Who Acts When Autonomous Weapons Strike?

The Act Requirement for Individual Criminal Responsibility and State Responsibility

Paola Gaeta*

Abstract

This essay examines the theories according to which 'actions' carried out by autonomous weapon systems enabled by strong artificial intelligence in detecting, tracking and engaging with the target ('intelligent AWS') may be seen as an 'act' of the weapon system for the purpose of legal responsibility. The essay focuses on the material act required for the commission of war crimes related to prohibited attacks in warfare. After briefly presenting the various conceptions of the act as an essential component of the material element of criminal offences, it argues that the material act of war crimes related to prohibited attacks is invariably carried out by the user of an 'intelligent AWS'. This also holds true in the case of so-called 'unintended engagements' during the course of a military attack carried out with an intelligent AWS. The essay moves on to examine the question of whether, in the case of the use of intelligent AWS by the armed forces of a state, the 'actions' of intelligent AWS — including those not intended by the user — are attributable to the state. It demonstrates that under a correct understanding of the concept of 'act of state' for the purpose of attributing state responsibility under international law, such attribution is unquestionable. It underlines that, suggesting otherwise, would bring to a breaking point the possibility of establishing violations by states of international humanitarian law in the conduct of hostilities.

* Professor of International Law, the Graduate Institute of International and Development Studies, Geneva (Switzerland). This essay is published as part of the research project that I have led, entitled 'Lethal Autonomous Weapon Systems and War Crimes: Who Is To Bear Criminal Responsibility for Commission?' (Project 10001C_176435), and funded by the Swiss National Science Foundation (SNSF). I would like to thank Guido Acquaviva, Claus Kreß and Thomas Weigend for providing detailed and prompt feedback on an advanced draft of this paper. [paola.gaeta@graduateinstitute.ch]

Journal of International Criminal Justice (2024), 1 of 23 https://doi.org/10.1093/jicj/mqae001

1. Introduction

Much has been written on the issue of the responsibility gap associated with the development of autonomous weapons systems (AWS). The debate, though presented in general terms,[1] more specifically concerns weapon systems that are enabled by 'strong' artificial intelligence[2] intended for targeting, i.e. detection, tracking and engaging with the target (hereinafter 'intelligent AWS'). These weapon systems, once activated, operate (or can operate) without the supervision or control of the user in performing their assigned tasks and functions. Weapon systems of this type are still at an early stage of development, given the difficulty of ensuring that they can be used in a manner compliant with the relevant rules of international humanitarian law.[3] Due to the specific characteristics of the algorithms, based on self-learning methods, the way in which the system performs its assigned tasks and functions cannot be fully predicted by the programmer or user. Intelligent AWS that present a high risk of unpredictability in the execution of crucial functions in the targeting cycle could therefore be indiscriminate, and thus prohibited by international humanitarian law.[4] For instance, their use might not guarantee compliance with the principle of distinction, which in the international

1 There are those who correctly point out that in the debate on the risks of responsibility gaps arising from the use of AWS, generalizations should be avoided given the different technological characteristics of AWS (see for instance N.G. Wood, 'Autonomous Weapon Systems and Responsibility Gaps: A Taxonomy', 25 *Ethics and Information Technology* (2023) 16.)

2 Strong artificial intelligence is defined as that based on algorithms using machine learning and other data-driven learning methods. For a short description of these methods and for further reference, see the A. Greipl's contribution to this issue of the *Journal*. See also amplius S.-S. Hua, 'Machine Learning Weapons and International Humanitarian Law: Rethinking Meaningful Human Control', 51 *Georgetown Journal of International Law* (2019) 117.

3 See for example Y. Dinstein, 'Autonomous Weapons and International Humanitarian Law', in W. Heintschel von Heinegg, R. Frau, and T. Singer (eds), *Dehumanization of Warfare: Legal Implications of New Weapon Technologies* (Springer, 2018) 15, at 17–20. On the international regulation of prohibited means of warfare, see the essay of A. Cassese, 'Means of Warfare: The Traditional and the New Law', reprinted in the Anthology of this special issue of the *Journal*, in particular where the author presents both the progress and the failures on the matter of the negotiations that led to the adoption of the 1977 First Additional Protocol to the 1949 Geneva Conventions.

