the degree and kind of care that might reasonably have been expected' in the circumstances), and *erroneous or inadvertent* (unintended) attacks against protected persons or objects.[22] Some experts have observed that 'nothing in the formulation of the primary rules prohibiting indiscriminate attacks and attacks with excessive incidental effects indicates that their scope is limited to deliberate attacks', but also note that 'the prohibition codified in Articles 51(2) and 52(1) of Protocol I—that civilians and civilian objects "shall not be the object of attack"—may be seen as covering only deliberate attacks'.[23]

Others debate the 'fault' requirement, especially in the context of erroneous or inadvertent attacks.[24] Some argue that because a breach of this rule requires that civilians be *made the object* of an attack, inadvertently or erroneously attacking civilians or civilian objects with an AWS does not constitute a breach of the principle of distinction.[25] However, some types of inadvertent attacks—including those that reflect recklessness or indifference as to whether a target is civilian or military in character—could amount to a violation of the prohibition on indiscriminate attacks, which is a specification of the principle of distinction.[26] Settling this interpretative question

[22] Boothby, W., *The Law of Targeting* (Oxford University Press: Oxford, 2012) pp. 176–77.

[23] Sassòli, M. and Quintin, A., 'Active and passive precautions in air and missile warfare', *Israel Yearbook on Human Rights*, vol. 44 (2014), p. 16; and Dinstein, Y., *The Conduct of Hostilities under the Law of International Armed Conflict*, 3rd edn (Cambridge University Press: Cambridge, 2016), p. 117.

[24] Geiss (note 10); Dinstein (note 23), p. 117; and Jain, A., 'Autonomous military capabilities, errors and responsibility under IHL', *Journal of International Criminal Justice*, vol. 20 (forthcoming 2023).

[25] Casey-Maslen, S. and Haines, S., *Hague Law Interpreted: The Conduct of Hostilities Under the Law of Armed Conflict* (Hart: Oxford, 2018), pp. 108, 157–58; and Jain (note 24).

[26] Dinstein (note 23), p. 117; and Jain (note 24).

around whether a breach has occurred is increasingly relevant to AWS, as their use is arguably associated with an increased risk of inadvertent harm to civilians.[27] The question of risk is addressed in the discussion below in this section, 'Distinguishing accidents from errors that constitute breaches of IHL obligations'.

The principle of proportionality prohibits an attack which may be *expected* to cause collateral damage that is excessive with respect to the anticipated military advantage; it does not prohibit attacks that otherwise *cause* excessive collateral damage. Thus, unless excessive collateral damage is *expected*, launching a disproportionate attack does not of itself constitute a breach of this rule. Since decisions to use force involving AWS may be made well in advance, an important question is how far in advance a proportionality assessment can be made to reasonably foresee the likely effects. Related to that, a fundamental interpretative question in the use of AWS is how to determine when the principle of proportionality has been violated and, notably, the standard to which errors in the conduct of proportionality analysis should be evaluated.[28] Another complicating factor is the extent to which a proportionality assessment involving an AWS may rely on technical indicators rather than, or in addition to, qualitative human assessments. These questions have critical implications for determining what AWS use constitutes a violation of the principle of proportionality that triggers state responsibility.

Parties to a conflict are obliged to comply with the principle of precautions in attack—that is, to take 'constant care' to spare civilians in military operations and to take several precautionary measures regarding attacks.[29] These are, however, open-textured obligations that even without the involvement of AWS raise several interpretative questions. For example, what does it mean 'to take *all feasible* precautions', which is a context-dependent and due diligence standard, and what standards of training and levels of technical knowledge are required of those planning or deciding upon attacks?[30] Moreover, the nature of AWS as preprogrammed weapons that entail key targeting decisions to be taken at earlier phases of the targeting cycle, may reconfigure the temporal aspect of the duty to take feasible precautions.[31] A key question is whether the temporal scope of the rule is limited to the *moment* of attack or extends to *phases preceding the attack*, including in the programming phase. The latter interpretation relies partly on the fact that the duty of constant care refers more generally to 'the conduct of military operations', not just attacks.[32] Questions around the temporal thresholds for AWS also apply to the principles of distinction and proportionality. Whether the principles could apply to phases preceding an attack depends on how broadly the concept of attack is interpreted. For example, Article 49 in AP I defines an attack but does not specify when an attack begins and ends.[33] These interpretations have implications for *who* is bound by these rules, which become a question of *whose conduct* can trigger state responsibility (section II in this chapter).

In determining what specific acts or omissions constitute a breach of the principles of distinction, proportionality and precautions in attack when AWS are involved, a key focus is the extent to which the necessary evaluative assessments are delegated

[27] See e.g. Seixas-Nunes (note 15); Geiss (note 10); and Scharre, P., *Autonomous Weapons and Operational Risk* (Center for a New American Security: Washington, DC, 2016), p. 5.

[28] Jain (note 24).

[29] AP I (note 11), Art. 57(1).

[30] See e.g. Longobardo (note 17); and US Department of Defense (DOD), *Department of Defense Law of War Manual* (US DOD, Office of General Counsel: Washington, DC, 2015), pp. 192–94 (§5.2.3.2).

[31] 'Targeting cycle' is a term adopted by many militaries to capture the entire targeting process and includes steps spanning from 'decide and 'detect' to 'deliver' and 'assess'.

[32] AP I (note 11), Art. 57(1); and View expressed by legal scholar, Interview with the authors, Online, 22 Feb. 2022.

[33] AP I (note 11), Art. 49. See Boulanin, V., Bruun, L. and Goussac, N., *Autonomous Weapon Systems and International Humanitarian Law: Identifying Limits and the Required Type and Degree of Human–Machine Interaction* (SIPRI: Stockholm, 2021), pp. 23–24.

to machines that rely on data, sensors and algorithms, or are the result of context-based value judgements made by humans.[34] This issue matters for state responsibility because the way in which the principles of distinction, proportionality and precautions in attack are interpreted—especially regarding what they demand in terms of the exercise of human agency, such as the extent to which a state supervises or otherwise controls an AWS—will inform what constitutes a breach of IHL.

The standards for assessing whether an AWS by nature or by use is indiscriminate

Under IHL, indiscriminate attacks are prohibited. This prohibition forbids the indiscriminate *use* of a weapon, and also provides two criteria to define weapons that are *by nature* indiscriminate, namely: (*a*) the weapon is incapable of being directed against a specific military objective; and (*b*) the effects of the weapon cannot be limited as required by IHL. However, 'there are differing views on whether the rule itself renders a weapon illegal or whether a weapon is illegal only if a specific treaty or customary rule prohibits its use'.[35]

In the context of AWS, there are two questions: on what basis may an AWS be deemed by *nature* indiscriminate; and on what basis would the *use* of an AWS amount to an indiscriminate attack? Answering the first question depends on conceptual notions, such as what 'by nature' means, and the technical standards of AWS, which are not laid down in IHL with great specificity. The second depends on what a user of an AWS should know and do in relation to its deployment, which is also unsettled in certain key respects.

There are divergent opinions as to whether AWS are by nature indiscriminate.[36] This is a complex question whose answer cannot be expressed in definitive terms under existing IHL, for at least two reasons. First, there are multiple ways in which to evaluate the 'indiscriminate nature' of AWS: (*a*) type of weapon payload; (*b*) basis of target recognition (i.e precision of target profile); and (*c*) the applicable standard of reliability and foreseeability (i.e the extent to which the behaviour and effect of the AWS can be predicted).[37] Second, it is difficult to establish hard metrics for each of these variables, partly because the determination of the acceptable threshold is context-dependent and subject to different understandings. For example, is an AWS that fails to identify the target 5 per cent of the time considered indiscriminate? Should AWS performance be benchmarked to human performance?[38] Arguably these questions cannot be fully answered in the abstract and need to be evaluated on a case-by-case basis and relative to similar types of weapons.[39]

[34] See e.g. CCW Convention, GGE, 'Joint submission on possible consensus recommendations in relation to the clarification, consideration and development of aspects of the normative and operational framework on emerging technologies in the area of lethal autonomous weapons systems', submitted by Austria, Brazil, Chile, Ireland, Luxembourg, Mexico, and New Zealand, June 2021; and ICRC, 'International Committee of the Red Cross (ICRC) position on autonomous weapon systems: ICRC position and background paper', *International Review of the Red Cross*, no. 915 (Jan. 2022).

[35] ICRC (note 12), 'Rule 71. Weapons that are by nature indiscriminate'.

[36] See e.g. Thurnher, J. S., 'Means and methods of the future: Autonomous systems', eds P. A. L. Ducheine, M. N. Schmitt and F. P. B. Osinga, *Targeting: The Challenges of Modern Warfare* (Asser Press: The Hague, 2016); and Boothby, W. H., 'Highly automated and autonomous technologies', ed. W. H. Boothby, *New Technologies and the Law in War and Peace* (Cambridge University Press: Cambridge, 2018), pp. 137 and 146.

[37] Predictability is not the same as reliability, which refers to the extent to which a system does or does not fail: 'Even exceptionally reliable systems that fail rarely might still occasionally fail in very unpredictable ways because the range of failures that the system can exhibit is wide.' Holland Michel, A., *The Black Box, Unlocked: Predictability and Understandability in Military AI* (United Nations Institute for Disarmament Research, UNIDIR: Geneva, 2020), p. 5.

[38] Intimacies of Remote Warfare, 'The ambiguities of precision warfare', Utrecht University, 12 June 2020; Henderson, I., Keane, P. and Liddy, J., 'Remote and autonomous warfare systems: Precautions in attack and individual accountability', ed. J. D. Ohlin, *Research Handbook on Remote Warfare* (Edward Elgar: Cheltenham, 2019), pp. 5–6; and Copeland, D. and Sanders, L., 'Holding autonomy to account: Legal standards for autonomous weapon systems', *Articles of War*, 15 Sep. 2021.

[39] Backstrom, A. and Henderson, I., 'New capabilities in warfare: An overview of contemporary technological developments and the associated legal and engineering issues in Article 36 weapon reviews', *International Review of the Red Cross*, vol. 94, no. 886 (Summer 2012).

The question of whether an AWS is by nature indiscriminate depends partly on whether its performance, behaviour and effects are sufficiently and reasonably foreseeable by one or more humans for the duration of the attack or operation. Compliance with the rule against indiscriminate attacks requires the human user of a weapon to sufficiently foresee its likely effects and to administer the weapon—including its effects—during use.[40] When it comes to human ability to foresee a weapon's effects, it is unclear what kind and degree of knowledge about the weapon and its environment of use is required—IHL does not contain an explicit technical knowledge requirement. However, the technical complexity associated with many AWS raises the question of whether, to comply with IHL, users of an AWS need a specific technical understanding of the weapon.[41] Also, several factors related to the intended and expected environment of use, such as weather conditions as well as the presence of civilians and civilian objects, have consequences for the foreseeability of the behaviours and effects of an AWS. In this regard, experts have pointed to the inability of simulated environments to properly test complex environments.[42] In light of these complexities, there needs to be a further elaboration on the required conceptual and technical standards of foreseeability in the use of an AWS, so that the corresponding standards and degree of care for compliance with IHL, and the prohibition on indiscriminate attacks, can be identified.

On the question of how AWS should be administered during use, IHL provides no explicit guidance on the types and degrees of human–machine interaction needed to comply with the prohibition on indiscriminate attacks. This question has been discussed extensively in the GGE; while states have yet to reach a consensus, they agree that requirements for human involvement are context-dependent.[43] Further clarification is needed to identify what use cases involving AWS would constitute a breach of the prohibition on indiscriminate attacks.

Distinguishing accidents from errors that constitute breaches of IHL fundamental obligations

While AWS in some circumstances potentially provide better protections to civilians through more precise and accurate targeting than traditional types of weapons, their use is also apparently associated with a risk of accidents and errors that inflict unintended harm or injury to protected individuals and objects.[44] 'Normal accident' theory suggests that in tightly coupled complex systems—such as modern military weapon systems, including AWS—accidents are 'inevitable' over a long enough time horizon.[45] It has been argued that AWS could be more prone to accidents, with sources of accidents arising from the risk of hacking, unexpected interactions with the environment, simple malfunctions and software errors.[46]

To what extent a harmful event resulting from an accident or error gives rise to state responsibility is debated, and depends, among other factors, on the sources of failures and standards of care. For example, at the GGE in the context of New Zealand comments about the risk of accidents that result in civilian casualties, as opposed to systems that are *designed* to commit violations, the United States has argued that such accidents are 'sometimes tragic and unavoidable' but do not necessarily imply a vio-

[40] Boulanin, Bruun and Goussac (note 33), p. 13.

[41] Boulanin, Bruun and Goussac (note 33), pp. 23–24.

[42] Boulanin, V, 'Implementing Article 36 weapon reviews in the light of increasing autonomy in weapon systems', SIPRI Insights on Peace and Security (2015).

[43] CCW Convention, GGE, 'Commonalities in national commentaries on guiding principles', 2020.

[44] Atherton, K., 'Understanding the errors introduced by military AI applications', Brookings TechStream Blog, 6 May 2022; Seixas-Nunes (note 15); and ICRC, *International Humanitarian Law and the Challenges of Contemporary Armed Conflicts* (ICRC: Geneva, 2019), p. 32.

[45] Perrow, C., *Normal Accidents: Living with High-risk Technologies* (Oxford University Press: Oxford, 1999); and Crootof (note 10), p. 113.

[46] Scharre (note 27), p. 5.

lation of IHL.[47] In the specific context of AWS, the US view is that unintended harm to civilians and protected persons arising from an accident or equipment malfunction 'is not a violation of IHL as such'; however, Switzerland argues that states 'remain legally responsible for unlawful acts and resulting harm caused by autonomous weapon systems they employ, including due to malfunction or other undesired or unexpected outcomes'.[48]

To distinguish a tragic, but not unlawful, accident from a breach of IHL fundamental rules arising from the use of AWS, it is first of all important to understand the different types of errors and failures that could lead to accidents.[49] Developing technical and legal criteria to distinguish between accidents and breaches of IHL is thus critical to establishing state responsibility arising from unintended harm involving AWS.

For technical criteria, designers and engineers could in the development phase come up with a list of possible failures and categorize them as, for example, 'accident' or 'error'. Such categorization would improve the ability to assess the sources of failures and whether someone is to be held responsible. However, this categorization task arguably becomes increasingly complex due to the technical complexities and the unpredictability associated with AWS.

The legal criteria depend on at least two related issues. First, they depend on the aforementioned interpretative debates around whether and to what extent attacks erroneously or inadvertently directed against civilians or civilian objects or errors in conducting proportionality assessments constitute breaches of IHL giving rise to state responsibility. Second, whether an error in targeting amounts to a breach of IHL ultimately depends on the kind and degree of care, and the associated standards, that must be exercised when taking precautions in attack. In other words, what is the required degree of care that decision makers, such as planners, commanders and operators (who may also overlap), have to exercise to ensure that the AWS only attacks intended military objectives and to avoid or at least minimize civilian harm? Part of the answer lies in the obligations to verify that the target is a military objective and not protected by IHL, and 'to do everything that is practicable or practically possible' (feasible) to prevent civilian harm.[50] The specific actions and omissions required by these obligations depend on the type of military objective and the environment of use of an AWS, and on the standard of care required by the obligation to take precautions.

Objectives that are military by nature (such as a military base) might, for instance, allow for more reliance on technical indicators and automated information in target verification, than objectives that are military by location, purpose or use (such as a border area, bridge or building), which require a different kind and degree of target verification.[51]

Some experts have suggested that the required standard of care is the 'reasonable commander standard'—whether the decision maker did what a *reasonable* person would have done in the circumstances and with the information *reasonably* available to them at the relevant time.[52] Along the same lines, others have argued that, for a

[47] Acheson, R. and Pytlak, A., 'Autonomous weapons and questions of ethics, control and accountability', *CCW Report*, vol. 10, no. 4 (3 June 2022).

[48] CCW Convention, GGE, 'US proposals on aspects of the normative and operational framework', Working paper submitted by the USA, CCW/GGE.1/2021/WP.3, 27 Sep. 2021, p. 4; and CCW Convention, GGE, 'A "compliance-based" approach to autonomous weapon systems', Working paper submitted by Switzerland, CCW/GGE.1/2017/WP.9, 10 Nov. 2017, p. 6.

[49] View expressed by legal experts, Experts workshop, Online, 8 Feb. 2022; see also Holland Michel, A., *Known Unknowns: Data Issues and Military Autonomous Systems* (UNIDIR: Geneva, 2021).

[50] Sassòli and Quintin (note 23) p. 12.

[51] Boothby (note 22), pp. 123–35.

[52] See e.g. Boothby (note 22), pp. 171–72 and 191–92; Rogers, A. P. V., *Law on the Battlefield* (Manchester University Press: Manchester, 2012), pp. 150–51; Commission Reporting to the Prosecutor for the International Criminal Tribunal

person to be a target, a commander must be 'reasonably convinced that the individual is a combatant'.[53] These standards, however, remain debated in IHL scholarship. Uncertainties around these standards are partly a result of the lack of clarity around standards of intent, knowledge and foreseeability demanded by IHL rules.[54] Others have argued that the presumption of civilian status—despite not being recognized by all states as international customary law, and not applicable to states not parties to the additional protocols—could be a useful standard.[55] The presumption of civilian status states that in situations of doubt as to whether a person or object is civilian, the person or object must be presumed to be civilian and thus an attack should not be launched.[56] However, the threshold of doubt triggering such presumption is not fully settled.[57]

What acts or omissions amount to a breach of the 'facilitative' IHL obligations?

States are also obliged to respect another set of rules in the development and use of AWS. These norms are expressed in the form of positive actions that states must carry out to secure respect for IHL fundamental rules, and are thus preventative, or 'facilitative', in nature.[58] These include obligations to conduct a legal review of new weapons, means and methods of warfare; provide legal advisers to their armed forces; and disseminate IHL to the armed forces (table 2.1). These norms are recognized as having customary international law status. The exception is the obligation to carry out legal reviews, which means, in practice, that state responsibility for violations of this obligation is only engaged for a state party to AP I. However, the GGE's guiding principle (e) recognizes the importance of conducting legal reviews in the development and use of AWS. Thus states not bound by AP I still acknowledge this obligation, even though the non-binding nature of the guiding principles means non-party states cannot be held legally responsible for not conducting a legal review.[59]

To establish the acts and omissions that give rise to state responsibility for failing to comply with the facilitative IHL obligations in the context of AWS, this subsection explores two issues: the lack of clear metrics to assess non-performance of these rules and challenges around imposing state responsibility for their non-performance.

Lack of clear metrics for assessing what constitutes non-performance of a facilitative obligation

A state's non-performance of a facilitative obligation constitutes a breach that may engage its international responsibility. However, determining whether such a breach has occurred is difficult because there are no clear metrics for assessing what conduct amounts to non-performance of a facilitative obligation. Besides uncertainties

for Yugoslavia on the NATO Bombing Campaign in 1999, Final Report to the Prosecutor, 13 June 2000, para. 28; and Jain (note 24).