4 The term international humanitarian law is used to encompass both the so-called Hague Law, namely the rules of international law regulating the conduct of hostilities, and the so-called Geneva Law, concerning instead the protection of persons and objects in the hands of the enemy party to the armed conflict. On the legal significance of this distinction, see R. Kolb, 'Of Hague Law and Geneva Law', *Articles of War—Lieber Institute, West Point* (published online on 29 November 2023), available online at https://lieber.westpoint.edu/of-hague-law-geneva-law/ (visited 30 December 2023). For the purpose of this essay, the term international humanitarian law is however used also to describe rules related to the conduct of hostilities, in particular those concerning prohibited military attacks, which traditionally belong to the so-called Hague Law.

On the prohibition of weapons of indiscriminate nature in international humanitarian law, see among others A. Cassese, 'The Prohibition of Indiscriminate Means of Warfare', in R.J. Akkermann et al. (eds), *Declaration of Principles: A Quest for Universal Peace* (Leyden, 1977) 171. The essay in question was also republished in P. Gaeta and S. Zappalà (eds), *The Human Dimension of International Law—Antonio Cassese: Selected Papers* (Oxford, 2008) 172.

humanitarian rules relating to the conduct of hostilities requires that the parties to an armed conflict shall at all times distinguish between civilians and combatants, as well as the observance of other relevant rules on prohibited attacks.[5]

Currently, weapon systems enabled by 'strong' artificial intelligence include certain types of loitering munitions (also known as 'suicide or kamikaze drones'), which have been used in many recent and ongoing conflicts.[6] At present, further developments in the area of intelligent AWS cannot be ruled out, and these extend the scope and possibilities of systems like intelligent loitering munitions.

Generally speaking, the difficulties that may arise with regard to the attribution of responsibility for harm caused by systems enabled by strong artificial intelligence are manifold and concern many fields of law, due to the progressive development of such systems in numerous fields of human activity.[7] The crux of the matter is the impossibility of assigning responsibility to the programmer or user, whether culpable or malicious, arising from harm caused by such systems, given the inherent unpredictability of the way in which the system performs the function and task assigned to it.[8] As far as the arms

5 See W.H. Boothby, 'Highly Automated and Autonomous Technologies', in W.H. Boothby (ed.), *New Technologies and the Law in War and Peace* (Cambridge University Press, 2019) 137, at 146. The author correctly points out that a 'weapon system would not be indiscriminate by nature if its intended circumstances of use were to be limited to areas where, and times during which, relevant civilian objects are known to be absent' (*ibid.*, footnote 26); this would be so even if the technology enabling such weapon system 'is found, during tests, … to be incapable of differentiating between civilian and military objects with the result that it would attack either of them without distinguishing them' (at 146, in the text).

In the description of the principle of distinction in the text above, the term 'combatant' is used to include members of non-state armed groups who do not enjoy the status of combatants under international humanitarian law.

6 I. Bode and T.F.A Watts, 'Loitering Munitions: Flagging an Urgent Need for Legally Binding Rules for Autonomy in Weapon Systems', *Humanitarian Law and Policy*, 29 June 2023, available online at https://blogs.icrc.org/law-and-policy/2023/06/29/loitering-munitions-legally-binding-rules-autonomy-weapon-systems/ (visited 30 December 2023). These munitions were initially conceived and used to search out and destroy radar systems within a limited period of time and in predetermined areas. However, technology has evolved. For example, loitering munitions are currently in use that, within a limited time and space, can search out and destroy various types of objects, some of which may not be military objectives by nature, and anti-personnel loitering munitions (also usable in densely populated areas).

7 One thinks for instance of the development of autonomous ships, autonomous space vehicles, autonomous surgical robots, robot dogs performing rescue and relief operations, just to mention a few examples. The debate on the so-called 'responsibility gap' arising from the development and use of intelligent technological systems thus pervades various branches of law.

8 In a seminal paper of 2004, Andres Matthias was the first to draw our attention to the challenges of ascribing responsibility for harm caused by systems enabled by learning automata. He noticed that learning automata are specifically designed to achieve a result by learning through interaction with the environment. Inevitably, therefore, the machines empowered by these automata will 'have to make "mistakes" during operation'. A. Matthias, 'The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata', 6 *Ethics Information Technology* (2004) 175–183. These mistakes, if they cause unlawful harm or infringe protected values cannot therefore be traced back by their developers or users. This

industry is concerned, the issue has been addressed mainly with respect to the responsibility gap that would arise from so-called 'unintended engagements' in armed conflict,[9] namely from military attacks carried out with AWS resulting in harm to persons or objects not intended by the human operator but caused by a failure of the system.[10]

Here the debate centres primarily on the possibility of criminal responsibility gaps in respect of war crimes, which, as is well known, consist of serious violations of international humanitarian law. Such crimes can only be committed during an armed conflict and must present a nexus with the conflict. In principle, therefore, if an intelligent AWS does not operate in compliance with international humanitarian law because of a failure of the system, the programmer of the AWS cannot be responsible for war crimes. The programmer's activities usually take place in peacetime and are not linked to a specific armed conflict in which the weapon system could be used.[11] A

creates a dilemma: to accept the creation and use of these intelligent machines and also accept the fact that there are no persons responsible for the harmful 'mistakes' caused by these intelligent machines; or to renounce the development and use of these intelligent machines, and thus also the benefits they are able to bring to our societies.

The dilemma presented by Matthias, as mentioned earlier, affects many areas of human activity and branches of law due to the progressive and pervasive presence of artificial intelligence systems in post-modern societies.

9 The term 'unintended engagement' is used by the U.S. Department of Defense to describe the risk of AWS selecting and striking targets other than those intended by the operator, resulting in fratricide, civilian casualties, or unintended escalation in a crisis. See P. Scharre, *Report on Autonomous Weapons and Operational Risk*, Center for a New American Security (2016), available online at https://www.cnas.org/publications/reports/autonomous-weapons-and-operational-risk (visited 30 December 2023), at 18. The author rightly notes that although the risk of AWS failures causing unintended engagement is low, qualitatively it is very high considering the 'aggregate damage potential resulting from potentially multiple systems falling victim to the same failure mode at the same time' (at 23). The aggregate damage potential of AWS is related to the fact that 'a software flaw in any one system is likely to be replicated across all other identical systems' (at 19). Because of this, 'the consequences of a fleet of autonomous weapons failing in a manner that led to multiple unintended engagements could be catastrophic' (at 23).

10 For an overview of the different causes of failures of artificial intelligence-enabled systems, see J. Kwik, *Lawfully Using Autonomous Weapon Technologies: A Theoretical and OperationalPerspective,* PhD thesis, University of Amsterdam (on file with the present author) Chapter 5. Briefly, failures of artificial intelligence-enabled systems are due to a variety of factors that include component failures (training data issues, input data issues, out of distribution, drift and bias) and system failures (not reduceable to one component of the system).

11 This is true in general with respect to any weapon system, not only with respect to very technologically advanced systems such as intelligent AWS. For the view according to which programmers of AWS may be held responsible for war crimes, see M. Bo, 'Are Programmers In or "Out of" Control? The Individual Criminal Responsibility of Programmers of Autonomous Weapons and Self-driving Cars', in S. Gless, and H. Whalen-Bridge (eds), *Human-Robot Interaction in Law and its Narratives: Legal Blame, Procedure, and Criminal Law* (Cambridge University Press, forthcoming 2024). The author assumes that programmers of algorithms enabling AWS partly decide how compliance with the rules of international humanitarian law is embedded in the algorithm and how AWS might interact with their environment after activation. As a result, programmers can exercise a significant level of control over AWS,

potential avenue for the criminal responsibility of the programmer when there is a failure of the system during use of an intelligent AWS in hostilities, might be to 'individualise' the obligation enshrined in Article 36 of the 1977 First Additional Protocol to the 1949 Geneva Conventions on the protection of the victims of warfare (hereinafter 'First Additional Protocol'). That article states that '[i]n the study, development, acquisition or adoption of a new weapon, means or method of warfare', States Parties to the First Additional Protocol have an obligation to determine 'whether its employment would, in some or all circumstances, be prohibited by [the] Protocol or by any other rule of international law applicable to the High Contracting Party'. If one conceptualizes the obligation under Article 36 as an obligation that is also addressed to individuals involved in the process of studying, developing, acquiring, or adopting a new weapon, its serious violation could give rise to their criminal responsibility for the failure to determine whether the new weapon is illegal.