[53] Haque, A., 'Killing in the fog of war', *Southern California Law Review*, vol. 86, no. 63 (2012).

[54] See e.g. Chengeta, T., 'Accountability gap, autonomous weapon systems and modes of responsibility in international law', *Denver Journal of International Law and Policy*, vol. 45, no. 1 (2016), p. 24; Copeland and Sanders (note 38); and Dunlap, C. J., 'Accountability and autonomous weapons: Much ado about nothing?', *Temple International & Comparative Law Journal*, vol. 30, no. 1 (2016), p. 70. See also Schmitt, M. N. and Schauss, M., 'Uncertainty in the law of targeting: Towards a cognitive framework', *Harvard National Security Journal*, vol. 10, no. 1 (2021); and Henderson, I. and Reece, K., 'Proportionality under international humanitarian law: The "reasonable military commander" standard and reverberating effects', *Vanderbilt Journal of Transnational Law*, vol. 51, no. 3 (2018).

[55] US DOD (note 30), pp. 200–01 (§5.4.3.2).

[56] Bothe, M. et al., *New Rules for Victims of Armed Conflicts: Commentary on the Two 1977 Protocols Additional to the Geneva Conventions of 1949*, 2nd ed (Martinus Nijhoff Publishers: The Hague, 2013) p. 336.

[57] Schmitt, M. N. (gen. ed.), *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (Cambridge University Press: Cambridge, 2017), pp. 424, 448–50; and Program on Humanitarian Policy and Conflict Research at Harvard University, *HPCR Manual on International Law Applicable to Air and Missile Warfare* (Cambridge University Press: Cambridge, 2013) pp. 90–92. For a survey of some such analyses, see, Schmitt and Schauss (note 54).

[58] See AP I (note 11), Art. 80; and ICRC, 'Commentary [to AP I, Art. 80] of 1987: Measures for execution', 1987, para. 3297.

[59] CCW Convention, GGE, 'Guiding principles' (note 3).

around *what conduct* is required by facilitative obligations, the unique characteristics of AWS also raise additional questions around *when and how*—that is, through which processes—the facilitative obligations are implemented. For example, the obligation to conduct legal review has been extensively debated in the GGE, but the prospects of holding a state responsible for non-performance will remain limited until further clarity and consensus are reached regarding the specificities of its application. That is because there are few (if any) internationally agreed standards regarding how to conduct a legal review of a new weapon, means or method of warfare; what, specifically, should be reviewed?[60] In addition to existing uncertainties around the legal review of any weapon, AWS pose new questions regarding the timing and basis for conducting a review.[61] This is especially the case with AWS that include 'self-learning' capabilities, which likely require more frequent, and potentially continuous, reviews. While there is some agreement that any modifications which alter the functioning, behaviour and effects of an AWS in a way that affects the application of IHL would most likely trigger a new review, it remains unclear how such a modification is identified and on what parameters.[62]

The obligation to provide legal advisers to the armed forces requires states to make legal advisers available 'when necessary' and at the 'appropriate [command] level'.[63] These flexible terms already result in various interpretations concerning who should receive legal advice and when. This lack of clarity around temporal aspects of implementation is particularly reinforced in the case of AWS, where its preprogrammed nature suggests that states need to make legal advisers available in the design and programming phase.[64] A further question is what degree of technical knowledge legal advisers need to possess, when the effects of an AWS depend on how its sensors and software interact with the environment and how they recognize preprogrammed target profiles and technical indicators.[65]

The obligation to disseminate IHL 'as widely as possible', including integrating it into military instruction, is reflected in all four Geneva Conventions and the two additional protocols (table 2.1). However, the obligation is silent on the methods for its effective implementation.[66] In relation to AWS, clarification is especially needed in terms of the *type* of training the obligations require. For example, the extent to which training and instruction obligations also extend to training in specific weapons, means and methods—including technical understandings of complex weapon systems such as AWS—is unclear. AP I does not seem to go as far as requiring specific military training for specific weapons, but this is arguably an implicit requirement flowing from obligations to 'ensure respect' and 'take all feasible precautions'.[67] Despite inconsistencies, state practice shows that military manuals increasingly contain provisions regarding autonomous weapons, cyber operations and unmanned systems.[68] This supports the contention that training on the use of specific weapons

[60] Boulanin, Bruun and Goussac (note 33), pp. 28–35; and ICRC, *Guide to Legal Reviews* (forthcoming 2022 or 2023).

[61] Boulanin, Bruun and Goussac (note 33), pp. 29–35.

[62] Boulanin, Bruun and Goussac (note 33), p. 33; Farrant, J. and Ford, C. M., 'Autonomous weapons and weapon reviews: The UK Second International Weapon Review Forum', *International Law Studies*, vol. 93 (2017); and CCW Convention, GGE, 'Chairperson's Summary', CCW/GGE.1/2020/WP.7, 19 Apr. 2020.

[63] AP I (note 11), Art. 82.

[64] Vazquez, A., 'LAWS and lawyers: Lethal autonomous weapons bring LOAC issues to the design table, and judge advocates need to be there', *Military Law Review*, vol. 228, no. 89 (Mar. 2020), p. 119.

[65] Boulanin, Bruun and Goussac (note 33),p. 37.

[66] Rossi, A., 'Training armed forces in IHL: Just a matter of law?', *Opinio Juris*, 8 Oct. 2020; and Longobardo, M., 'Training and education of armed forces in the age of high-tech hostilities', eds E. Carpanelli and N. Lazzerini, *Use and Misuse of New Technologies: Contemporary Challenges in International and European Law* (Springer: Cham, 2019).

[67] Longobardo (note 66), p. 84.

[68] See e.g. US DOD (note 30),pp. 352–55 (§6.5.8–6.5.9) and ch. 16; German Federal Ministry of Defence, *Law of Armed Conflict: Manual* (Federal Ministry of Defence: Berlin, May 2013), p. 74, para. 486; and British Ministry of

is needed to help ensure respect for IHL.[69] These interpretative questions become increasingly relevant in the context of AWS, where clarifying the acts, and particularly the omissions, that flow from a state's training and instructions in AWS development and use is critical for assessing whether the state is responsible for non-compliance with the obligation. While several states have already stressed the need to provide AWS-specific training to their armed forces, debate in the GGE around AWS has not systematically addressed the question.[70] Clarification around the issue is warranted, not least because a breach of the facilitative training obligation could lead to violations of fundamental rules of IHL.[71]

The final critical facilitative obligation is the duty to investigate alleged grave breaches of the Geneva Conventions (which amount to war crimes) and to prosecute or extradite suspected perpetrators, regardless of their nationality and other jurisdictional links. However, the methodologies for how states investigate war crimes and the scope of whom to prosecute remain subject to interpretation. The AWS-specific issues related to this obligation are extensively addressed in chapter 4.

The difficulty of imposing state responsibility for non-performance of facilitative obligations

The facilitative IHL obligations are obligations of conduct. While obligations of result, such as the duty to criminalize war crimes, require states to obtain a certain result, obligations of conduct require states to 'make every effort' towards a certain goal or outcome. For obligations of conduct, states are not responsible 'for a possible failure of their efforts as long as they have done everything reasonably in their power to fulfil these obligations'.[72] Determining whether an obligation of conduct has been breached can be challenging because the assessment is based on the lack of performance of a certain *behaviour* rather than on failure to obtain a certain *result*.[73] Moreover, some obligations of conduct, such as rules regarding the dissemination of and training in IHL, must be discharged with due diligence.[74] However, what states have to do to discharge due diligence obligations, and in turn which omissions would amount to a breach of the obligations, remains unclear in certain respects.[75] This is an inherent challenge of IHL that, like many others, is exacerbated by AWS.

Other inherent challenges are the lack of a monitoring system and lack of transparency concerning the measures states take to fulfil these positive obligations, such as their processes for conducting legal reviews or investigating serious breaches, which they are not obliged to share. The lack of information compounds the difficulty in ascertaining the occurrence of a breach giving rise to state responsibility. For example, in relation to the obligation to conduct a legal review, if a state's review finds that a weapon is illegal, the state is not obliged to make its findings public, and consequently is 'not bound to reveal anything regarding new weapons which are being developed or manufactured' (except to the extent required by the 'implementing laws and regulations' articles in the Geneva Conventions).[76]

Defence, *UK Air and Space Power: Joint Doctrine Publication 0-30*, 2nd edn (Development, Concepts and Doctrine Centre: Swindon, 2017), paras 2.3, 2.20, 2.21 and 4.15.

[69] Longobardo (note 66), pp. 86–87.

[70] See e.g. US DOD, Directive no. 3000.09, 21 Nov. 2012 (updated 8 May 2017), Enclosure 4, para. 3; and CCW Convention, GGE, 'Australia's system of control and applications for autonomous weapon systems', Working paper submitted by Australia, CCW/GGE.1/2019/WP.2/Rev.1, 26 Mar. 2019.

[71] United Nations, 'Eritrea–Ethiopia Claims Commission, Partial Award: Central Front—Ethiopia's Claim 2', *Reports of International Arbitral Awards*, vol. 26 (28 Apr. 2004).

[72] ICRC, 'Commentary [to GC I (note 11)] of 2016, Article 1: Respect for the Convention' (note 13), para. 119.

[73] Longobardo (note 17), pp. 183–84.

[74] Longobardo (note 17), p. 186.

[75] ICRC, 'Commentary [to the GC I (note 11)] of 2016, Article 1: Respect for the Convention' (note 13), para. 164.

[76] ICRC, 'Commentary [to AP I (note 11), Art. 36] of 1987: New weapons', 1987, para. 1481. For the exceptions, see e.g. GC I (note 11), Art. 48,

Another challenge is that state responsibility for violations of facilitative IHL obligations is rarely invoked unless such violations are connected with a breach of fundamental IHL rules. In theory, a violation of any rule of IHL, either fundamental or facilitative, triggers the responsibility of the state that has committed it. However, in practice, the violation of a facilitative obligation is in itself unlikely to injure another state and thus, in the absence of an injured state, state responsibility is unlikely to be invoked (box 2.1).[77]

II. Attributing the breach of an international law obligation: Whose conduct in the development and use of AWS could trigger state responsibility?

For state responsibility to be triggered, a breach of IHL must be attributable to the state. Articles 4 to 11 of the ARSIWA set forth the criteria for determining which persons' and entities' wrongful acts are attributable to a state and trigger its international responsibility (box 2.2).

In general terms, state responsibility is triggered by state 'organs', persons and entities over which the state has authority or control, or whose actions the state endorses. It is important to note that states are collective entities that act through human agents, and it is thus human acts and omissions that trigger state responsibility. However, several states contend that designating particular persons and entities as agents of a particular state is left to the discretion of the state.[78] Nevertheless, this section considers how the responsibility to comply with IHL is shared and diffused among different actors in the context of the development and use of AWS. It addresses three different avenues of attribution: when and under which conditions a state's responsibility in the development and use of AWS is engaged by, respectively, state agents, private actors and other states' armed forces.

State agents

States perform their obligations to comply with IHL across a number of agents whose conduct is directly attributable to the state.[79] However, the unique characteristics of AWS raise three key questions around such attribution.

The first question relates to the distribution of responsibilities in the use and development of AWS. According to some experts, the potential for AWS to transform, and perhaps reconfigure, traditional decision-making structures is significant, with implications for who within the military chain of command and control, including at the different levels of the military hierarchy and within the political sphere, makes decisions on the use of force that involves an AWS.[80] That is because decision-making processes leading to the use of force may be distributed across a large number of actors at the strategic, operational and tactical levels, both before and during an attack. Conversely, others argue that AWS do not pose challenges to existing divisions of role and responsibilities within command and control structures.[81] In their view, while a newly introduced weapon will always create new roles and responsibilities, and will trigger adjustment in structures, the command structures arguably

[77] Aside from breaches of *erga omnes* obligations (i.e. obligations owed to all), injury is a requirement for invoking state responsibility. ARSIWA (note 7), Arts 42 and 48.

[78] View expressed by state representatives, Experts workshop, Online, 8 Feb. 2022.

[79] ARSIWA (note 7), Art. 4.

[80] See e.g. Gillespie, T., 'Good practice for the development of autonomous weapons: Ensuring the art of the acceptable, not the art of the possible', *RUSI Journal*, vol. 65, no. 5–6 (2021); Seixas-Nunes (note 15); and View expressed by legal advisers and scholars, Experts workshop, Online, 9 Feb. 2022.

[81] See e.g. Henderson, Keane and Liddy (note 38), p. 21; and View expressed by state representatives, Interviews with the authors, Online, Feb.–Apr. 2022.

Box 2.2. Persons or entities whose conduct is attributable to the state

There are a range of persons or entities whose conduct is attributable to the state, including:

- Persons or entities who act as organs of the state
- Persons or entities exercising elements of governmental authority
- Organs placed at the disposal of a state by another state
- Persons or entities empowered to exercise elements of the governmental authority (even if they exceed their authority or contravene instructions)
- Persons or groups acting under instructions, direction or control of a state
- Persons or groups acting in the absence or default of the official authorities
- An insurrectional movement which becomes the new government of a state
- Agents otherwise acknowledged and adopted by a state as its own

Source: International Law Commission, 'Responsibility of states for internationally wrongful acts', Draft articles, Text adopted by the Commission at its 53rd session, 23 Apr. to 1 June and 2 July to 10 Aug. 2001, subsequently adopted by the United Nations General Assembly through Resolution No. 56/83 of 12 Dec. 2001, Arts 4–11.

remain the same.[82] Regardless of the weapon, users of AWS will still need to follow schemes of authorization, as allowing an AWS to override the usual decision-making processes 'is contrary to current practice in modern armed forces'.[83] Nevertheless, the preprogrammed nature of AWS, where some decisions are made at an early stage and distributed across the development and use phases, prompts the need to look closer at the larger decision-making structures, extending beyond the armed forces.[84] That is, AWS potentially make the roles and conduct of developers and decision makers increasingly relevant.[85]

The second question relates to the fact that 'human conducts' trigger state responsibility; specifically, what type of human conduct is required? The type is not defined in the law of state responsibility, leaving open questions as to whether the interdependency between human conduct and machine behaviour in the use of an AWS would have implications for attributing a violation of IHL to the state.[86]

The third question is, who is ultimately responsible for a breach in the context of complex decision-making structures, including decisions made by autonomous systems? Any targeting process—even without AWS—is inherently complex, wherein many individuals make numerous decisions that are both interlinked and distributed across different points in time and space.[87] Views differ as to the exercise of IHL-mandated evaluative decisions in targeting. One view is that in multi-layered decision-making structures, IHL demands that a single person, such as the commander, be responsible for the decision to use an AWS, including all the associated legally mandated value judgements (see chapter 3). The alternative view holds that implementation of a state's IHL obligations in the use of AWS could be considered a group exercise, residing with multiple people and shared across the chain of command and control.[88]

These tensions around the extent to which obligations under IHL are implemented by one person acting as an agent of the state or by a network of state agents have clear implications for the attribution of state responsibility. With AWS, the ability of a single commander to comply with IHL significantly depends on a set of decisions made at

[82] View expressed by state representatives, Interviews with the authors, Online, Feb.–Apr. 2022.

[83] Henderson, Keane and Liddy (note 38), p. 21.

[84] View expressed by legal scholars and state representatives, Experts workshop, Online, 8 Feb. 2022.

[85] Boutin (note 10), pp. 13–14; and View expressed by legal scholars and state representatives, Experts workshop, Online, 8 Feb. 2022.

[86] Lewis, D.A., 'On "responsible A.I." in war: Exploring preconditions for respecting international law in armed conflict', eds S. Voeneky et al., *The Cambridge Handbook of Responsible Artificial Intelligence: Interdisciplinary Perspectives* (Cambridge University Press: Cambridge, 2022); and Boutin (note 10).

[87] CCW Convention, GGE, 'Australia's system of control and applications for autonomous weapon systems' (note 70), p. 3.

[88] Boulanin, Bruun and Goussac (note 33),pp. 16–17.

earlier stages by other agents. The characteristics of AWS exacerbate the fact that in any complex system, decisions made by one link in the chain will affect the decisions of others further along the chain.[89] If one link in the chain fails to adequately consider the nature of, or the effects of implementing, AWS technologies, or fails to help take all feasible precautions in using an AWS in attack, it will be difficult for either the *commander*, on one view, or the *overall chain*, on the other view, to function as legally required. This could lead to one or multiple breaches of IHL by different actors as well as situations of diffuse responsibility. In both cases, it may be difficult to identify where a non-fulfilment of duties, attributable to the state, lies.[90]

State responsibility is a collective form of responsibility that does not aim at the individualization of responsibility but at the state apparatus. What matters in that context is the larger setup of decision-making processes and division of responsibilities *within* and *beyond* the chain of command, and whether that setup allows implementation and compliance with IHL. In other words, states have a meta-responsibility to ensure a proper scheme of a chain of command and control, and potentially to create new roles and functions within that chain, for the lawful development and use of AWS.[91] While the GGE generally agrees on the importance of ensuring a responsible chain of command in the development and use of AWS, there is less clarity around what that would look like in the context of AWS—and specifically whose conduct in decisions involving AWS would trigger state responsibility. Clarification of a state's decision-making structures in the development and use of AWS are important for assessing whether the state has a proper scheme of roles and responsibilities in place to ensure its AWS are developed and used in compliance with IHL.

Private actors

The preprogrammed nature of an AWS supposes that its effects will not only be determined partly by multiple people along the military command-and-control chain (users at different levels, including weapon operators) but also partly by private actors, including the many software engineers, developers and programmers (collectively, 'developers') involved in the development phase. As the steps taken and decisions made in the development phase of an AWS may have a direct impact on its effects and behaviour when used, it becomes relevant to consider whether the conduct of both state and private developers could trigger state responsibility for an IHL violation. To discuss this issue, it proves useful to shed some light on the distinction between state responsibility for the conduct of state agents and state responsibility for the conduct of private actors.

In principle, breaches of IHL committed by private actors are not imputable to the state. The conduct of a private actor such as a software developer is not the conduct of a state agent unless the developer acts 'on the instructions of, or under the direction or control' of the state, or the state 'acknowledges and adopts the conduct as its own'.[92] It therefore becomes important to look at how private actors have been incorporated into a state system that develops and uses AWS. For example, a private actor involved in developing an AWS that is incapable of being directed against a specific military objective as part of a state project would have their conduct attributable to that state, and the state is responsible for the breach of IHL.[93]

[89] Ekelhof, M., 'Lifting the fog of targeting: "autonomous weapons" and human control through the lens of military targeting', *Naval War College Review*, vol. 71, no. 3 (2017), p. 83.
[90] Ekelhof (note 89), p. 83.
[91] View expressed by experts and state representatives, Experts workshop, online, 8 Feb. 2022.
[92] ARSIWA (note 7), Arts 4, 8 and 11.
[93] View expressed by legal expert, Experts workshop, Online, 8 Feb. 2022

Another scenario is where the state fails to conduct a sufficient legal review, or does not conduct one at all, of a new AWS developed wholly or partly by private actors. In this situation, state responsibility would be triggered by the state's failure to fulfil its legal review obligations, and not for the unlawful conduct of the developers.