Leaving aside this hypothesis, which remains to be verified and explored, the problem of the responsibility gap with respect to war crimes due to unintended engagements of intelligent AWS has always focused on the user (be it the operator or the military commander who decides on the use of the weapon) for the war crimes related to prohibited attacks. The debate has focused particularly on the *mens rea* required for the commission of such crimes.[12] However, given the nature and degree of autonomy of intelligent AWS, the question also arises as to whether prohibited attacks caused by a failure of the system can be considered as the act of the user for establishing the *actus reus* of a war crime. A similar issue arises if one moves up to the level of the international responsibility of the state party to the armed conflict using the intelligent AWS. Under the default regime, state responsibility under international law arises if the state has committed an internationally wrongful act. In turn, this requires that an act of the state is in breach of an international obligation incumbent on that state. Since there exists such an act of the state in cases where the conduct of a person or group of persons is attributable to that state, could one potentially attribute to the state prohibited attacks caused by a failure of the system and not by a human conduct?

This article will approach these two issues in turn.

---

control that continues even when the systems are used by the end user. According to the author, by virtue of this control, programmers could be responsible for war crimes if they could understand and foresee the risk of a crime committed with autonomous systems technology.

12 Basically, it has been observed that when the rules on war crimes require the (direct or indirect) intent of the agent, it would not be possible to hold the latter criminally responsible for the absence of the cognitive element due to the impossibility of foreseeing all the possible consequences of the use of such intelligent AWS. On this point see more extensively M. Bo, 'Autonomous Weapons and the Responsibility Gap in the Light of the *Mens Rea* of War Crime of Attacking Civilians in the ICC Statute', in 19 *Journal of International Criminal Justice* (2021) 275.

2. 'Actions' of the Intelligent AWS and the Material Act of War Crimes

It is well known that, in modern criminal systems, the essential constituent elements of the offence are: the external, or objective element, called the *actus reus* in the Anglo-American tradition; and the subjective element, i.e. the mental state of the potentially responsible subject required by the criminal law. According to the conventional theory, the *actus reus* consists of an essential component, i.e. the act required to constitute the offence (and this act can also be an omission) and it may also require circumstances attending the act and the result of the act. This apparently simple assertion actually conceals very complex issues, as demonstrated by the long-standing debate in criminal doctrine on the general theory of the offence and, as far as we are concerned, on what is to be understood by an act.[13]

Without going into the details of this intricate debate, it is worth exploring a well-established aspect, or assumption, of it: that modern criminal law systems and concepts of criminal responsibility are built around human actions and volition.[14] This is axiomatic with respect to the criminal responsibility of natural persons. However, it is also indirectly true in the case of the criminal responsibility of legal persons. The theoretical explanations put forward for the responsibility of the latter are in fact based on the imputability of acts and volitions of natural persons acting on behalf of the entity to the latter, or alternatively, for the criminogenic or improper organization of the entity, which is, in any case, the result of human acts and volitions that created it.[15]

13 For an excellent survey of the various theories proposed, see J. Keiler, 'Actus Reus', in P. Caeiro, S. Gless, and V. Mitsilegas (eds), *Elgar Encyclopedia of Crime and Criminal Justice* (Elgar online: visited 20 December 2023; forthcoming in hardback in 2024).

14 In the late Middle Ages up to the early modern period, cases of trials of animals for crimes against humans have been documented: see W.W. Hyde, 'The Prosecution and Punishment of Animals and Lifeless Things in the Middle Ages and Modern Times', in 64 *University of Pennsylvania Law Review and American Law Register* (1916) 696. More recently, see P. Dinzelbacher, 'Animal Trials: A Multidisciplinary Approach', in 32 *The Journal of Interdisciplinary History* (2002) 405.

15 On this point, see among others T. Weigend's contribution to this special issue of the *Journal*.

Regarding the justification that the personhood of legal entities is linked to the circumstance that they are composed of natural persons, one may cite an interesting case concerning a petition for writ of habeas corpus filed before a US court by *Nonhuman Rights Project Inc.* for two chimpanzees (Hercules and Leo). In dismissing the petition, the judge noted that the two chimpanzees do not possess attributes sufficient to establish legal personhood. He recognized that legal personhood does not have to be synonymous with human, as in the case of corporations and partnerships that have been deemed persons for certain purposes. He noted, however, quoting with approval an *amicus curiae*, that 'these corporations are still composed of humans'. (In *Matter of Nonhuman Rights Project Inc. v. Stanley*, 29 July 2015, 2015 NY Slip Op 25257 [49 Misc 3d 746], also available at https://www.nycourts.gov/REPORTER/3dseries/2015/2015_25257.htm (visited 30 December 2023))