A third scenario involves states' due diligence obligation to ensure respect for IHL, where a state could be held responsible for failing to take all necessary measures to prevent violations of IHL by private actors, including developers. Again, the state is not directly responsible for the breaches of IHL committed by developers, but for its own failure to fulfil its due diligence obligations.[94] As discussed above, these obligations are ill-defined, but arguably could include the introduction of domestic regulation and IHL compliance 'by design' in the development of an AWS.[95]

Other states

Another issue concerns whether the international responsibility of a state can be triggered by violations of IHL by *other* states. AWS are rarely developed and used in isolation, so states have to consider their obligations concerning other states and responsibility for the conduct of other states. For example, when and under which conditions would a state be held responsible for unlawful civilian harm or death stemming from an AWS that it transferred to another state?[96] This example brings the question into the scope of the Arms Trade Treaty (ATT), under which states parties are required to either prohibit all types of transfers if they *know* weapons are going to be used to commit war crimes, or to assess the risk of weapons exports being used to commit or facilitate a serious violation of IHL.[97]

One starting point is the debate about whether the obligation to ensure respect for IHL entails an external dimension involving other states (see section I in this chapter). According to a broad interpretation of the external dimension of the duty to ensure respect, Common Article 1 of the Geneva Conventions imposes both negative obligations—not to encourage, aid or assist other states to violate IHL—and positive obligations to prevent, investigate and suppress violations of IHL.[98] This means that a state providing AWS to another state, *knowing* that the latter will use it for violations of IHL, could bear responsibility for having violated its due diligence obligations to prevent violations of IHL (and possibly the ATT), regardless of the state's intentions.[99] As already highlighted, the content of due diligence obligations is open-ended. Therefore, whether a state has violated its due diligence obligations would be determined on a case-by-case basis taking into account factors such as the 'gravity of the breach, the means reasonably available to the State, and the degree of influence it exercises over those responsible for the breach'.[100]

The primary IHL obligation to ensure respect for IHL operates in parallel with Article 16 of the ARSIWA, which provides that states are responsible for *knowingly* aiding or assisting another state in the commission of any internationally wrongful act.[101] The words 'with knowledge of the circumstances of the internationally wrongful act' in Article 16 have been interpreted as requiring a fault element of 'intent' on the

[94] Gaeta, Viñuales and Zappalá (note 8), p. 252; and View expressed by legal expert, Interview with authors, Online, 25 Feb. 2022.

[95] Boutin (note 10), p. 31; and Views expressed by experts and state representatives, Experts workshop, Online, 8 Feb. 2022.

[96] Boutin (note 10), p. 30.

[97] Arms Trade Treaty, opened for signature 3 June 2013, entered into force 24 Dec. 2014, Arts 6(3) and 7(1)(i).

[98] ICRC, 'Commentary [to GC I (note 11)] of 2016, Article 1: Respect for the Convention' (note 13), paras 159 and 164.

[99] ICRC, 'Commentary [to the GC I (note 11)] of 2016, Article 1: Respect for the Convention' (note 13), para. 192.

[100] ICRC, 'Commentary [to the GC I (note 11)] of 2016, Article 1: Respect for the Convention' (note 13), para. 150. See also Seixas-Nunes (note 15), p. 456; and Longobardo (note 17), pp. 46–47.

[101] ARSIWA (note 7), Art. 16.

part of the assisting state; that is, for state responsibility to be triggered through giving aid or assistance, the state must have *intended* to facilitate a violation of international law by the other state.[102] Whether 'intent' is required is, however, debated; one view is that the better interpretation 'is really about ensuring that supplying the weapons contributed materially to the wrongful act'.[103]

To strengthen compliance with IHL in these contexts, states could usefully clarify how they understand their due diligence obligations flowing from the duty to ensure respect for IHL by other states and what precisely is entailed in Article 16. As a starting point for the discussion on ensuring that state responsibility is attributed in the event of a breach, states could focus on scenarios of transfers of weapons.

III. Summary

State responsibility for internationally wrongful acts is a key framework for securing respect for IHL. As such, the framework of state responsibility contains the potential to retain human responsibility in the development and use of AWS. This is especially due to its broad scope: state responsibility can be triggered by the breach of any IHL provisions applicable to both the development and use of AWS; importantly, state responsibility is triggered not only by acts but also by omissions (i.e. what the state *failed to do* to respect or ensure respect for IHL); and finally, the framework of state responsibility serves to account for the collective and systemic nature of IHL implementation, in which multiple agents of the state participate.

However, the ascription of state responsibility is subject to several limitations that, while pre-dating AWS, are further challenged by certain aspects of AWS. First, existing uncertainties embedded in the open-texted nature of primary obligations of IHL (fundamental as well as facilitative rules) may make it difficult to identify what acts or omissions in the development and use of AWS would amount to a breach giving rise to state responsibility. Second, it is unclear how IHL fundamental rules, such as the prohibition on indiscriminate attacks, translate into conceptual and technical standards of foreseeability of effects and behaviour in the use of AWS. Third, the determination of *whose* acts and omissions trigger state responsibility becomes increasingly obscure in the context of AWS, which especially highlights the need to clarify the roles and responsibilities of agents involved in the larger decision-making structures, including developers and private actors.

For state responsibility to fulfil its crucial preventative function against IHL violations—and to release its potential to help ensure human responsibility in the development and use of AWS—dedicated efforts are needed to clarify how the framework of state responsibility for IHL violations applies to AWS. Such effort would need to address and clarify *what* (and *whose*) acts and omissions would trigger state responsibility in the context of AWS, with a particular focus on what types of unintended and unforeseen effects would, or should, engage the responsibility of the state.

[102] International Law Commission, 'Draft articles responsibility of states for internationally wrongful acts, with commentaries', *Yearbook of the International Law Commission, 2001*, vol. 2, no. 2 (2001), Commentary on Art. 16, para. 5.

[103] Boivin, A., 'Complicity and beyond: International law and the transfer of small arms and light weapons', *International Review of the Red Cross*, vol. 87, no. 859, (2005), p. 471.

3. Individual criminal responsibility for war crimes that involve the use of AWS

The framework of individual criminal responsibility for violations of international law holds that individuals can be held criminally liable for the commission of certain crimes that are considered to be of international concern. Among the spectrum of international crimes that could be committed with AWS—genocide, crimes against humanity, war crimes and aggression—this chapter focuses on war crimes. Serious violations of IHL amount to war crimes, which are codified and defined in such international instruments as the 1949 Geneva Conventions, Additional Protocol I, and the Rome Statute of the International Criminal Court (Rome Statute) (box 3.1) as well as in national laws on war crimes and customary international law (box 3.2).[104]

Individual criminal responsibility for war crimes has been discussed within the GGE as one of the applicable responsibility frameworks in the development and use of AWS.[105] Compared to state responsibility for internationally wrongful acts, individual criminal responsibility for war crimes has gained more attention in the policy debate as well as in the literature. However, the debate has yet to sufficiently elaborate on how the legal elements that underpin individual criminal responsibility would be established in relation to war crimes involving the use of AWS.

This chapter outlines the conditions necessary to trigger individual criminal responsibility and explores how these legal elements would be established in the specific case of AWS. The chapter focuses on war crimes committed with the *use* of AWS in the conduct of hostilities, particularly the war crimes of unlawful attacks stemming from violations of the fundamental IHL rules on distinction and proportionality (see table 2.1). The chapter is consequently structured around three questions based on the legal elements needed to establish a war crime: a material element and a mental element, which in turn depend on the applicable mode of responsibility (box 3.1). Section I addresses *what conduct* would fulfil the material element of a war crime in the use of AWS. Section II discusses what *mental state* would fulfil the required mental element of war crimes. Section III covers the question of *whose* conduct in the development and use of AWS could engage individual criminal responsibility, looking specifically at command responsibility, as well as some modes of participation in war crimes that could capture the criminal responsibility of developers. Section IV summarizes the chapter's main findings. A further question, how the legal elements of a war crime are established and scrutinized, is addressed in chapter 4.

I. What conduct in the use of AWS would fulfil the material element of a war crime?

This section explores what conduct in the *use* of AWS would fulfil the material element of a war crime. It outlines four main issues that could make the assessment of the material element of a war crime involving AWS a difficult task: (*a*) the different conduct requirements in AP I and the Rome Statute; (*b*) some disputed IHL elements in war crimes provisions; (*c*) problems deriving from the application of the war crime of indiscriminate attacks to the AWS context; and (*d*) problems associated with the criminalization of omissions (i.e. failures to act) in the context of AWS.

[104] Rome Statute of the International Criminal Court, opened for signature 17 July 1998, entered into force 1 July 2002 (Rome Statute).

[105] See e.g. CCW Convention, GGE, 'Report of the 2019 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems', CCW/GGE/2019/3, 25 Sep. 2019; CCW Convention, GGE, 'Revised chair's paper', 20 Sep. 2021; and CCW Convention, GGES, 'Chairperson's summary', 19 Apr. 2021.

Box 3.1. Definition, codification and elements of war crimes of unlawful attacks under international instruments

Definitions and codification

There are different codifications and definitions of war crimes of unlawful attacks across the 1949 Geneva Conventions (GCs), Additional Protocol I (AP I), and the Rome Statute of the International Criminal Court (Rome Statute), as well as in customary international law.[a]

Grave breaches of the Geneva Conventions and Additional Protocol I

Grave breaches are a limited number of violations of the GCs and AP I that are considered particularly serious; the list of grave breaches in AP I includes violations of the principles of distinction and proportionality.[b] Article 85(5) of AP I explicitly states that **grave breaches** of the Geneva Conventions and AP I **are war crimes**. While all grave breaches are war crimes, not all war crimes constitute a grave breach. In fact, beyond grave breaches, a violation of the Geneva Conventions and the additional protocols could amount to a war crime if it is 'serious' and is criminalized under international law.[c]

War crimes in the Rome Statute

The category of war crimes extends to those captured in Article 8 of the Rome Statute, which contains an exhaustive list of the war crimes that fall under the jurisdiction of the International Criminal Court. In relation to violations of the rules of the conduct of hostilities, and unlawful attacks specifically, Article 8 not only includes violations of the principles of distinction and proportionality but also goes beyond the list of grave breaches set out in Article 85 of AP I.[d]

Elements needed to establish a war crime of unlawful attack

Elements needed to establish that a war crime of unlawful attack has occurred vary across the international instruments, but typically comprise four elements:

- *Attributable conduct*. The conduct must be attributable to one or more natural persons engaging in conduct relating to an armed conflict. Artificial agents, such as machines, cannot bear individual criminal responsibility under current international and national criminal law.
- *A material element*. The attributable conduct and its consequences must constitute a serious violation of international humanitarian law (IHL) that amounts to a war crime, such as making a civilian population the object of an attack.
- *A mental element*. The alleged perpetrator must have acted 'wilfully' (under AP I, Article 85), or 'intentionally' (under the Rome Statute, Art. 8(2)(b)(i) and (iv)).
- *Mode of responsibility*. The proscribed conduct must be carried out through one of the recognized modes of responsibility, which typically fall into two categories: *perpetration* and *participation* in the commission of a crime.[e] For example, under Article 25 of the Rome Statute, perpetration of a war crime entails committing the crime (whether as an individual or jointly with one or more other individuals), or ordering, soliciting, or inducing the commission of the crime; while participation in a war crime includes aiding, abetting or otherwise assisting in the crime's commission or attempted commission, including providing the means for its commission. A separate mode of responsibility is command responsibility, which is the doctrine holding that a leader is criminally responsible for the war crimes of their subordinates (see box 3.3).

[a] International Committee of the Red Cross (ICRC), Customary IHL Database, [n.d.], 'Rule 156. Definitions of war crimes'.

[b] The definition of a grave breach is found in the Geneva Conventions and additional protocols: GC I, Art. 50; GC II, Art. 51; GC III, Art. 130; GC IV, Art. 147; and AP I, Art. 11. Article 85 of AP I contains a definition of grave breaches stemming from the violation of the principles of distinction and proportionality. On the Geneva Convention and its additional protocols see note 11 on p. 6 of the main text.

[c] ICRC, 'What are "serious violations of international humanitarian law"?', Explanatory note, 2012; and Gaeta, P., 'The interplay between the Geneva Conventions and international criminal law', eds A. Clapham, P. Gaeta and M. Sassoli, *The 1949 Geneva Conventions: A Commentary* (Oxford University Press: Oxford, 2015), p. 742.

[d] The *Elements of Crimes* define and specify the elements of international crimes under the Rome Statute and complement their interpretation. International Criminal Court (ICC), *Elements of Crimes* (Official Records of the Assembly of States Parties to the Rome Statute of the International Criminal Court, First session, New York, 3–10 Sep. 2002, part II.B).

[e] de Hemptinne, J. et al. (eds), *Modes of Liability in International Criminal Law* (Cambridge University Press: Cambridge, 2019).

Diverging material elements in Additional Protocol I and the Rome Statute

Under AP I and the Rome Statute, violations of certain general prohibitions and restrictions on the conduct of hostilities give rise to war crimes (box 3.1). While the range of war crimes that could arise in the use of AWS is broad, this chapter focuses primarily on war crimes of unlawful attacks against civilians and civilian objects as paramount

examples of war crimes stemming from violations of the principles of distinction and proportionality (table 2.1).[106]

The material elements of the war crimes of unlawful attacks set out in AP I and the Rome Statute differ. In AP I, an unlawful attack that stems from the violations of the principles of distinction or proportionality must result in civilian death or injury to be defined as a war crime.[107] Under the Rome Statute a result of death or injury is not a requirement of the material element for a war crime. Article 8(2)(b)(i) of the Rome Statute criminalizes 'Intentionally directing attacks against the civilian population as such or against individual civilians not taking direct part in hostilities'. Moreover, Article 8(2)(b)(ii) criminalizes 'Intentionally directing attacks against civilian objects, that is, objects which are not military objectives' and is equally formulated as a crime of conduct.[108] This divergence has implications for prosecuting war crimes that involve the use of an AWS. With any weapon there are challenges in establishing a causal link between an unlawful attack and its consequences, but these are exacerbated in the case of AWS. The nature of AWS—especially the fact that how an AWS arrives at a particular decision is likely to be obscure, known as the 'black box' problem (see section II in chapter 4)—makes determining the causality nexus between an attack with an AWS and the resulting deaths or injuries potentially difficult.[109]

Other differences in the material elements of war crimes concern the formulation of war crimes as stemming from violations of the proportionality rule. Article 85(3)(b) of AP I criminalizes 'Launching an indiscriminate attack affecting the civilian population or civilian objects, in the knowledge that such attack will cause excessive loss of civilian life, injury to civilians or damage to civilian objects', where it defines 'excessive' 'in relation to the *concrete* and *direct* military advantage anticipated from the attack'.[110] The Rome Statute defines the violation in the same terms but with a higher threshold for what is considered disproportionate as '*clearly* excessive in the relation to the *concrete* and *direct overall* military advantage anticipated'.[111] Some experts consider the Rome Statute definition as more stringent, in that it requires greater consideration of 'the context in which the attack takes place when considering the concrete and direct military advantage anticipated from the destruction, capture or neutralization of the attacked military objective'.[112]

Finally, it is important to note that the different implementations of war crimes provisions in national laws, especially which definitions a state chooses to incorporate, influences states' approaches to the enforcement of criminal responsibility for war crimes. These different approaches have implications for how states allocate human criminal responsibility for war crimes involving use of an AWS and potentially dictate how states approach the regulation of AWS.

[106] For examples see AP I (note 11), Art. 85(1)(d), (2); and Rome Statute (note 104), Art. 8(2)(b)(iii), (v), (vi), (xxiv). For violations of specific and general rules prohibiting or restricting specific weapons, means and methods of warfare criminalized under the Rome Statute, see Art. 8(2)(b)(xvii), (xvii), (xix) (xix) . See also Boulanin, Bruun and Goussac (note 33), pp. 4–5.

[107] AP I (note 11), Art. 85(3).

[108] Rome Statute (note 104), Arts 8(2)(b)(i) and (ii).

[109] Egeland, K., 'Lethal autonomous weapon systems under International Humanitarian Law', *Nordic Journal of International Law*, vol. 85, no. 2 (2016).

[110] AP I (note 11), Art. 57(2)(a)(iii) (emphasis added).

[111] Rome Statute (note 104), Art. 8(2)(b)(iv) (emphasis added).

[112] Olásolo, H., *Unlawful Attacks in Combat Situations: From the ICTY's Case Law to the Rome Statute* (Martinus Nijhoff: Leiden, 2008), p. 83.

Disputed interpretation of IHL-rooted elements in war crimes provisions

Some crucial elements of war crimes have their roots in IHL fundamental rules that are open to interpretation. This unresolved issue of war crimes law compounds the determination of material elements of war crimes involving AWS.[113]

For instance, as discussed above, there are different interpretations of '(clearly) excessive incidental damage' and 'concrete and direct (overall) military advantage anticipated'. The IHL benchmark for determining when a civilian loses their protected status, and thus when they could become a legitimate military objective, is the concept of direct participation in hostilities. However, there are divergent interpretations as to the range of activities included within the notion of active participation in hostilities, as well as 'whether the activities carried out by certain categories of non-combatants amount to active participation in the hostilities'.[114]

Another debated concept that is crucial for attributing criminal responsibility for war crimes involving AWS is the concept of 'attacks that are not directed against a specific military objective', discussed next.

The difficulty of establishing the war crime of indiscriminate attacks

The war crime of indiscriminate attacks deserves particular attention in the context of AWS. This is due, among other reasons, to the potentially increased risk of errors in target identification associated with the use of AWS.[115] However, several issues make it difficult to attribute responsibility for the war crime when AWS are involved.

First, the category of indiscriminate attacks as a war crime is rather unclear. Violations of the principles of distinction and proportionality in attack are both described as indiscriminate attacks, but neither AP I nor the Rome Statute defines a specific war crime of indiscriminate attack. Scholars and the case law of international criminal courts define an attack as 'indiscriminate' if it is not directed against a specific military objective; if it uses an indiscriminate weapon (i.e. one incapable of distinguishing between civilian objects and military objectives); and if it is carried out without taking the necessary precautions to spare civilians, especially by failing to seek precise information on the target of the attack.[116] The lack of a specific war crime provision regarding indiscriminate attacks as a violation of the principle of distinction creates uncertainties as to the elements of this war crime, with serious implications for the attribution of individual criminal responsibility for war crimes involving AWS.

Second, how is the requirement for an attack to be 'directed at a specific military objective' to be construed in the context of AWS, when the concept of 'specific military objective' is subject to different interpretations? For example, under Article 51(5)(a) of AP I, distinct military objectives cannot be treated as a single objective. Does this entail that an attack using an AWS that is programmed to attack multiple and distinct military targets without human authorization or supervision is indiscriminate? Or, instead, would this attack be in principle lawful if the different targets and locations to which the AWS is aimed are part of a larger, identifiable and coherent military objective? Further questions arise in relation to how the precision of target profiles

[113] Gaeta, P., 'Serious violations of the law on the conduct of hostilities: A neglected class of war crimes?', eds F. Pocar, M. Pedrazzi and M. Frulli, *War Crimes and the Conduct of Hostilities: Challenges to Adjudication and Investigation* (Edward Elgar: Cheltenham, 2013), pp. 26–28; and MacDougall, C., 'Autonomous weapon systems and accountability: Putting the cart before the horse', *Melbourne Journal of International Law*, vol. 20, no. 1 (2019).