A. Criminal Theories of the Act and Action of Artificial Intelligence Systems

In doctrine, there are those who theorize that artificial intelligence systems can be considered criminally responsible subjects,[16] rejecting the thesis that is expressed in the maxim *machina delinquere (et puniri) non potest* (machines cannot commit crimes and cannot be punished).[17] Regarding the *actus reus*, this doctrine does not hesitate in asserting that the tasks accomplished by such systems are comparable to a human act for the purposes of criminal responsibility.[18] The premise underlying such assertions is a purely materialistic conception of the criminal act, which disregards any connection with its voluntariness. Indeed, it is argued that for the purposes of criminal responsibility, the act is simply the 'material performance through factual-external presentation, whether willed or not'.[19] Accordingly, it is affirmed that 'artificial intelligence technology is capable of performing "acts", which satisfy the conduct requirement' and that '[t]his is true not only for strong artificial intelligence technology, but for much lower technologies as well'.[20] Therefore:

> [w]hen a machine, (e.g., robot equipped with artificial intelligence technology) moves its hydraulic arms or other devices of its, it is considered an act. That is correct when the movement is a result of inner calculations of the machine, but not only then. Even if the machine is fully operated by human operator through remote control, any movement of the machine is considered an act.[21]

The materialistic conception of the act stands in antithesis to the traditional theory according to which the criminally relevant act is that of bodily movement or non-movement based on the will of the person, also understood as an expression of the freedom of self-determination of that person and of their lordship over their body.[22] Those who follow this traditional approach categorically deny that artificial intelligence systems, including those capable of acting in physical space based on self-learning algorithms, can conduct a criminally relevant act.[23] This is because their actions would still originate from a self-learning algorithm established by the programmer and developer,

16 See, in particular, G. Hallevy, *Responsibility for Crimes Involving Artificial Intelligence Systems* (Springer, 2015). See also F. Lagioia and G. Sartor, 'AI Systems under Criminal Law: a Legal Analysis and a Regulatory Perspective', 33 *Philosophy & Technology* (2020) 433. For a critical assessment of this theory, see among others T. Weigend's contribution to this special issue of the *Journal*.

17 This Latin maxim is a paraphrase of the well-known maxim used to argue that legal persons cannot be criminally responsible (*societas delinquere (et puniri) non potest*). On this point, see for example A. Cappellini, 'Machina Delinquere Non Potest? Brevi Appunti su Intelligenza Artificiale e Responsabilità Penale', disCrimen, 27 March 2019.

18 See Hallevy, *supra* note 16, at 60–63; Lagioia and Sartor, *supra* note 16, at 439–441.

19 See Hallevy, *supra* note 16, at 61.

20 *Ibid.*

21 *Ibid.*

22 See for all M.S. Moore, *Act and Crime: The Philosophy of Action and its Implications for Criminal Law* (Oxford University Press, 2010).

23 See for instance C. Piergallini, 'Intelligenza artificiale: da "mezzo" a "autore" del reato', *Rivista italiana di diritto e procedura penale* (2020) 1745, at 1766–1767.

who would therefore remain the 'real' drivers of the action.[24] The unpredictability in the choices and actions of the intelligent system are therefore understood as an unpredictability necessitated by this algorithm and not a manifestation of intelligent action. In other words, an artificial intelligence system 'does not act, but it is acted upon'.[25]

The impossibility of considering the actions of artificial intelligence systems as acts in the sense of criminal law is also reached if one adheres to the theory that an act is any conduct that has an impact in the social sphere.[26] As has rightly been observed, current artificial intelligence systems 'are still too young to have gathered ... [the] "critical mass" of social meaning and importance' necessary for the recognition of their own and independent action.[27]

In any case, the possibility of considering the action of the artificial intelligence system as an act of the system itself — an essential prerequisite for making the system subject to criminal law — seems to be confined only to theoretical debate. As things stand, when there is a fact of criminal relevance resulting from the action of intelligent systems, it is the user's responsibility that is at stake — at least in situations where the user is required to exercise direct supervision over the system's operation. For example, in the United States, the driver of a Tesla travelling in AutoPilot mode was convicted for the car accident that occurred in 2019 in a suburb of Los Angeles causing the death of two people in a Honda Civic.[28] This is the first prosecution in the United States for vehicular homicide caused by a car travelling autonomously. It seems, however, that this circumstance was not relevant in determining the criminal responsibility of the Tesla driver, precisely because the driver is still required to be vigilant, and the AutoPilot self-driving system is marketed by Tesla as a driver assistance system.[29] If one considers that AutoPilot

24 *Ibid.*, at 1767.