[114] Olásolo (note 112), pp. 107–15; see also Boulanin, Bruun and Goussac (note 33),p. 21.

[115] Views expressed by legal experts, Experts workshop, Online, 9 Feb. 2022; Boulanin, Bruun and Goussac (note 33), p. 22; and Ohlin, J. D., 'The combatant's stance: Autonomous weapons on the battlefield', *International Law Studies*, vol. 92 (2016).

[116] Dörmann, K., *Elements of War Crimes under the Rome Statute of the International Criminal Court: Sources and Commentary* (Cambridge University Press: Cambridge, 2003), pp. 131–32; and *Situation in the Democratic Republic of the Congo in the Case of The Prosecutor v. Bosco Ntaganda* (ICC, Trial Chamber VI, 8 July 2019), para. 921.

and whether some types of military objectives—those defined as military by their location, purpose or use, rather than being military by nature—preclude the use of AWS in attack.[117]

Third, when does the use of an AWS amount to an attack using a weapon that is indiscriminate by nature? This challenge relates to the difficulties in determining tolerable standards of precision, reliability and foreseeability.[118]

Problems associated with the criminalization of omissions in the context of AWS

Unlike command responsibility and other modes of participation where omissions, or failures to act, are an uncontroversial material element, the mode of individual perpetration of a war crime is less clear on whether a failure to act gives rise to individual criminal responsibility. In the context of AWS, the 'commission by omission' problem specifically relates to the question of whether a failure to suspend an unlawful attack with AWS equates to actively launching an unlawful attack. For example, does an individual's failure to override the autonomous targeting functions of an AWS, to abort the system's mission, or to suspend an attack using an AWS that was or should have been expected to be unlawful, constitute a war crime?[119] Clarifying to what extent and under which conditions such failures trigger criminal responsibility is also relevant for identifying types and degrees of human–machine interaction required in the use of AWS.

The question of whether omissions, such as a failure to suspend an attack with AWS expected to be unlawful, is a mode of commission of war crime pre-dates the AWS conversation and remains unsettled. Some national laws provide a legal basis for the criminalization of war crimes committed by omissions; Article 86 of AP I requires states to 'repress grave breaches . . . which result from a failure to act when under a duty to do so'. However, whether there is a rule of customary international law on 'commission by omission' is subject to debate: some argue that such a rule 'is available to states that wish to apply or implement it, but they are free to do otherwise'.[120] It is similarly contentious as to whether omissions are criminalized within the Rome Statute: a general provision on omission liability was ultimately rejected in the adoption of the Rome Statute because a consensus could not be reached by delegates on the requirements of omissions.[121]

Some omissions, such as failing to gather or use available information to verify targets, could amount to violations of the IHL primary rule on the duty to take precautions (table 2.1).[122] However, war crime provisions do not criminalize failures to take precautions. Failures to take precautions could, in principle, be taken into account as contextual elements to prove that a commander had the intent to target civilians.[123] Even so, there are still questions as to what, in concrete terms, is demanded by the principle of precaution on the part of the commander, especially what information the precautionary rule on target verification requires in the use of AWS. A specified actual target? A particular type of target? Objects and features that may be expected to trigger the AWS during the attack? In this context, it would be useful to clarify the distinction between omissions that might give rise to criminal responsibility for

[117] Boulanin, Bruun and Goussac (note 33), p. 22.

[118] See chapter 2 under the heading 'The standards for assessing whether an AWS by nature or by use is indiscriminate'.

[119] See Bo, M., 'Failures to exercise human control over autonomous weapons systems-related attacks and criminal responsibility for war crimes', *Journal of Criminal Justice* (2023) (forthcoming).

[120] Gaeta, P., 'Grave breaches of the Geneva Conventions', eds A. Clapham, P. Gaeta and M. Sassoli, *The 1949 Geneva Conventions: A Commentary* (Oxford University Press: Oxford, 2015), p. 627.

[121] Ambos, K., 'Omissions, in particular command responsibility', *Treatise on International Criminal Law: Volume 1: Foundations and General Part* (Oxford University Press: Oxford, 2013), p. 189.

[122] AP I (note 5), Art. 57(2)(a)(i) and (b)

[123] Dörmann (note 116), p. 132.

'commission by omission' and failures to take precautions that do not trigger criminal responsibility.

II. What standards of intent and knowledge in the use of AWS would fulfil the mental element of a war crime?

The second element for the attribution of individual criminal responsibility is the mental element, or 'guilty mind' (*mens rea*). Like the material element, a fundamental difficulty for the establishment of the mental element of a war crime, involving AWS or not, is that the element is codified differently in AP I and the Rome Statute. Under Article 85(3) of AP I, war crimes stemming from violations of the rule of distinction and proportionality must be committed 'wilfully'. In contrast, paragraphs (i) and (iv) of Article 8(2)(b) of the Rome Statute require that the prohibited attack is executed 'intentionally'; there is some debate as to whether intentionality under Article 8 coincides with the general *mens rea* requirement of 'intent and knowledge' under Article 30 of the Rome Statute.[124] In addition, definitions and interpretations of the mental element adopted in national criminal or military laws either coincide or depart from international law standards (box 3.2).

In the context of AWS, these differences in codification and interpretation do not matter so much in situations where direct intent (*dolus directus*) can be established. An example is where the user deliberately programs and launches an AWS to attack a civilian population. If the user's intent is established then it is undisputedly covered by the mental elements of 'wilfulness' under AP I and 'intentionality' under the Rome Statute.[125] Moreover, situations of indirect intent (*dolus indirectus*), where the user launches an attack using an AWS and is 'practically' or 'virtually' certain that the attack will be directed against civilians or result in civilian death or injuries, are covered by these mental elements.[126] However, it could be problematic where the use of an AWS enables the user to shield their 'intent' or knowledge.

Difficulties emerge in cases where harm, or the risk of harm, to protected people or objects is caused by an AWS user's insufficient care, control or diligence. In such cases, the mental element that needs to be established is recklessness, *dolus eventualis* (a special kind of intent involving foreseeing and accepting the consequence and risks of actions) or negligence. The extent to which conduct (act or omission), and risk-taking behaviour in particular, establishes these mental elements and whether that triggers individual criminal responsibility depends on the implementation and interpretations adopted in the legal framework (box 3.2).

This section outlines some key issues raised by the determination of the mental element in three different scenarios of unintended harm to protected persons and objects resulting from the use of AWS: (*a*) risk-taking behaviours, including recklessness and *dolus eventualis*, where the harm was foreseen; (*b*) negligent behaviours, where the harm was foreseeable; and (*c*) accidents, where the harm was unforeseen.

[124] See Werle, G. and Jessberger, F., '"Unless otherwise provided": Article 30 of the ICC statute and the mental element of crimes under international criminal law', *Journal of International Criminal Justice*, vol. 3, no. 1 (2005); and Bo, M., 'Autonomous weapons and the responsibility gap in light of the *mens rea* of the war crime of attacking civilians in the ICC statute', *Journal of International Criminal Justice*, vol. 2, no. 19 (2021).

[125] ICRC, 'Commentary [to AP I (note 11), Art. 85] of 1987: Repression of breaches of this protocol', para. 3474; International Criminal Court (ICC), *Elements of Crimes* (Official Records of the Assembly of States Parties to the Rome Statute of the International Criminal Court, First session, New York, 3–10 Sep. 2002, pp. 9–30 (Art. 8); and *Situation in the Democratic Republic of Congo in the case of the Prosecutor v. Germain Katanga and Mathieu Ngudjolo Chui* (ICC), Pre-Trial Chamber I, 30 Sep. 2008, para. 271.

[126] Olásolo (note 112), p. 218; and ICRC (note 125), para. 3474.

Box 3.2. Establishing the mental element of a war crime

Mental elements of a crime under national laws

There are various types and degrees of mental elements and different understandings of these notions across and even within national criminal law systems. Generally speaking, the mental elements are:

- under civil law systems (e.g. most European states): intent in the first degree, or direct intent (*dolus directus*); intent in the second degree or indirect intent (*dolus indirectus*); *dolus eventualis*; and negligence (*culpa*)
- under common law systems (e.g. Australia, the United Kingdom and the United States): direct intent (intentionally or purposely); indirect intent (knowingly or intentionally); recklessness; and negligence.

In addition, some countries recognize some form of strict or absolute liability offences.

Direct intent

Direct intent refers to the perpetrator's aim and is characterized by their 'purposeful will' to engage in the prohibited conduct or bring about the forbidden results.[a]

Indirect intent

In civil law systems, the perpetrator must foresee that it is certain or highly probable that specifically forbidden consequences will flow from their conduct. In common law systems, this mental state is known as acting 'knowingly' (USA), or still broadly defined as 'intentionally' (Australia and the UK). According to the US Model Penal Code, 'knowingly' means the perpetrator is 'practically certain' that their conduct will cause the forbidden result; in the UK the perpetrator only needs to be 'virtually certain' that the forbidden result will occur.[b]

Dolus eventualis and recklessness

There is no uniform definition of *dolus eventualis*, but generally this mental element means the perpetrator foresees the risk that a forbidden consequence is likely to occur and nevertheless proceeds with their actions. That is, *dolus eventualis* is defined by the perpetrator's particular subjective posture towards the result: they must accept or approve the forbidden consequence, or be reconciled or make peace with its occurrence, or be indifferent to its occurrence.[c]

Recklessness means the perpetrator foresees that their conduct may bring about the forbidden consequence but nevertheless takes a deliberate and unjustifiable risk of bringing it about. There are long-standing debates over the difference between the concept of *dolus eventualis* in civil law systems and recklessness in common law systems. According to some, the difference lies in that recklessness 'requires an affirmative aversion to the harmful side-effect'.[d]

Negligence or culpa

Negligence or *culpa* refers to a lack of foresight.[e] US military law refers to both culpable and simple negligence as being degrees of carelessness, and defines simple negligence as 'the absence of due care, that is, an act or omission of a person who is under a duty to use due care which exhibits a lack of that degree of care of the safety of others which a reasonably careful person would have exercised under the same or similar circumstances'.[f] Others argue that negligence is the fault of not knowing when there is a duty to know.[g]

The ground rule is that criminal offences must be committed intentionally and whether a criminal offence is punishable on the basis of negligence must be explicitly provided for by law.

The mental element in international law

There are different mental elements for war crimes across the 1949 Geneva Conventions, Additional Protocol I (AP I) and the Rome Statute of the International Criminal Court (Rome Statute).

Wilfully

Under Article 85(3) of AP I, behaviour that constitutes a war crime of unlawful attack must have been *committed wilfully*. According to the International Committee of the Red Cross, this mental element means that the perpetrator 'must have acted consciously and with intent', meaning that their mind was 'on the act and its consequences' (also known as 'criminal intent' or 'malice aforethought'), and that 'this encompasses the concepts of "wrongful intent" or "recklessness", viz., the attitude of an agent who, without being certain of a particular result, accepts the possibility of it happening'; however, the element does not cover 'ordinary negligence or lack of foresight', when an agent acts without having their mind 'on the act or its consequences'.[h] For the war crime of directing an attack against civilians, international criminal tribunals have similarly held that 'the perpetrator has to act consciously and with intent, willing the act and its consequences. This encompasses the concept of recklessness but not negligence'.[i]

Intent and knowledge

Under Article 30 of the Rome Statute, the general mental element is 'intent and knowledge', where *intent* is defined in terms of a person's conduct as 'that person means to engage in the conduct' and of a consequence as 'that person means to cause that consequence or is aware that it will occur in the ordinary course of events'; and *knowledge* 'means awareness that a circumstance exists or a consequence will occur in the ordinary course of events'. However, under Article 8(2)(b), war crimes of unlawful attacks require the specific mental element of being committed 'intentionally'. Whether the mental element for war crimes covers only direct and indirect intent, or possibly also covers *dolus eventualis*, is subject to debate.[j]

[a] Finnin, S., 'Mental elements under Article 30 of the Rome Statute of the International Criminal Court: A comparative analysis', *International and Comparative Law Quarterly*, vol. 61, no. 2, pp. 330–31.

[b] Finnin (note a), p. 332.
[c] Fletcher, G. P., *Rethinking Criminal Law* (Oxford University Press: Oxford, 2000), pp. 445 and 446.
[d] Fletcher (note c), pp. 445 and 446.
[e] International Committee of the Red Cross (ICRC), 'Commentary [to the Protocol Additional to the Geneva Conventions of 12 Aug. 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I), 8 June 1977, Art. 85] of 1987: Repression of breaches of this protocol', para 3474.
[f] US Joint Service Committee on Military Justice (JSC), *Manual For Courts-Martial, United States (2019 edition)* (JSC: Washington, DC, 2019), p. IV-147, para. 103(c)(2).
[g] Fletcher (note c), p. 182.
[h] ICRC (note e), para. 3474.
[i] See e.g. *Prosecutor v. Radavan Karadžić* (ICTY, Trial Judgment (public redacted version), 24 Mar. 2016), vol. 1, para. 456; *Prosecutor v. Pavle Strugar* (ICTY, Appeals Chamber, 17 July 2008), para. 270.
[j] Ambos, K., *Treatise on International Criminal Law* (Oxford University Press: Oxford, 2013), pp. 277 and 278; and Bo, M., 'Autonomous weapons and the responsibility gap in light of the *mens rea* of the war crime of attacking civilians in the ICC statute', *Journal of International Criminal Justice*, vol. 2, no. 19 (2021).

Risk-taking behaviours, including recklessness, in the use of AWS

Whether certain risk-taking behaviour engages individual criminal responsibility for a war crime has important implications for the use of AWS. The defining conceptual features of AWS as preprogrammed weapons—where the parameters for target identification, selection and engagement are determined in advance—have raised the question as to what their users *need to know* to exercise IHL-mandated evaluations in attack, such as those demanded by the principles of distinction, proportionality and precautions, to ensure that the attack is lawful. That is, at what point does the conduct of a user of an AWS who knows there is a risk of harm to civilians, and nevertheless *engages in risk-taking behaviour* in an attack, amount to a war crime? The answer depends on (*a*) the interpretation of the mental element of war crimes and whether the user's knowledge of risks and risk-taking conduct is sufficient to trigger criminal responsibility for war crimes; and (*b*) the standards and type of knowledge demanded by IHL on the part of those deciding on and launching attacks.

First, whether and which form of risk-taking behaviours entail criminal responsibility for war crimes depends on the legal framework. For example, a user who foresaw the risk of an attack using an AWS being indiscriminate, but was reckless as to this result and carried out the attack, might be liable for the war crime under AP I but not under the Rome Statute. The International Committee of the Red Cross (ICRC) accepts in its commentary on Article 85 of AP I that the mental element of 'willingly' encompasses recklessness; while international criminal tribunals have agreed with this interpretation (see box 3.2), it remains debated.[127] The Rome Statute does not mention recklessness in Article 30, only 'intent and knowledge', and many scholars have concluded that Article 30 generally excludes recklessness.[128] Whether 'intent and knowledge' overlaps with 'intentionally' under Article 8 and whether the Rome Statute accepts the related notion of *dolus eventualis* is also debated (see box 3.2). If so, it would hold criminally responsible any user deploying an AWS who accepted the high probability that the ensuing attack could violate the rules on distinction and proportionality.[129] However, states could consider *dolus eventualis* as a sufficient mental element under national laws.

[127] Massingham, E. and McKenzie, S., 'Testing knowledge: Weapons reviews of autonomous weapons systems and the international criminal trial', eds E. Palmer et al., *Futures of International Criminal Justice* (Routledge: Abingdon, 2021); Ohlin, J., 'Targeting and the concept of intent', *Michigan Journal of International Law*, vol. 13, no. 1 (2013); and Bo, M., 'The human–weapon relationship in the age of autonomous weapons and the attribution of criminal responsibility for war crimes', Conference paper presented at We Robot 2019, University of Miami Law School, Apr. 2020.
[128] Massingham and McKenzie (note 127); and Ohlin (note 115), p. 22.
[129] Bo (note 124); and Ohlin (note 115),p. 22.

Second, whether the mental elements of intent, knowledge or recklessness can be established depends partly on the standard and type of knowledge on the part of the user of an AWS, as demanded by IHL. While acknowledging the different scope of decisions taken by different individuals involved in the use, including commanders and weapon operators, the question is, what should the user(s) know before deploying an AWS in an attack? Should the user know about what advance programming was done, such as inputting target profiles, or the internal functioning of an AWS? Some argue that the complexity of AWS requires heightened technical knowledge of how a weapon functions and how a result is achieved, while others hold that what is required is knowledge about the effects of the use of an AWS concerning a concrete attack.[130]

One aspect that should be considered in this debate is that the implementation of IHL is, to a large extent, a systemic and collective activity. Therefore, it is unrealistic to expect that every single actor knows everything; rather, actors involved in IHL implementation necessarily have to rely on decisions made previously by other actors. As with other weapons, users of AWS are entitled to assume that others before them have acted in good faith, for example by developing an AWS in line with extant departmental testing, evaluation criteria and the information and guidance in weapons manuals. Experts have argued that Article 36 legal reviews can assist in establishing the level of knowledge required. Because these legal reviews are likely to provide recommendations on the lawful and unlawful use of an AWS, they can translate into weapons manuals that provide knowledge to users of the AWS. Use of an AWS that departs from the circumstances authorized in the legal review as weapons manual can be proven as conduct showing the user's intent or knowledge as the mental element of criminal responsibility.[131]

A related question is what the user should know about the environment of use, including the level of knowledge and prediction about how an AWS will react to the environment. Experts consulted as part of this project suggested that the user's knowledge must be evaluated with respect to the normal and expected use in the context of its anticipated operating environment. However, others question whether it is reasonable to expect the user to know 'all situational variables and all the different ways they can be processed'.[132]

It follows that the standard of knowledge required of an AWS user—and therefore which risk-taking behaviours could amount to war crimes—varies from system to system and is contingent on the environment of use. Clarifying these standards of knowledge for different scenarios will be crucial for establishing individual criminal responsibility in the use of AWS that results in a war crime. An example scenario is where the design of an AWS precludes foreseeability of effects and risks, such that its users are prevented from acquiring the necessary type of knowledge to implement IHL, which would underline the question as to whether fielding such an AWS would be lawful in the first place. AWS provide, in that regard, a vehicle to explore anew, and potentially help resolve, enduring legal debates about whether recklessness is an accepted mental element for war crimes, as well as the standards of intent, knowledge, behaviour and care that are demanded by the fundamental IHL targeting rules.

[130] For the former view see e.g. Seixas-Nunes (note 15), p. 433. The latter view is that of experts consulted as part of this project.

[131] It has been suggested that the ability to establish *mens rea* can be improved through legal reviews. See Massingham and McKenzie (note 127); and Dunlap (note 54).

[132] Buchan, R. and Tsagourias, N., 'Autonomous cyber weapons and command responsibility', *International Law Studies*, vol. 96, no. 64 (2020), p. 661.

Negligent behaviours in the use of AWS

Whether negligence or lack of due care satisfies the mental element for war crimes is a long-standing debate that becomes of crucial relevance in the context of AWS.[133] The IHL obligation to take precautions in attack is arguably formulated as a duty of care and demands 'the greatest care a combatant can reasonably take under circumstances where non-combatant harm is not intended but is a real risk'.[134] Negligence is not a mental element for war crimes under AP I or the Rome Statute (box 3.2). However, the implications of negligent behaviour for individual (criminal) responsibility vary across national systems. Some states criminalize some war crimes committed negligently.[135] In others, negligent behaviour in the deployment of AWS (e.g. failing to take precautions in attack by omitting to seek precise information on the targets) could amount to an administrative offence, or be punishable by disciplinary measures.[136] And in some states, such as the USA, ordinary laws that cover death or harm arising from negligence could cover some negligent behaviours during conflict.[137] (A complicating factor is the combatant privilege immunity under Article 43(2) of AP I, which states that 'lawful combatants must then be given immunity from domestic prosecution for violent acts committed strictly in accordance with IHL, even though they would normally be crimes under the domestic law of the competent State(s)'.[138])

It remains unclear what acts or omissions would amount to being negligent in the case of use of AWS. Possible examples of negligent behaviour are disregarding weapons manuals and failures to take precautions in attack. However, in the case of AWS, proving negligent behaviour may be difficult. The interplay of automation bias and complacency (respectively, the tendency to trust automated systems and insufficient monitoring of automated output), complex human–machine interfaces and opaque design of autonomous functions, may prevent users from becoming aware of some risks associated with AWS or affect how the AWS might respond to edge cases (i.e. cases outside of 'normal or expected use').[139] These issues can be an obstacle to holding individuals responsible on the basis of negligence.[140]

Further elaborations on what conduct in the use of AWS amounts to negligent behaviour—and therefore entails criminal punishment as opposed to disciplinary, administrative or other measures—are warranted.

Unforeseen accidents involving the use of AWS

The third set of questions arises from incidents involving an AWS that inflict harm on civilians or protected persons or objects, where the harm is the result of unforeseen behaviour or effects of the AWS. As mentioned in section 1 of chapter 2 ('Distinguishing accidents from errors that constitute breaches of IHL fundamental obligations') 'normal accident' theory suggests that in tightly coupled complex systems—such as AWS—accidents are 'inevitable' over a long time span. Moreover, because complex systems often interact in ways that are unanticipated and nonlinear, the rate of such

[133] Gaeta, Viñuales and Zappalá (note 8); Crootof (note 10); and Seixas-Nunes (note 15), p. 453.

[134] Rudesill, D. S., 'Precision war and responsibility: Transformational military technology and the duty of care under the laws of war', *Yale Journal of International Law*, vol. 32, no. 1 (2007), p. 528; and Sassòli and Quintin (note 23).

[135] Crootof, R., 'War torts: Accountability for autonomous weapons', *University of Pennsylvania Law Review*, vol. 164, no. 6 (2016), p. 1383.

[136] ICRC (note 125), para. 3474.

[137] See e.g. 10 U.S.C. §§ 47.X.918–19 (2012).

[138] Yanev, L., 'Jurisdiction and combatant's privilege in the *MH17* trial: Treading the line between domestic and international criminal justice', *Netherlands International Law Review*, vol. 68 (2021).

[139] Parasuraman, R., Molloy, R. and Singh, I. L., 'Performance consequences of automation-induced "complacency"', *International Journal of Aviation Psychology*, vol. 3, no. 1 (1993).

[140] Bo (note 124), pp. 296 and 297.

accidents often cannot be predicted accurately in advance. System designers may not be able to accurately assess the probability of a failure, and hidden failure modes may lurk undetected.[141] For these reasons, concerns have been raised about whether AWS are more prone to 'normal' accidents than other complex weapons systems.[142] With AWS there is also an increased risk of technical components acting unpredictably, adversarial interference, user errors, data errors, or communications failures.[143] Many argue that the unpredictability brought about by AWS is qualitatively different from the 'ordinary' unpredictability of traditional weapons.[144] Others argue that because the operational environment in which AWS are meant to be deployed is 'deeply unstructured and unpredictable', AWS are 'bound to act unpredictably, no matter how well they are designed'.[145] Some machine learning algorithms, such as those that would allow the AWS to keep 'learning on the job' (online learning), could also make it harder to understand the decisions made by the AWS and so give rise to more unpredictability.[146] For this reason, some states have declared they would never deploy AWS driven by such algorithms.[147]

Under the legal framework of individual criminal responsibility (box 3.2), an individual cannot be held responsible for the unpredictable behaviour and effects of an AWS because the ability to foresee is a necessary requirement of the mental elements of intent and knowledge.[148] Thus, the crucial questions are: (*a*) what types of accidents involving an AWS would qualify as unforeseeable, and therefore not criminally attributable; and (*b*) what types of accidents should have been foreseeable through, for example, testing and training, and therefore could amount to negligent and risk-taking behaviours that in some instances could give rise to individual criminal responsibility? The answers to these questions are complicated by the fact that the unpredictability and complex nature of AWS mean that an accident could have several causes, which are often interrelated.[149] Moreover, the extent of unpredictability in the use of AWS is debated.[150]

It is therefore of the utmost importance that states better categorize possible accidents with AWS and seek a deeper understanding of what types of accidents could, and should, have been foreseen. This would help delineate the contours of negligent and risk-taking behaviours that could potentially trigger individual criminal responsibility.

III. *Whose* conduct in the development and use of AWS could trigger criminal responsibility for war crimes?

Decisions to use force involving AWS are likely to be spread across more people over a larger geographical and temporal scope than those involving traditional weapons. This diffused decision-making makes it relevant to consider modes of responsibility

[141] Scharre (note 27), p. 25.

[142] Crootof (note 135), p. 1373; and Holland Michel (note 49).

[143] Crootof (note 10).

[144] Acquaviva, G., 'Autonomous weapons systems controlled by artificial intelligence: A conceptual roadmap for international criminal responsibility', *Military Law and the Law of War Review*, vol. 60, no. 1 (2022), p. 110.

[145] Amoroso, D. and Giordano, B., 'Who is to blame for autonomous weapons systems' misdoings?', eds Carpanelli and Lazzerini (note 66), p. 215.

[146] Online learning is a training regime where the system, once launched, keeps using its input from its operating environment to optimize and refine its behavior. It is in constrast to offline learning where the system only learns during a definite training phase prior to deployment. See Hughes, J., 'The law of armed conflict issues created by programming automatic target recognition systems using deep learning methods', *Yearbook of International Humanitarian Law* (Asser: The Hague, 2019), p. 99; and Boulanin, V. and Verbruggen, M., *Mapping the Development of Autonomy in Weapon Systems* (SIPRI: Stockholm, Nov. 2017), p. 16.

[147] View expressed by state representatives, Experts workshop, Online, 10 Feb. 2022.

[148] Bo (note 124); Crootof (note 135), p. 1373; and View commonly expressed by legal experts, Experts workshop, Online, 9 Feb. 2022.

[149] Crootof (note 10), p. 13.

[150] Holland Michel (note 37).

pertaining to *command responsibility* and to *participation in a war crime* (such as aiding and abetting), rather than direct individual commission of the crime.[151] In the AWS context, the focus for each mode, in the GGE and the literature, is on the roles of, respectively, military commanders and developers.[152] The doctrine of command responsibility has a long history in IHL and international criminal law (box 3.3), and so may be considered an obvious avenue for ascribing responsibility for war crimes involving AWS. In contrast, the question of whether developers of an AWS may be held responsible for facilitating a war crime is novel. This section discusses the two possibilities in turn.

The role of military commanders under the doctrine of command responsibility

A military commander can be held responsible for the commission of a war crime under the modes of perpetration or participation (box 3.1) or command responsibility (box 3.3).[153] Under the doctrine of command responsibility, commanders are responsible for failing to prevent or punish war crimes committed by their subordinates, rather than violating the principles of distinction and proportionality through their own actions. For these reasons, command responsibility is often presented as a solution to the purported accountability gap for war crimes involving AWS.[154] As such it allows commanders to be held accountable for the breadth of macro- and micro-level decisions they take about AWS use—including, potentially, how target profiles or indicators are programmed, the authority to deploy and context for use. However, new questions arise in relation to the application of command responsibility with AWS.

First, and fundamentally, for *whose* conduct is the commander responsible in preventing war crimes in the use of AWS? That is, is an AWS considered a subordinate of the commander who deploys it? How are responsibility and decision-making allocated within the chain of command and control, and how does it change when AWS are deployed?

Second, since the material and mental elements to establish command responsibility are different than for individual commission (box 3.2), what are the standards of intent, knowledge and behaviour that could trigger command responsibility for war crimes involving AWS?

The following subsections outline some key issues these questions raise.

The commander–subordinate relationship in the context of AWS

One of the requirements to establish the material element of command responsibility—that is, the failure to prevent or punish a war crime—is the existence of a commander–subordinate relationship. This relationship is traditionally a human-to-human relationship. How should it be construed in the context of AWS?

One view is that the relationship between commanders and AWS resembles the relationship between commanders and their human subordinates.[155] If correct, a commander could be held responsible for failing to prevent and punish crimes *committed by AWS*. Another view is that the relationships are not analogous and that

[151] de Hemptinne, J. et al. (eds), *Modes of Liability in International Criminal Law* (Cambridge University Press: Cambridge 2019).

[152] See e.g. Zhang (note 19); Henderson, Keane and Liddy (note 38); McFarland (note 10); and Amoroso and Giordano (note 145).

[153] Rome Statute (note 104), Art. 25.

[154] Kraska, J., 'Command accountability for AI weapon systems in the law of armed conflict, *International Law Studies*, vol. 97, no. 407 (2021); Henderson, Keane and Liddy (note 38); and Amoroso and Giordano (note 145).

[155] Jain, J., 'Autonomous weapons systems: New frameworks for individual responsibility', eds Bhuta, N. et al, *Autonomous Weapons Systems: Law, Ethics, Policy* (Cambridge University Press: Cambridge, 2016,), p. 312.

Box 3.3. The doctrine of command or superior responsibility

The doctrine of command or superior responsibility stipulates that superiors—both military and civilian leaders—can be held criminally responsible for the war crimes their subordinates commit. Thus, a commander' or superior's criminal responsibility arises for failing to *prevent* or *repress* criminal acts committed by those under their command.

The doctrine of command responsibility is laid down in international law:

- Additional Protocol I lays down in Articles 86 and 87 an international legal obligation on parties to an armed conflict to require commanders to take all necessary measures to prevent war crimes from being committed or, if war crimes have been committed, to initiate disciplinary or penal sanctions against the perpetrators.[a]
- The Rome Statute provides in Article 28 that both military commanders and superiors are 'criminally responsible for crimes within the jurisdiction of the [International Criminal] Court' committed by, respectively, forces or subordinates under their command, authority and control, as a result of the commander's or superior's 'failure to exercise control properly over such forces' or subordinates, where: (i) the commander/superior 'either knew or, owing to the circumstances at the time, should have known' that the forces/subordinates 'were committing or about to commit such crimes'; and (ii) the commander/superior 'failed to take all necessary and reasonable measures' within their power 'to prevent or repress their commission or to submit the matter to the competent authorities for investigation and prosecution'.

[a] International Committee of the Red Cross, Customary IHL Database, [n.d.], 'Rule 153. Command Responsibility for failure to prevent, repress or report war Crimes'.

command responsibility is triggered only by 'chargeable offences' being committed by human subordinates.[156] To be chargeable, an offence must be committed with the required *mens rea*.[157] For this to apply to an AWS requires the AWS to be capable of acting with intent or knowledge—and also begs the question of whether an AWS can be punished. This view is contrary to the GGE's guiding principles that, for example, 'accountability cannot be transferred to machines' and 'emerging technologies in the area of lethal autonomous weapons systems should not be anthropomorphized'.[158] Moreover, individual criminal responsibility, at least under the Rome Statute, relates only to natural persons.

A second fundamental question relates to the core element of the commander–subordinate relationship: the 'effective command and control' on the part of the commander vis-à-vis the subordinate. Military commanders are those 'who are formally authorized to exercise military command or have *de facto* command authority within a military organization (*de iure* and *de facto* commanders)'.[159] However, the decisive factor of the 'effective command and control' requirement is the 'actual capability to effectively influence the conduct' of the subordinates.[160] Determining the capacity of commanders to effectuate their command and control in the use of AWS is a new and complex issue. The answer depends on how the question concerning the superior–subordinate relationship is addressed—that is, whether commanders exercise effective command and control only over a human subordinate operating an AWS or also over the AWS itself. Effective command and control over human subordinates is arguably achieved through institutional and organizational means.[161] In the case of an AWS, effective command and control might depend on the weapon's programming and design, and to what extent commanders are able to modify and correct its instructions, and to supervise and override the system. Some have argued that a commander retains effective control if the AWS remains under their close

[156] Bo., M, 'Meaningful human control over autonomous weapon systems: An (international) criminal law account', *OpinioJuris*, 18 Dec. 2020; and Chengeta (note 54), p. 32.

[157] Crootof (note 135).

[158] CCW Convention, GGE, 'Guiding principles' (note 3), paras (b) and (i).

[159] Werle, G. and Jessberger, F., *Principles of International Criminal Law* (Oxford University Press: Oxford, 2020), p. 691.

[160] Werle and Jessberger (note 159), p. 691.

[161] Buchan and Tsagourias (note 132), p. 651.

supervision 'through a constant and real-time monitoring mechanism' that allows the commander 'to adjust the algorithm to modify instructions, assign new tasks, or correct glitches', as well as having 'the ability to abort operations or deactivate [the AWS] if it starts to behave unexpectedly or once it has successfully completed its mission'.[162] In other words, what types and degrees of human–machine interaction are needed not only for compliance with IHL but also for establishing the 'effective command and control' requirement for applying command responsibility?

Addressing the commander–subordinate relationship in the context of AWS and the related question of 'effective command and control' over the use of AWS is fundamental to ensuring that the doctrine of command responsibility in instances of war crimes involving AWS applies to the commander who deployed the AWS.

The standards of knowledge required of a commander to satisfy the mental element of a war crime

Another set of complexities relates to the mental element of command responsibility. Command responsibility has a significantly lower mental element requirement than individual commission of war crimes: commanders are criminally responsible if they *knew or should have known* that subordinates committed, were committing or were going to commit a war crime (box 3.3); the Statute of the International Criminal Tribunal for the former Yugoslavia (ICTY Statute) uses the phrase 'had reason to know'.[163] In other words, the mental element of knowledge but also the element of negligence applies in that commanders could be held responsible for *failing to acquire knowledge*.[164] For this reason, the doctrine of command responsibility provides a useful framework in the context of AWS since 'at least the issue of intent and knowledge as required by other . . . modes of liability, could appear to be circumvented'.[165]

Despite a lower mental element requirement, determining the extent to which a commander has a proactive duty to acquire information concerning war crimes involving AWS remains complex. This determination of the mental element for command responsibility raises questions which form part of old debates but are also AWS-specific. According to an interpretation given by the ICTY, customary international law allows a presumption of negligent lack of knowledge if the commander had information that 'put him on notice of offences committed by subordinates', while under the Rome Statute it is crucial to determine whether the commander would, in the exercise of their duties, have gained knowledge of the commission of the crime by their subordinates.[166] But what, specifically, does a commander's proactive duty to seek and scrutinize information concerning an attack with an AWS entail? Some argue that the duty to acquire knowledge would include, for example, keeping up to date with technological developments.[167] According to this view, commanders could be considered negligent if they knew that a software update available to them could make an AWS more accurate, yet failed to implement it.[168] It follows that the standard of knowledge required of commanders concerning AWS may vary from system to system and be contingent on the environment of use.

[162] Buchan and Tsagourias (note 132), p. 657.

[163] Statute of the International Criminal Tribunal for the former Yugoslavia, UN Security Council Resolution 827, S/RES/827, 25 May 1993, Art. 7(3).

[164] Zhang (note 19); and Buchan and Tsagourias (note 132), p. 661.

[165] Acquaviva (note 144).

[166] See Werle and Jessberger (note 159), p. 272, paras 697 and 698.

[167] Buchan and Tsagourias (note 132), p. 661.

[168] Buchan and Tsagourias (note 132), pp. 661–62.

The role of AWS developers as participants in a war crime

The preprogrammed nature of AWS makes the development stage a key phase where critical decisions about targeting are made. Developers (i.e. engineers, designers and programmers) play a significant role in defining the behaviour of an AWS. Thus, it has been argued in the international policy debate that developers are among the range of subjects to be considered for the attribution of individual criminal responsibility for war crimes involving AWS.[169] However, it remains unclear whether and how developers may be held responsible for participating in the commission of a war crime that involved an AWS they helped to develop.[170]

This question raises two sets of issues. The first relates to traceability and attribution. Development of AWS is distributed across several actors, including 'different teams of individuals, different producers or even asynchronously and for different purposes'.[171] Tracing back responsibility to one of these individuals and identifying who among the various programmers, designers, data labellers and others are responsible for a certain act or omission that led or contributed to a war crime is a significant challenge. A complicating factor is that military commanders could adapt the parameters of an AWS during deployment, thus potentially blurring the distinction between developers and users. The second set of issues, which is the focus here, concerns the determination of the material and mental elements of developers' criminal responsibility for war crimes.

Establishing conduct that gives rise to developers' participation in a war crime involving an AWS

Developers can be held to account under modes of criminal responsibility that refer to either perpetration of or participation in a crime (box 3.1), but modes referring to *participation* seem most suitable for holding developers to account for their role in contributing to unlawful attacks involving AWS. A useful provision is Article 25(3)(c) of the Rome Statute on 'aiding and abetting', because the material element 'can be remote from the time and location of where the crime in question is committed'.[172] Article 25(3)(c) holds an individual criminally responsible 'for the purpose of facilitating the commission' of a crime if the individual 'aids, abets or otherwise assists in its commission or attempted commission, including providing the means for its commission'. A crucial question is thus what specific conduct on the part of the developers of an AWS amounts to aiding, abetting, assisting or providing the means for committing a war crime arising from the use of the AWS during an armed conflict? This question raises three main issues.

The first relates to whether the temporal and geographical distance between the developers' conduct and the armed conflict is sufficiently close to establish the required nexus for war crimes. The Rome Statute requires that: 'The conduct took place in the context of and was associated with an international armed conflict'.[173] This indicates that war crimes law applies 'from the initiation of . . . armed conflicts and beyond the cessation of hostilities until a general conclusion of peace is reached' and that a sufficient nexus must be established between the conduct and the armed conflict.[174]

[169] McFarland, T. and McCormack, T., 'Mind the gap: Can developers of autonomous weapons systems be liable for war crimes?', *International Law Studies*, vol. 90, no. 1 (2014); and Amoroso and Giordano (note 145).

[170] Bo, M., 'Are programmers in or "out of" control? The individual criminal responsibility of programmers of autonomous weapons and self-driving cars', ed. S. Gless, *Human–Robot Interaction in Law and its Narratives: Legal Blame, Criminal Law, and Procedure* (Cambridge University Press, 2022) (forthcoming).

[171] Amoroso and Giordano (note 145), p. 215.

[172] Ventura, M., 'Aiding and abetting', eds de Hemptinne et al. (note 151), p. 185.

[173] Rome Statute (note 104), Art. 8; see also International Criminal Court (note 125).

[174] Dörmann (note 116), pp. 19 and 20.

Therefore, where the developers' conduct occurs before the beginning of an armed conflict, it could be difficult to prove the threshold for the temporal applicability of IHL and the contextual element of war crimes.[175] However, it is conceivable that developers would, in some contexts, program or upgrade AWS software during an armed conflict, or that developers exercise such control over the capabilities of an AWS that their conduct could causally be linked to the commission of war crimes.[176]

The second issue relates to a developer's responsibility for the technical characteristics of an AWS. Arguably, one pathway through which the responsibility of developers could be engaged is through their participation in developing an AWS that is *indiscriminate by nature*, and therefore prohibited under IHL. However, this pathway is problematic given that the basis for deeming AWS as inherently indiscriminate remains uncertain. Moreover, it would have to be shown that the development of the AWS, and therefore those involved in that development, had a *substantial effect* on the commission of a war crime.[177] This raises complex questions related to the establishment of a causal link between the design of an AWS and unlawful consequences arising from its use.[178] It also brings to the fore long-standing debates on what exactly constitutes a *substantial contribution* or *effect* in aiding and abetting.[179]

The third issue concerns a developer's contribution to an attack through their role, for example in the parametrization of the AWS targeting parameters. Arguably, their responsibility could be engaged if it could be established that they contributed to the *indiscriminate use* of a weapon that is not *indiscriminate by nature*. However, this would depend on each developer's actual role and contribution in the *use* phase, which would, among other factors, in turn depend on their integration into the chain of command and in the targeting process.[180]

Establishing the mental element of developers' participation in a war crime involving an AWS

Assisting the commission of a war crime requires a different and lower mental element threshold than the individual commission of a war crime. To ascribe criminal responsibility to a developer of an AWS on the ground of aiding, abetting, assisting or providing the means for committing a war crime that involved the AWS, it would be necessary to prove that the developer *knew* of the existence of the armed conflict in which the AWS was being deployed, *and* that their conduct would assist the commission of a war crime. However, the diffuse nature of AWS development, being fragmented among different actors both internal and external to armed forces, makes it challenging to prove these knowledge requirements on the part of developers. Moreover, the Rome Statute provisions have a higher mental element, where the developer must have acted 'for the purpose of facilitating the commission' of a war crime.[181] The implications of this requirement, 'purpose' in addition to knowledge, continue to attract controversy among commentators.[182] This elevated mental element is arguably difficult to prove and a challenge to holding developers criminally responsible for war crimes.[183]

[175] McFarland and McCormack (note 169), p. 374.
[176] Bo (note 170); and McFarland and McCormack (note 169), p. 366.
[177] Werle and Jessberger (note 159), p. 669.
[178] Ventura (note 172), p. 177; and Bo (note 170).
[179] Ventura (note 172).
[180] View expressed by some state representatives and legal experts, Experts workshop, Online, 8 Feb. 2022; and Buchan and Tsagourias (note 132), p. 659.
[181] Werle and Jessberger (note 159); McFarland (note 10), p. 674; and McFarland and McCormack (note 169), p. 380.
[182] Ventura (note 172), pp. 74ff.
[183] McFarland and McCormack (note 169).

IV. Summary

Establishing individual criminal responsibility for war crimes is not a straightforward exercise, and AWS are not making it easier. AWS reopen unresolved legal disputes around how the elements of a war crime can, or should, be established. With regard to the material element, AWS highlight the legal uncertainty around what constitutes the commission of an indiscriminate attack, and whether omission is a mode of commission of a war crime. AWS also point to the lack of criminalization of the violation of the duty to take precautions.

Regarding the mental element for war crimes, AWS underline existing uncertainties regarding the criminalization of risk-taking behaviours, such as whether being reckless about the effects of an attack amounts to a war crime.

The preprogrammed nature of AWS combined with the complex network of actors involved in the development and use of AWS prompts questions about the different modes through which individual criminal responsibility can be established. In particular, to what extent should commanders be held responsible for the use of increasingly complex and unpredictable weapon systems? And under what conditions can individuals involved at earlier stages, such as developers and programmers, be held responsible for the commission of a war crime?

If human responsibility is to be retained through the framework of individual criminal responsibility, clarification is needed in at least two parallel aspects. First, what conduct the rules of IHL prohibit, require and permit needs clearer articulation to establish what conduct across the life cycle of an AWS, including the development phase, would fulfil the material element of a war crime. Second, the standards of intent, knowledge and care on the part of the different actors, spanning from developers to commanders and users, required by IHL and by the rules on individual criminal responsibility, needs clear elaboration.

4. Challenges and opportunities for investigating IHL violations involving AWS

From key provisions of the Geneva Conventions and additional protocols, it flows that states must be able to discern, scrutinize and attribute conduct that is a potential breach of IHL (box 4.1). In the case of a grave breach—that is, a serious violation of IHL that amounts to a war crime—states are obliged to initiate investigations, 'search for' individuals who have committed (or otherwise facilitated) a breach, and 'repress' such conduct. In the case of any other IHL violation (not amounting to a grave breach), states are obliged to 'suppress' such violations. These obligations are reflected in IHL as well as in customary international law, and failing to conduct investigations of allegations of serious violations of IHL, including war crimes, constitutes non-compliance with a state's international obligations.[184] Ensuring the practical ability to investigate harmful incidents is, therefore, critical to attributing human responsibility, including state responsibility and individual criminal responsibility for unlawful conduct.

In relation to AWS, it remains both unclear and underexplored what implications AWS have on states' practical ability to investigate potential IHL violations and to hold humans accordingly responsible. The GGE has briefly addressed this issue by agreeing that 'states must provide mechanisms to ensure accountability for any violations of their obligations under international law . . ., including by providing for investigations of reasonable suspicions of violations and bringing perpetrators to justice'.[185] Along the same lines, some states have explicitly called for the need to establish procedures and mechanisms for reporting incidents that involve AWS.[186]

This chapter provides an overview of existing investigation practices and considers the implications of AWS for the ability to trace back conduct, that is, to discern, scrutinize and attribute violations of IHL. Section I outlines the contours of existing processes and mechanisms in place for investigating IHL violations. Section II discusses the implications of AWS on the practical ability to investigate and explain harmful incidents, and ultimately to attribute responsibility, by exploring the implications of collecting and assessing evidence around the incident in question. Finally, section III summarizes the chapter's main findings. The focus is on the investigation process only; accountability mechanisms to institute prosecutions or other legal proceedings directed at either states or individuals are outside the scope of this report.

I. Existing mechanisms and processes for investigating IHL violations

Assessing the implications of AWS for a state's ability to discern, scrutinize and attribute IHL violations first warrants a better understanding of existing mechanisms and processes to investigate potential IHL violations. IHL requires states to repress grave breaches and suppress other violations of IHL, but it does not specify the methods or standards for doing so. While the need to have robust and effective domestic processes in place for investigating violations of IHL is reflected in military manuals and reports, there are no universal guidelines for how investigations into potential IHL

[184] United Nations, General Assembly, Human Rights Council, 'Human rights in Palestine and other occupied Arab territories: Report of the United Nations Fact-Finding Mission on the Gaza Conflict', A/HRC/12/48, 15 Sep. 2009, p. 35.

[185] CCW Convention, GGE, CCW/GGE.1/2021/CRP.1 (note 10), annex III, p. 13.

[186] CCW Convention, GGE, 'Outline for a normative and operational framework on emerging technologies in the area of LAWS', Working paper by France and Germany, 27 Sep. 2021, CCW/GGE.1/2021/WP.5; CCW Convention, GGE, 'US proposals on aspects of the normative and operational framework', Working paper by the USA, 27 Sep. 2021, CCW/GGE.1/2021/WP.3; and CCW Convention, GGE, 'Elements for a future normative framework conducive to a legally binding instrument to address the ethical humanitarian and legal concerns posed by emerging technologies in the area of (lethal) autonomous weapons (LAWS)', Commentary submitted by Brazil, Chile and Mexico, 2019, p. 7.

Box 4.1. States' obligations to repress grave breaches and suppress any other violation of the Geneva Conventions and the additional protocols

States parties to the Geneva Conventions (GCs) and additional protocols (APs) are obliged to **repress any act that is a grave breach**—that is, a serious violation amounting to a war crime. This entails a state's obligations to punish the responsible individuals, notably to adopt legislative measures to 'provide for effective penal sanctions', and to search for and bring before their own courts or surrender to another contracting party those allegedly responsible for a grave breach, regardless of their nationality.[a]

States are also obliged to **suppress any act that is a violation of IHL** other than a grave breach. This entails a wide range of measures including the institution of judicial or disciplinary proceedings, or the adoption of legislative, administrative or other regulatory measures to prevent, prohibit and punish violations of the GCs and APs other than grave breaches. This also includes the criminalization of violations of the GCs and APs beyond the list of grave breaches.[b]

[a] Geneva Conventions and additional protocols (note 11): GC I, Art. 49; GC II, Art. 85; GC III, Art. 129; GC IV, Art. 146; AP I, Art. 85. See also International Committee of the Red Cross, Customary IHL Database, [n.d.], 'Rule 158. Prosecution of war crimes'.

[b] GC I, Art. 49(3); GC II, Art. 50; GC III, Art. 129; GC IV, Art. 146; AP I, Art. 86(1); ICRC, 'Commentary [to GC I (note 11)] of 2016, Article 49: Penal sanctions', 2016, paras 2896–97. See also Gaeta, P., 'The interplay between the Geneva Conventions and international criminal law', eds A. Clapham, P. Gaeta and M. Sassoli, *The 1949 Geneva Conventions: A Commentary* (Oxford University Press: Oxford, 2015).

violations should be carried out.[187] And generally, states may exercise discretion in terms of how they wish to conduct their inquiries, including how (and what) evidence is gathered and according to which thresholds of evidence it is assessed.[188] Moreover, states are not obliged to share detailed information about their processes and rarely do so. As a result, existing national mechanisms to investigate unlawful conduct not only vary greatly but may also be subject to little public oversight. However, drawing on publicly available information and interviews with states, this section aims to provide an overview of known types of investigations and the variety of existing mechanisms for collecting evidence.

Types of investigations and how they are initiated

The types of investigations that can be initiated vary depending on the type of incident. An overall distinction should, however, be made between two broad categories: internal investigations and external investigations. The most common type is the internal investigation carried out within a state, where the state investigates conduct by its own agents or other similarly positioned individuals. Internal investigations can inquire into all types of IHL violations. Some are characterized as routine assessments triggered by the mere application of force, while other investigations are triggered only by allegations of grave breaches. External investigations refer to a more limited set of situations where a harmful incident potentially amounting to a grave breach is investigated by a state with no jurisdictional links to the incident or by external bodies.

Apart from routine investigations, investigations are triggered by internal or external reports about potential breaches. Processes for internal reporting may be through 'self-reports' from members of the armed forces or via standardized collateral damage assessments conducted by commanders.[189] External reports about potential

[187] For guidelines in military manuals and reports see e.g. US Department of the Army, 'Criminal investigation activities', Army Regulation 195-2, 21 July 2020; Open Society Justice Initiative, 'Comparative analysis of preliminary investigation systems in respect of alleged violations of international human rights and/or humanitarian law', 10 Aug. 2010; and Schmitt, M. N., 'Investigating violations of international law in armed conflict', *Harvard National Security Journal*, vol. 2 (2011).

[188] Regarding the 'sufficient evidence' threshold see ICRC, 'Commentary [to GC I (note 11)] of 2016, Article 49: Penal sanctions', 2016, para. 2861.

[189] See e.g. McNerny, M. J. et al., *US Department of Defense Civilian Casualty Policies and Procedures: An Independent Assessment* (RAND Cooperation: Santa Monica, CA, 2022), p. 10; Khalfaoui, A. et al., *In Search of Answers: US Military Investigations and Civilian Harm* (Center for Civilians in Conflict and Columbia Law School Human Rights Institute:

violations of IHL can come from other states, members of the civilian population or civil society organizations. Some states have established permanent mechanisms to facilitate external reporting.[190] Once a state receives a report, it will usually make an initial assessment as to whether there is reasonable ground to initiate an investigation into the incident. In the USA, for instance, if the report is categorized as 'credible' (in contrast to 'non-credible'), an investigation will be launched.[191] It is important to note that the investigative process does not necessarily distinguish between different types of responsibility, such as state or individual criminal responsibility, ahead of its launch. The starting point is the harmful incident and the aim is to first establish the facts and then subsequently identify the responsible agents (if any). The following, non-exhaustive list provides an overview of the types of investigations that may be initiated both internally and externally.

Post-strike assessments

Some states conduct automatic assessments following every application of force, regardless of the result. Such investigations—usually labelled 'post-strike', 'collateral damage' or 'battle damage' assessments—are, therefore, perhaps the most common type of 'investigation'. In many militaries, these are standardized procedures that serve as the final stage of the targeting cycle. Generally, the aim is to evaluate conduct, learn from mistakes and assess potential claims of civilian harm. Depending on the findings, assessments can result in criminal prosecutions, claims for compensation or reparation, or *ex gratia* payments.[192]

Safety investigations

Some states, such as the USA, have established processes to facilitate safety investigations.[193] These are usually triggered by technical mishaps and serve to inquire into the technical performance of systems. In the USA, for example, safety investigations can trigger a subsequent 'accident investigation' that may take additional steps by also interviewing witnesses and carrying out additional testing if needed.[194]

Administrative investigations

Administrative investigations are a widespread practice in many states (including Australia, Canada, the UK and the USA) that serve as 'fact-finding' missions to determine the facts around an incident.[195] Administrative investigations can inquire into both the alleged commission of war crimes (and potentially trigger a referral to criminal investigation) and a broader set of potential breaches. Their broader scope allows administrative investigations to identify both systemic issues and individual conduct that may not amount to a serious violation but that need to be addressed by non-judicial means.[196] Administrative investigations can therefore help states to address and punish non-criminal violations of IHL as well as war crimes.

New York, 2020) pp. 13–14; and US Joint Chiefs of Staff, 'Methodology for combat assessment', CJCSI 3162.0, 8 Mar. 2019.

[190] Bijl, E., 'Civilian harm reporting mechanisms: A useful means to support monitoring and accountability?', PAX for Peace, 2022.

[191] US Department of Defense, 'Annual report on civilian casualties in connection with United States military operations in 2020', 29 Apr. 2021, p. 6.

[192] ICRC, Customary IHL Database, [n.d.], 'Practice relating to rule 150. Reparation'.

[193] See e.g. US Air Combat Command, 'Air Force Safety and Accident Board investigations', 25 Jan. 2019.

[194] US Air Combat Command (note 193).

[195] See Lubell, N., Pejic, J. and Simmons, C., *Guidelines on Investigating Violations of International Humanitarian Law: Law, Policy and Good Practice* (Geneva Academy of International Humanitarian Law and Human Rights and the ICRC: Geneva, 2019), p. 11; Schmitt (note 187), p. 79; and McNerny et al. (note 189), p. ix.

[196] Lubell, Pejic and Simmons (note 195), pp. 32–35.

Internal criminal investigations

Through their national authorities and in accordance with their national procedures, states are obliged to search for and bring before their own courts or surrender to another contracting party those allegedly responsible for a grave breach of the Geneva Conventions and additional protocols.[197] States are under an obligation to initiate criminal investigations aimed at determining the facts around a harmful incident that is suspected of amounting to a war crime committed by their own nationals or on their territory.[198] The investigation's result may trigger the prosecution of individuals if sufficient evidence (according to national standards) to bring a criminal charge is collected.[199] In some states, criminal investigations into violations of IHL are carried out by the military justice system, which may be based on a specific code of military procedure. Many states run internal disciplinary hearings within their military justice systems that may then evolve into criminal hearings. But, overall, state practices vary according to the separation of criminal versus disciplinary jurisdictions.

External criminal investigations

External investigations can come about due to the obligation of states to exercise universal jurisdiction over war crimes. In fact, the aforementioned obligation to search for and prosecute alleged offenders before a state's own courts must be carried out 'regardless of their nationality' or any other jurisdictional link.[200]

Other external investigations over IHL violations can be carried out through permanent or ad-hoc bodies with international or regional jurisdiction. Permanent bodies include the International Criminal Court (ICC), the International Court of Justice (ICJ) and the International Fact-Finding Commission.[201] Ad-hoc bodies include commissions of inquiry and fact-finding missions into specific incidents or conflicts. Such international investigative bodies can be established by the UN Security Council, General Assembly, Human Rights Council, Secretary-General and the High Commissioner for Human Rights. Usually, they are established through human rights law bodies, and therefore particularly mandated to investigate human rights violations.[202] Ad-hoc international tribunals can also be established to investigate and prosecute war crimes and other crimes against international law; past examples include the International Criminal Tribunal for Yugoslavia and the International Criminal Tribunal for Rwanda.

[197] Geneva Conventions (note 11): GC I, Art. 49; GC II, Art. 50; GC III, Art. 129; and GC IV, Art. 146; and AP I, Art. 85(1).

[198] 'Prosecutions for grave breaches could of course be based on . . . accepted titles of jurisdiction, such as territoriality, active and passive personality or the protective principle'. ICRC, 'Commentary [on GC I (note 11)] of 2016: Article 49: Penal sanctions' (note 188), para. 2862.

[199] Lubell, Pejic and Simmons (note 195), p. 11; and ICRC, 'Commentary [on GC I (note 11)] of 2016: Article 49: Penal sanctions' (note 188), para. 2864.

[200] Geneva Conventions (note 11): GC I, Art. 49; GC II, Art. 50; GC III, Art. 129; and GC IV, Art. 146.

[201] AP I (note 11), Art. 90(2)(c)(i); International Humanitarian Fact-Finding Commission, [n.d.]; and Pfanner, T., 'Various mechanisms and approaches for implementing international humanitarian law and protecting and assisting war victims', *International Review of the Red Cross*, vol. 91, no. 874 (2009).

[202] See e.g. United Nations, Human Rights Council, 'International commissions of inquiry, commissions on human rights, fact-finding missions and other investigations', [n.d.]; Council of the European Union, Council Decision 2008/901/CFSP of 2 December 2008 concerning an independent international fact-finding mission on the conflict in Georgia, 2008/901/CFSP, *Official Journal of the European Union*, L323/66, 3 Dec. 2008; United Nations, Human Rights Council, 'Report of the Human Rights Council on its fifteenth special session', A/HRC/S-15/1, 25 Feb. 2011, ch. 1, 'S-15/2. Situation of human rights in the Libyan Arab Jamahiriya'; United Nations, Human Rights Council, 'Commissions of inquiry into alleged human rights and IHL violations in occupied Palestinian territory and Israel', Resolution S-20/1, 27 May 2021; and United Nations, Human Rights Council, 'Commissions of inquiry into alleged human rights and IHL violations in Ukraine', Resolution 49/1, 7 Mar. 2022.

Using all feasible means to conduct effective investigations

Regardless of the type of investigation, a state is encouraged to use all feasible means to inform its inquiries.[203] The ability to collect, access and preserve information is critical to establishing the facts surrounding an incident and gathering sufficient evidence to address IHL violations and to identify and punish wrongdoers. Though there are no international guidelines on how this is ensured, it is widely recognized that investigations must be 'effective', indicating that they must be conducted in good faith and use all feasible means to establish the facts and causes surrounding an incident. Other universal principles pertaining to 'independence, effectiveness, promptness, and impartiality' also apply to investigations.[204]

How this translates into practice is, however, left to the discretion of each state. States may resort to a variety of sources to inform their inquiries, including operational data, field intelligence and visual imagery, as well as information provided by external sources, such as organizations operating in the area or local news outlets.[205] To the extent feasible, and if the investigatory body so wishes, an on-scene commander may be required to collect and preserve information relevant to a potential investigation related to an armed conflict.[206] However, if the armed conflict is still ongoing during the time of investigation or if there is limited physical access to the site, collecting on-scene information may not be considered 'feasible'.[207]

Investigatory bodies may also wish to access operational logs and internal reports generated before and after the incident. To this end, it is considered good practice for states to implement 'forward-looking' recording and reporting mechanisms to document conduct, as these can help inform a potential investigation. Such mechanisms are critical to ensure that 'nothing disappears along the way' and that everything can be 'scrutinized and examined later'.[208] Types of reporting mechanisms include the submission of weekly reports and reports connected to use-of-force decisions. Mechanisms for recording decisions and actions could include digital logs and cameras attached to armed forces during military operations.

Based on analysis of the data acquired during the investigation, and the national standards of evidence, the investigatory body will decide whether further action is needed. If there is a finding of responsibility, judicial, disciplinary or administrative proceedings may follow.

Challenges related to existing investigatory mechanisms and processes

Effective investigations into harmful incidents in armed conflict are critical measures to give effect to states' obligations under international law. However, existing mechanisms and processes face several challenges and limitations.

First, questions around how (and what) evidence is gathered and assessed, especially by states' internal investigators, have revealed discrepancies between the allegations raised by external reports and the conclusions reached by internal investigatory bodies.[209] The discrepancy (combined with recurring harmful incidents) has generated criticism around existing methodologies for internal investigations and

[203] Lubell, Pejic and Simmons (note 195), p. 7.

[204] Lubell, Pejic and Simmons (note 195), p. 7. See also United Nations, General Assembly, Human Rights Council (note 184) p. 35.

[205] McNerny et al. (note 189), p. 14.

[206] Lubell, Pejic and Simmons (note 195), p. 16; and Schmitt (note 187), p. 80.

[207] Lubell, Pejic and Simmons (note 195), p. 26; and McNerny et al. (note 189), p. 15.

[208] View expressed by a governmental legal adviser, Interview with the authors, Online, 16 Feb. 2022. See also Lubell, Pejic and Simmons (note 195), p. 19.

[209] See e.g. Khan, A., 'The civilian casualties files', *New York Times*, 6 Apr. 2022; and Crootof (note 10), p. 55.

states' efforts to sufficiently understand and address causal links. For example, some observers have argued that internal investigations are often ineffective because of insufficient or non-existent recordings, bias towards military sources over civilian sources, and a general lack of standardized procedures; and that investigations are usually treated as stand-alone events and therefore fail to identify systematic issues or learn from past experience.[210]

Another challenge pertains to the ability of external investigators to access relevant information and gather evidence, which depends on the willingness of states to share information related to the incident being investigated. The exception is, however, cases where investigators are operating under the authority of a binding decision of the UN Security Council.

Finally, existing mechanisms are limited in the sense that they are often only triggered by allegations of serious violations amounting to war crimes, such as violations of the principles of distinction and proportionality. According to observers, the focus on serious violations is problematic as it risks 'jettisoning critical parts of the broader IHL infrastructure which do not lend themselves to criminalization'.[211] It is important to recall that states remain responsible for respecting and ensuring respect for all IHL provisions, even if several states currently fail to adopt the necessary administrative or non-judicial disciplinary measures to do so.

II. Implications of AWS for investigating violations of IHL

In developing and using AWS, states must ensure they have the technical ability to trace back conduct relating to an incident and identify the wrongdoers (if any). Many states have explicitly taken steps towards this (e.g. the USA has made 'traceability in AI' a priority), while others (e.g. Portugal) have expressed the view that 'the use of force must be planned and executed in such a way that it can always be retraceable to the human being operating the machine, in order to prevent any accountability gaps for violations of international law'.[212] Ensuring traceability in the context of AWS is fundamental for states to comply with their obligations to repress and suppress violations of IHL and, more broadly, to uphold and observe their commitments under international law. Yet, the question of how to ensure effective investigations of incidents involving the use of AWS has received little attention in the debates about responsibility and AWS.[213] This section aims to address this question through a review of the opportunities and risks that the introduction of autonomy presents for the investigation of violations of IHL.

Implications for evidence gathering

The ability to gather evidence is critical to an effective investigation. The following subsections in turn address aspects of AWS that have implications on the ability to gather evidence.

[210] See e.g. Crootof (note 10), p. 54; Hartig, L., 'What counts as sufficient transparency on civilian casualties in Somalia', *Just Security*, 20 Apr. 2020; and McNerny et al. (note 189).

[211] Crootof (note 10), pp. 55 and 59.

[212] US Department of Defence (DOD), 'DOD adopts 5 principles of artificial intelligence ethics', DoD News, 25 Feb. 2020; and Portugal, 'Commentaries by Portugal on "Operationalising all eleven guiding principles at a national level"', Aug. 2020.

[213] See e.g. Bose, U., 'The black box solution to autonomous liability', *Washington University Law Review*, vol. 92, no. 5 (2015).

Challenges related to opacity of complex systems

Certain technical characteristics associated with AWS could potentially undermine the ability to gather evidence pertaining to a harmful incident. For instance, reliance on machine learning is particularly associated with certain traceability issues. Machine learning is an approach to software development that involves training the system to learn from data to improve its performance on specific tasks. Its main advantages are that it removes the need for hand-coded programming and it is efficient for automating tasks that require advance pattern recognition such as target recognition.[214] The disadvantage, especially in an investigative context, is that machine learning algorithms only allow users to understand system inputs and outputs, but not necessarily the process in between—how a system arrived at a certain conclusion.[215] This challenge, often referred to as the 'black box' of AI, constitutes a challenge from an evidence-gathering perspective, as it could prevent investigators from accessing information around why and how an AWS arrived at certain decisions.[216] A deeper understanding of what capabilities would complicate, or even prevent, investigators from sufficiently informing their inquiries is needed. States could consider to what extent making technical expertise available in the investigation process would improve the ability to access information of a more technical character. Moreover, states could also consider the deployment of measures aimed at preventing the 'black box' concerns in the design and acquisition of AWS.[217] That could include making assessments of opacity and understandability of the AWS part of the legal review process.

Opportunities for recording and documenting

Increased reliance on computing in warfare could potentially strengthen the ability to collect evidence. This is notably through forward-looking scrutiny measures related to recording and documenting conduct.[218] This is already the case with current technologies, where some militaries benefit from the increased wealth of information offered by audio- and video-recording technologies.[219] While the potential advantages may not be unique to AWS as such, auditable algorithms, digital trails and logs are among the technical features associated with AWS that contain the potential to help inform an investigation.[220] These features could form the basis of a 'glass box' (instead of a 'black box'), which would give investigators access to key information about what the systems did and on what basis.[221] Some states, therefore, associate AWS with significant opportunities concerning investigating and suppressing unlawful conduct. For example, the USA has argued that AWS 'could strengthen the implementation of IHL, by . . . facilitating the investigation or reporting of incidents involving potential violations, enhancing the ability to implement corrective actions, and automatically generating information on unexploded ordnance'.[222] The UK has similarly argued:

[214] Boulanin, V. et al., *Artificial Intelligence, Strategic Stability and Nuclear Risk* (SIPRI: Stockholm, 2020), pp. 10–11.

[215] Holland Michel (note 37).

[216] Holland Michel (note 37); Holland Michel (note 49); Deeks, A., 'The judicial demand for explainable artificial intelligence', *Columbia Law Review*, vol. 119 (2019), p. 1832; ICRC, 'ICRC position on autonomous weapon systems', 12 May 2021, p. 7; and Copeland and Sanders (note 38).

[217] Copeland and Sanders (note 38).

[218] See e.g. CCW Convention, GGE, Joint working paper submitted by Costa Rica, Panama, Peru, the Philippines, Sierra Leone and Uruguay, 2021, p. 6; CCW Convention, GGE, 'Implementing International humanitarian law in the use of autonomy in weapon systems', Working paper submitted by the USA, CCW/GGE.1/2019/WP.5, 28 Mar. 2019; and CCW Convention, GGE, Statement by the UK, 11 Apr. 2018.

[219] McNerny et al. (note 189), p. 16.

[220] Lewis, D., Blum, G. and Modirzadeh, N. K., 'War-algorithm accountability', Harvard Law School Program on International Law and Armed Conflict, Research briefing, Aug. 2016, p. 98; and Copeland and Sanders (note 38).

[221] Gubrud, M. and Altmann, J., 'Compliance measures for an autonomous weapons convention', International Committee for Robot Arms Control Working paper no. 2, May 2013.

[222] CCW Convention, GGE, 'Implementing international humanitarian law in the use of autonomy in weapon systems', CCW/GGE.1/2019/WP.5 (note 218), para. 2(c).

> Ultimately, if a decision has been taken to field any capability, there should be an auditable trail of the decision makers and a record of their assessments on the suitability of the system for use in a specific theatre or phase of operations. Accountability might even be improved if the automated recording systems that an autonomous system would need to legally operate provide better evidence to support subsequent investigation in the event of an incident.[223]

However, a better understanding of how these technical features could be utilized from an investigatory perspective is needed. States could usefully consult technical experts who could, among others things, provide insights on how recording mechanisms can be incorporated into AWS during the development and design phase. In similar discussions pertaining to civilian uses of autonomous vehicles, the implementation of flight data recorders and event data recorders have been suggested as a way to explain events and potentially assign responsibility.[224] Also, if new computing processes are being used to inform investigations, states would need to discuss how to ensure that such processes are properly incorporated into existing recording mechanisms and that their armed forces receive training in using and accessing such logs. There is also the sensitive question as to whether the information generated by digital logs may be made equally available to internal and external investigators. States should consider this aspect to prevent exacerbating existing problems related to external bodies' access to evidence.

Implications for assessing evidence

Besides the practical task of *gathering* evidence, a subsequent critical task for an effective investigation is *assessing* the evidence collected. The implications of AWS for the ability to assess evidence are considered below.

Challenges in discerning accidents from breaches

The complex technical characteristics of AWS raise particular challenges around distinguishing accidents from breaches of IHL. Evaluations of past investigations suggest that this is already a problem, with some being criticized for too often concluding that harmful incidents were due to regrettable accidents rather than systemic issues or individual misconduct.[225] However, the technical complexity of AWS risks exacerbating this challenge by making it increasingly hard to distinguish, for example, a product or design defect from a user's malicious intent or systemic negligence by the state.[226] To effectively investigate an incident, an investigator has to be able to understand how the weapon works, including its technical anatomy.[227]

In addition to the potential associated with auditing mechanisms incorporated into AWS design (the 'glass box'), one way to address this challenge would be for states to seek a better technical understanding of the types of incidents and technical failures that may be associated with AWS. Such a categorization exercise could help separate IHL violations from accidents (see discussion in chapters 2 and 3).[228] This measure could be applied as part of strengthened safety investigations.

Moreover, rigorous testing, evaluation and verification at the point of acquisition of an AWS could also improve the technical understanding of the system, which would

[223] CCW Convention, GGE, Statement by the UK (note 218), para. 6.

[224] Bose (note 213), p. 1348.

[225] See e.g. Fidell, E. R., 'The missing Kabul drone strike report', *Just Security*, 5 Nov. 2021; and Lewis, L., 'Hidden negligence: Aug. 29 drone strike is just the tip of the iceberg', *Just Security*, 9 Nov. 2021.

[226] Bose (note 213), p. 1338; and Lubell, Pejic and Simmons (note 195), p. 12.

[227] Schmitt (note 187), p. 84.

[228] Holland Michel (note 49), p. 21; Bose (note 213); and Dickinson, L., 'Lethal autonomous weapons systems: The overlooked importance of administrative accountability', eds E. Talbot Jensen and R. Alcala, *The Impact of Emerging Technologies on the Law of Armed Conflict* (Oxford University Press: Oxford, 2018), p. 46.

help investigators distinguish technical glitches from foreseeable and potentially unlawful accidents. These processes would feed into the due diligence obligations required in relation to AWS, to help identify and assess whether a state took all feasible steps to foresee and prevent the harmful incident.[229]

Implications related to assessing AWS users' intent, knowledge and due care

A critical task facing investigators when assessing any incident pertains to establishing the psychological attitude of those involved. Understanding what users of an AWS knew (or should have known) at the time of certain conduct is critical for attributing responsibility. However, the unique characteristics of AWS have several implications for this task.

First, as discussed in chapter 3, the technical complexity and unpredictability of AWS could undermine the ability to establish the mental element of alleged perpetrators of war crimes. One view is that a harmful incident could be traced back to a complex code, making the task of tracing harm back to a human's intent increasingly complicated.[230] The opposite view holds that AWS serve to effectuate the intent of the users—that is, users will program the system to attack a specific target when certain conditions are met, thus serving as an instrumentalization of the users' intent. Establishing the link between decisions and consequences will, according to this view, be made easier.[231] These opposing views highlight the need for states to elaborate on the standards of intent, knowledge and due care that IHL would require of AWS users. To this end, a better understanding of what information was made available to users, for example through weapons manuals, could improve the ability of investigators to assess whether users acted with intent, knowledge or negligence (box 3.2).

Second, having a clear scheme of responsibilities in place is important for assessing the intent, knowledge and due care of all those involved in decisions to use AWS. Schemes of responsibilities indicate touchpoints of decision-making that investigators can draw on to determine whether users had reason to know of or foresee potential harmful consequences of their decisions. However, as discussed in previous chapters, the preprogrammed nature of AWS entails some changes in how decisions to use force are (re)distributed across existing schemes of responsibility or chains of command. As reflected in the policy debate and consultations with states, however, the nature of the potential changes is not sufficiently understood. A deeper understanding of the ways in which AWS impact existing schemes of responsibilities is critical for states to conduct effective investigations. As a starting point, states could usefully share information about existing command and control structures and how these may be used to facilitate and inform investigations involving complex weapons systems and multiple user inputs.[232] Moreover, potential adjustments to consider could include a reorganization of the command chains that would, for example, allocate authority to authorize AWS at a higher level or ensure that developers too are captured in the structures.[233] Clarification could help not only trace back responsibility for IHL violations but also prevent them.

[229] View expressed by, among other states, the USA during informal GGE consultations in May 2022.
[230] McFarland (note 10), p. 163.
[231] View expressed by state representatives, Experts workshop, Online, 9 Feb. 2022.
[232] Holland Michel (note 49), p. 21.
[233] View expressed by state representatives, Experts workshop, Online, 8 Feb. 2022; and View expressed by state representatives, Interviews with the authors, Online, 22 Mar. 2022. The view is also reflected in the literature; for example, Massingham and McKenzie (note 127) illustrate the adjustments required by the AEGIS missile defence systems, categorized as a partially autonomous weapon system.

On the importance of administrative and safety investigations

Besides implications on *how* investigations are carried out, AWS may also have implications for *what* investigations are carried out. In addition to investigations of specific incidents that may be war crimes, other types of investigations are key in the context of AWS. As discussed in previous chapters, breaches of IHL involving AWS may be increasingly linked to collective failures to implement a web of interlinked obligations rather than an intentional breach, amounting to a war crime, carried out by an individual. Therefore, the ability to investigate conduct that allegedly does not amount to a serious violation of IHL but rather pertains to more systemic issues deserves particular attention. This points to the importance of conducting frequent and effective administrative investigations as a matter of routine. Administrative investigations have been subject to scant attention in the AWS debate but are relevant as they are triggered by a broader scope of incidents, rather than only those potentially amounting to a war crime.[234] As such, they readily serve to identify systemic issues about larger structures and processes flowing from the facilitative obligations of IHL.[235] Strengthening mechanisms to conduct administrative investigations will also address existing concerns raised over states' failures to adopt the necessary non-judicial and disciplinary measures in incidents not amounting to war crimes.

Moreover, as discussed in previous chapters, as complex systems AWS may be more prone to accidents than other weapons. Therefore, it is increasingly important to strengthen mechanisms for inquiring into technical aspects of AWS in the event of an unforeseen incident that causes harm. Such mechanisms should help to distinguish, for example, technical glitches from breaches, while also helping to improve the technical performance of the systems. This points to the importance of safety investigations as being increasingly relevant for accidents involving complex systems such as AWS.[236]

III. Summary

To retain human responsibility in the development and use of AWS, states must ensure they have the practical ability to trace unlawful conduct of IHL violations back to potential wrongdoers, including individuals and state agents. However, the extent to which AWS facilitate or undermine the practical ability to investigate IHL violations has so far been underexplored.

AWS are likely to have implications on the critical ability to *collect* and *assess* the evidence. Some of the technical features associated with AWS, such as digital logs and auditing mechanisms, may enhance both the collection and assessment of the evidence. In contrast, other technical features, especially those related to machine learning and opaque algorithms, may prevent investigators from accessing and collecting relevant information. Also, the unpredictability associated with the use of AWS could further complicate the task of assessing the evidence, such as whether an unintended incident should have been reasonably foreseen by the user(s). The ability to assess evidence in harmful incidents involving AWS also depends on how the legal frameworks of state responsibility and individual criminal responsibility are interpreted, as discussed in previous chapters. For example, how far back in time responsibility can be attributed, and what standards of intent and knowledge are needed for an act to be unlawful, serve as critical baselines for assessing evidence around incidents involving AWS. The need to clarify how the frameworks of state responsibility and individual

[234] An exception is Dickinson (note 228).
[235] Lubell, Pejic and Simmons (note 195), p. 12; and Schmitt (note 187), p. 79.
[236] View expressed by state representatives, Experts workshop, Online, 8 Feb. 2022.

criminal responsibility apply in the context of AWS is therefore also crucial from an investigatory perspective.

A deeper understanding around how states intend to ensure their practical ability to conduct investigations of incidents involving AWS, is a critical dimension of ensuring human responsibility. Moreover, approaching AWS from a traceability perspective—that is, ensuring that an AWS is 'discernible' 'scrutable' and 'attributable'—will be a useful avenue for identifying limits on AWS. An AWS that would preclude states from tracing back conduct would likely be incompatible with IHL and should not be deployed.

5. Key findings and recommendations

This report explored questions of how, in practice, states and individuals could be held responsible under existing law for IHL violations involving AWS. There are multiple legal frameworks through which human responsibility for IHL violations may be ensured. The report focused on the central legal frameworks: the rules governing state responsibility for internationally wrongful acts and individual criminal responsibility for war crimes. These rules fulfil different yet complementary functions with regard to the prevention of and accountability for IHL violations involving AWS. The rules governing state responsibility provide a framework for collective responsibility. They aim to provide accountability for any act or omission that would constitute a breach of a state's international obligations, and they cover the conduct of any agents whose acts or omissions are attributable to the state. The rules governing individual criminal responsibility for war crimes are meant to ensure an individualized form of accountability for certain serious violations of IHL. They provide a framework to prosecute individuals who commit or participate in, for instance, violations of the rules governing the conduct of hostilities. Together with IHL norms, these frameworks provide a comprehensive understanding of what international law demands from states and individuals to uphold respect for IHL in the development and use of AWS.

To generate useful insights for the policy process on the regulation of development and use of AWS, the report reviewed the conditions necessary to impose state responsibility for internationally wrongful acts and individual criminal responsibility for war crimes involving AWS. It addressed the following four questions:

1. *What act or omission* in the development and use of AWS would give rise to state responsibility or individual criminal responsibility (or both)?
2. *Whose conduct* in the development and use of AWS may engage state responsibility or give rise to individual criminal responsibility (or both)?
3. *What standards of intent, knowledge, behaviour and care* on the part of those involved in the development and use of AWS—including developers, decision makers, planners, commanders and operators—would give rise to state responsibility and individual criminal responsibility (or both)?
4. *How in practice* would unlawful conduct in the development and use of AWS be traced back to states and individuals? That is, how would IHL violations be discerned, scrutinized and attributed?

Section I of this chapter summarizes the answers that the report provides to these four questions, while section II provides a series of recommendations to the governmental and non-governmental experts that contribute to the international debate on AWS at the GGE and in other forums. These findings and recommendations aim to help determine which aspects of the normative and operational framework applicable to AWS may need to be further clarified or developed.

I. Key findings

Clarifications of IHL rules are needed to effectively understand what acts and omissions in the development of AWS would give rise to state responsibility for internationally wrongful acts and individual criminal responsibility for war crimes

Discerning what acts and omissions in the development and use of AWS would give rise to state responsibility and individual criminal responsibility remains challenging for two main reasons.

First, the rules governing state responsibility for internationally wrongful acts and individual criminal responsibility for war crimes are intrinsically linked to what the rules of IHL demand, permit and prohibit in the development and use of weapons, means and methods of warfare. The establishment of an internationally wrongful act giving rise to state responsibility depends on the normative standards established by IHL rules. The determination of whether a war crime triggering individual criminal responsibility has been committed depends also to a great extent on the interpretation of the fundamental rules of IHL. In the context of AWS, this means that the interpretation of what IHL rules require, permit or prohibit in the development and use of AWS is critical for the attribution of responsibility under these two frameworks. For instance, questions remain around whether and under which conditions AWS qualify as weapons that are indiscriminate by nature, and whether and under which conditions attacks that involve AWS amount to indiscriminate attacks. Other unresolved issues include what the IHL obligations to respect the principles of distinction, proportionality and precautions in attack demand in terms of human–machine interaction in the use of AWS; when compliance with these rules begins; and what is required from states to comply with 'facilitative' due diligence obligations aimed at securing respect for the fundamental rules of IHL.

Second, open interpretative questions around elements of these responsibility frameworks have significant implications for the establishment of state and individual criminal responsibility in the development and use of AWS. For instance, the requirement of 'human conduct' as a condition for attributing a violation of IHL involving AWS to a state, as well as how 'effective command and control' in the doctrine of command responsibility should be interpreted, are debated. This means that the basis for establishing that a state or individual violated the fundamental rules of IHL or failed to perform their duty under IHL may, in some cases, be unclear, or at least subject to different interpretations.

These problems highlight the need for the policy process on AWS to achieve more precision and common understanding around (*a*) what IHL compliance requires, permits and prohibits in the development and use of AWS, and the implications for the operation of the two responsibility frameworks; and (*b*) how each responsibility framework applies to the development and use of AWS.

Schemes outlining how responsibility is distributed among human agents in the development and use of AWS are needed for discerning *whose* acts or omissions engage state responsibility and individual criminal responsibility

The implementation of obligations under IHL is, in practice, a collective exercise carried out by multiple actors across different points in time and space. To ensure compliance with IHL, states should have schemes in place that delineate the roles and responsibilities of the people in charge of discharging their IHL obligations: *who* is responsible for doing *what*, *where* and *when*. The question of what such a scheme

should look like in the context of AWS remains largely unaddressed. In the GGE, states have agreed that responsibility for the development and use of AWS requires ensuring a responsible chain of human command and control, but have not elaborated on what constitutes such a chain in the context of AWS. A few states have enlarged at a general level on how they see the (re)distribution of roles and responsibilities for IHL compliance among the individuals involved in the development and use of AWS. However, many questions remain underexplored: (*a*) what is concretely demanded from those individuals; (*b*) when and where their roles and responsibilities start and end; (*c*) whether the responsibility for the use of AWS ultimately lies with a single person, be it a commander or another user; and (*d*) how these elements interact with one another.

The lack of common understanding around how the responsibility to comply with IHL in the development and use of AWS is allocated across multiple actors has practical implications for the identification and attribution of IHL violations that would trigger state responsibility or give rise to individual criminal responsibility (or both). First, concerning state responsibility, it is potentially difficult to determine (*a*) whether a state had a proper scheme in place to ensure that its obligations under IHL were duly implemented by all the agents involved in the development and use of AWS; (*b*) whether an agent of the state has breached any of their obligations under IHL; and (*c*) whose acts and omissions by (human) agents in the development and use of AWS could be considered a breach of IHL attributable to the state.

Second, concerning individual criminal responsibility, it is harder to trace back responsibility for the commission of a war crime within and beyond the military command-and-control chain. A critical issue, in that regard, is whether the responsibility for compliance with the principles of distinction, proportionality and precautions in attack is diffused across many actors or ultimately resides with the commander. States have presented mixed messages on this question. Some argue that decisions to use force involving AWS are distributed across a network of actors, while others argue that ultimately AWS are intended to carry out the intent of a single commander. Such uncertainty makes it difficult to discern the conditions under which a commander's responsibility for a war crime in the use of AWS could arise, and whether the applicable mode of responsibility is individual perpetration or participation in the commission of a war crime, or the doctrine of command responsibility. Similarly unclear are the conditions under which the acts or omissions of the other individuals involved in the development and use of AWS (such as developers) could amount to participation in the commission of a war crime.

Elaboration by states on these issues would not only strengthen the exercise of IHL compliance and prevent IHL violations, but also make it easier to detect and investigate unlawful conduct. Having a clear responsibility scheme in place that indicates points of decision-making would make it easier to determine which agents and individuals did not perform their legal obligations.

Defining the standards of intent, knowledge, behaviour and care required from the actors involved in the development and use of AWS is needed for the determination and attribution of responsibility

States are collective entities acting through human agents. War crimes are committed by individuals. Therefore, the standards of intent, knowledge, behaviour and care on the part of those involved in the development and use of AWS—including developers, decision makers, planners, commanders and operators—are critical to the task of establishing whether IHL was violated in a way that could give rise to state

responsibility or individual criminal responsibility. However, states have not settled the kinds and degrees of these standards.

For instance, in the context of state responsibility, it has been debated whether a violation of the principle of distinction can only be 'deliberate'—as opposed to an unintended violation flowing from lack of care or due diligence in taking precautionary measures on the part of the user of a weapon. This open interpretative question in turn depends on other unresolved issues that arise in the context of AWS use: what should the user of an AWS be able to reasonably foresee and do to ensure the effects of the weapon are directed at a specific military objective and that the weapon does not target civilians or other protected individuals or objects, or have disproportionate effects? This answer is critical for determining the basis on which state agents may breach any of the IHL fundamental rules.

With regard to individual criminal responsibility, it remains disputed: (*a*) whether, and on what basis, recklessness may satisfy the mental element of perpetrating or participating in the commission of a war crime; and (*b*) whether omissions, such as a failure to suspend an attack with AWS expected to be unlawful, may amount to war crimes. These uncertainties are not novel but find new resonance in the context of AWS. The preprogrammed nature of AWS raises questions not only about how the intent of the developers, decision makers, planners, commanders and operators may be formulated and then effectuated by the system during an attack, but also about what they should be able to foresee and do to ensure that the effect will not be unlawful. It is commonly agreed that a scenario where a commander or user of an AWS would *intentionally* pre-program the AWS to attack people and objects that are protected under IHL, or *intentionally* launch or not suspend an attack with an AWS that is expected to have indiscriminate and disproportionate effects, could give rise to state responsibility and individual criminal responsibility under existing provisions. In contrast, a scenario where an AWS attack results in unlawful effects as a result of negligence or recklessness on the part of the user is subject to debate. Such a scenario could involve situations where the user did not seek sufficient information about the specific target or the potential presence of protected objects and people in the target area, or ignored the possibility that the AWS might underperform in the specific environment of use and therefore potentially cause harm.

This finding stresses the need for the policy process on AWS to discuss and elaborate on the standards of intent, knowledge, behaviour and care that IHL compliance demands from the different actors involved in the development and use of AWS. Fleshing out what the different types of actors should reasonably foresee and do to ensure that the effects of AWS are lawful will not only strengthen IHL compliance but also facilitate the task of discerning when these actors intentionally engage in behaviour that triggers state responsibility or individual criminal responsibility.

Ensuring the practical ability to trace back IHL violations to potential wrongdoers is critical for retaining human responsibility in the development and use of AWS

States are obliged to repress grave breaches of IHL and suppress any other violations of IHL. To perform these obligations, states need to have investigation mechanisms in place that allow them to discern, scrutinize and attribute IHL violations. However, states enjoy discretion in how they choose to implement these obligations. National practices for investigating potential breaches of IHL therefore vary. The extent to which the characteristics of AWS affect states' practical ability to investigate potential

violations of IHL is an important, yet largely overlooked, dimension of the debate on human responsibility.

The defining features of AWS are likely to have implications for how responsibility for IHL violations involving AWS can be investigated and traced back to humans. The technical characteristics of AWS present both opportunities and challenges regarding the ability to gather and assess evidence during investigations into harmful incidents. On the one hand, the use of digital logs and automatic preservation of information could improve the ability to collect evidence. On the other hand, the so-called black box problem of AI could prevent investigators from accessing key information about the incident. Arguably, issues arising from the unpredictability and opacity associated with AWS and AI-based functions could constitute serious technical obstacles to an effective investigation, such as discerning whether a harmful incident is the result of a technical glitch or a human's unlawful conduct.

Second, the potential redistribution of roles and responsibilities associated with the use of AWS may make it increasingly difficult to trace back unlawful conduct by responsible humans in the chain of command and control. However, as suggested in the previous finding, there is an opportunity for states to (re)elaborate and formalize aspects of the decision-making process in the command-and-control chain where AWS are involved. Having a clear scheme in place that delineates the different roles and responsibilities in the decision-making process involving AWS would arguably make it easier to detect where in the chain a potential breach occurred.

Approaching AWS from a trace-back perspective is a useful avenue for identifying limits on AWS. For example, an AWS with features that prevent a state from conducting an effective investigation would likely be at odds with states' IHL obligations to repress grave breaches of IHL and suppress any other violations of IHL, and so such an AWS should not be developed or used. However, this has remained a largely overlooked dimension of the AWS debate on human responsibility. More focused discussions are needed to ensure that states secure their ability to trace back conduct when needed.

II. Recommendations

Further elaborate on what the rules of IHL demand, permit and prohibit in the development and use of AWS, and how the responsibility to implement these rules is to be allocated across multiple individuals

The GGE's further deliberations on *how* the norms of IHL should be respected will be a critical first step to discerning what, and whose, actions or omissions would engage state responsibility for an internationally wrongful act and individual criminal responsibility for a war crime. In practice, that means having structured and focused discussions around what IHL rules—fundamental as well as facilitative—demand from the different actors involved in the development and use of AWS. Such an exercise should be mindful of the fact that AWS could redistribute roles and responsibilities for IHL compliance across multiple dimensions: *who* may take *what* decisions *when* and *where*, and *how* these decisions may interact with one another. Clarifying the responsibility scheme within and beyond the chain of command and control would facilitate the task of preventing, investigating and punishing IHL violations by generating a greater understanding of: (*a*) what is demanded from the different actors involved in the development and use of AWS, including what they should reasonably foresee about the behaviour and effects of AWS; (*b*) how the decisions and behaviours of the different actors involved in the development and use of AWS interact; and (*c*) what and whose conduct triggers state responsibility and individual criminal responsibility for war crimes.

Such clarifications are all the more important as the concept of a responsible human chain of command and control is considered by many states and experts pivotal for the delineation of limits on autonomy in weapon systems—and possibly the definition of a threshold for a prohibition on AWS. Such discussions would provide a comprehensive understanding of what upholding respect for IHL in the development and use of AWS means—in terms of both compliance with IHL (forward-looking responsibility) and accountability for IHL violations (backwards-looking responsibility)—and therefore also provide concrete baselines for determining what types of AWS, or AWS use-cases, could be covered by a dedicated regulation.

Share information and exchange views about national practices that can foster respect for IHL and help trace back IHL violations involving AWS

A practical way to support respect for IHL is to share information and exchange views on national practices in place to implement IHL obligations, including the obligations to investigate and punish IHL violations.

Such an exercise would entail sharing information and discussing processes and procedures for not only the legal review of new weapons, means and methods of warfare but also the provision of legal advice and legal training to the armed forces. States could also elaborate on how they currently ensure compliance with IHL at a systemic level, for instance by elaborating on the roles and responsibilities of the different actors that may be involved in the decision to use an AWS—from the strategic to the operational and tactical levels.

Such exchanges and elaboration could lead to the production of documentation that would specify roles and responsibilities, provide specific guidance to the different actors involved in the development and use of AWS, and also create 'decision logs' that record who took what decisions, when and where. These documents would also significantly facilitate the investigation and punishment of IHL violations involving AWS. In relation to the measures related to tracing back conduct, states could usefully share information about the types of investigation processes they have developed internally to discern, scrutinize and attribute responsibility. In doing so, particular attention should be paid to the unique characteristics of AWS and how these affect the ability to conduct effective investigations.

Information-sharing could work as a virtuous circle in several regards. It primarily provides an opportunity for states to demonstrate compliance with IHL, as well as an important baseline to identify potential elements of best practices which could inspire other states to develop or refine their approach to due diligence obligations. It would also provide the empirical foundation for elaborating the normative and operational frameworks governing AWS. The outcome of discussions around existing practices could help states articulate guiding principles and generate language that could inform a future policy outcome.

Identify limits and requirements in the development and use of AWS that could help ensure human responsibility in practice

In-depth analysis of the existing rules and mechanisms governing state responsibility and individual criminal responsibility provides a useful lens through which states can identify potential limits or requirements for the development and use of AWS.

In terms of limits, states could seek to identify, conceptually and technically, what features or standards would make AWS indiscriminate by nature or would make it potentially difficult to use AWS in compliance with the principles of distinction,

proportionality and precautions in attack. States could also seek to recognize the technical features of AWS that would pose challenges to the task of tracing back responsibility for IHL violations to humans. The identification of such technical features could help more clearly delineate the contours of a two-track regulation on AWS, one that, as suggested by a number states at the CCW, would prohibit certain types of AWS on the one hand and regulate the development and use of all others on the other.

In terms of requirements, states could also seek to define the standards of intent, knowledge, behaviour and care that are demanded from the different actors involved in the development and use of AWS. This process would require adopting a holistic approach to the issue of human–machine interaction, elaborating on standards of intent and knowledge required in decisions taken at the critical junctures in the development and use of AWS, and assessing how decisions across the life cycle of an AWS depend and rely on each other. Such deliberation would be critical to fleshing out what the concept of a responsible human chain of command and control would entail in the context of AWS, including what would constitute a model of 'responsible reliance'—that is, how to ensure that the people involved in the development and use of AWS are enabled to rely on decisions and assessments made at earlier junctures in the decision-making chain and to safely assume that these assessments and decisions were made in good faith. Such an exercise could not only generate concrete recommendations for the development and use of AWS, but also facilitate the task of discerning, scrutinizing and attributing IHL violations in the development and use of AWS.

About the authors

Dr Marta Bo (Italy) is an Associate Senior Researcher within SIPRI's Armament and Disarmament research area. Marta is also a Researcher at the University of Amsterdam-Asser Institute in The Hague and Research Fellow at Graduate Institute Geneva where in July 2022 she completed a four-year research project on autonomous weapon systems and war crimes. Her research focuses on state responsibility and individual criminal responsibility for unlawful conducts in the use and development of autonomous weapon systems; war crimes; AI and criminal responsibility; automation biases and mens rea; disarmament and criminalisation. Marta leads, designs and implements capacity-building training programmes for judges and prosecutors in international and transnational criminal law, international humanitarian law, and human rights law. She has published on international and transnational criminal law, the International Criminal Court and the principle of complementarity, law of the sea and human rights, artificial intelligence and criminal responsibility, autonomous weapons, self-driving cars. Marta is a Member of the Steering Committee of the Antonio Cassese Initiative for Justice, Peace and Humanity and editor of the international criminal law section of the Leiden Journal of International Law.

Laura Bruun (Denmark) is a Researcher at SIPRI, working on emerging military technologies. Her focus is on how emerging military technologies, notably autonomous weapon systems (AWS) and military applications of artificial intelligence, affect compliance with—and interpretation of—international humanitarian law (IHL). Laura was a co-author of a recent SIPRI publication, Autonomous Weapon Systems and International Law: Identifying Limits and the Required Type and Degree of Human-Machine Interaction, SIPRI Report (2021). Before joining SIPRI in 2020, Laura worked for Airwars, an NGO based in London, where she monitored and assessed civilian casualty reports from US and Russian airstrikes in Syria and Iraq.

Dr Vincent Boulanin (France/Sweden) is a Senior Researcher leading SIPRI's research on emerging military and security technologies. His focus is on issues related to the development, use and control of autonomy in weapon systems and the military applications of artificial intelligence (AI). He regularly presents his work to and engages with governments, United Nations bodies, international organizations, research institutes and the media. Before joining SIPRI in 2014, Boulanin completed a doctorate in political science at the École des Hautes Études en Sciences Sociales [School of Advanced Studies in the Social Sciences], Paris. His recent publications include Autononous Weapon Systems and and International Law: Identifying Limits and the Required Type and Degree of Human-Machine Interaction, SIPRI Report, (2021, co-author); Artificial Intelligence, Strategic Stability and Nuclear Risk, SIPRI Report (2020, co-author); and Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control, SIPRI/ICRC Report (2020, co-author).

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