25 *Ibid.*, at 1769. See also D. Lima, 'Could AI Agents Be Held Criminally Responsible? Artificial Intelligence and the Challenges for Criminal Law', 69 *South Carolina Law Review* (2018) 677, at 682.

26 Act theory inspired by this conception is currently prevalent in countries such as Germany and the Netherlands: see Keiler, *supra* note 13.

27 See Lima, *supra* note 25.

28 N. Percy, 'Driver of Tesla on Autopilot Gets Probation for Crash that Killed 2 in Gardena', *Daily Breeze*, 30 June 2023, available online at https://www.dailybreeze.com/2023/06/30/driver-of-tesla-on-autopilot-gets-probation-for-crash-that-killed-2-in-gardena/ (visited 30 December 2023). The car was travelling in semi-autonomous mode (thanks to the Autopilot system) and the on-board software had crossed a red light at high speed, colliding with the Honda Civic, killing the two people on board.

Due to the progressive introduction of autonomous and semi-autonomous self-driving cars, criminal prosecutions of road homicides similar to the one just reported can be expected.

29 Autopilot is proposed by Tesla as a driver assistance system integrated into cars; as stated on the Tesla website, it requires active driver supervision and does not make the vehicle autonomous: see https://www.tesla.com/support/autopilot (visited 30 December 2023). In August 2021, the US National Highway Traffic Safety Administration began an investigation after a series of car accidents, some fatal, involving Autopilot. On 12 December 2023, the Administration issued a recall finding that there may be increased risk of a crash when Autosteer is engaged and drivers do not maintain responsibility for vehicle operation. Autosteer is functionality that, according to Tesla, keeps Model S in its driving lane when

technology is also the technology present in some types of autonomous weapon systems currently in use (e.g., *loitering munitions* such as Switchblade drones), one realizes that the use of the technology in question is not necessarily 'safer', but merely more convenient for the user (at least until something goes wrong).[30]

B. Unintended Engagements of Intelligent AWS and War Crimes Related to Prohibited Attacks

Excluding the thesis that the action of the intelligent system is an act of the system itself for criminal law purposes, one can now examine whether unintended engagements of the AWS resulting in prohibited attacks are acts of the user of the weapon system for the purpose of the *actus reus* of the relevant war crimes.

To this end, a few clarifications are in order. First, precisely because we are discussing the use of weapons in the conduct of hostilities, and assuming that these weapons do not by their nature operate indiscriminately, the potentially relevant war crimes are those relating to prohibited attacks. These crimes penalize attacks that are unlawful under international humanitarian law, such as attacking civilians (in violation of the principle of distinction) or attacking a military target causing civilian collateral damage disproportionate to the anticipated military advantage (in violation of the so-called principle of proportionality).[31] The *actus reus* of these war crimes

cruising at a set speed. The recall presses Tesla to make updates to ensure drivers are paying attention while using Autopilot (J. Ewing, C. Metz, and D. Bryson Taylor, 'Tesla Recalls Autopilot Software in 2 Million Vehicles', *The New York Times (Digital Edition)*, 13 December 2023.)

30 On the subject of Autopilot and its use in cars for driver assistance, Philip Koopman's recent statement for the Subcommittee on Innovation, Data, and Commerce (IDC) of the US House of Representatives is illuminating, particularly when he observes: 'Cars that mostly drive while providing insufficient driver monitoring and attention management are already causing injury and fatality crashes on our roads. So-called autopilot and related features are only convenience features. Automated steering control is not a safety feature, despite the marketing messaging you might have heard … . With regard to truly driverless vehicles, they have the potential–and I emphasise the word potential–to improve lives and expand mobility options, but only if designed and deployed carefully. That potential can only be realised after they are safe and responsible. Right now the technology is irresponsible, as news stories from San Francisco and other places tell us on a weekly basis. Nobody really knows yet how safe they will turn out to be, due to fundamental limitations of the technology. Most importantly, the 'AI' used in these vehicles is good at things it has been taught, and bad at surprises. But the real world is so full of surprises, we don't know if we can teach the computers enough to come out acceptably safe in the end.' (Written Testimony of Dr Philip Koopman IDC Subcommittee Legislative Hearing, 26 July 2023, available at http://users.ece.cmu.edu/~koopman/pubs/Koopman2023_EC_Testimony_AV_Safety.pdf (visited 30 December 2023)).

31 The list of war crimes related to prohibited attacks contained in Art. 8(2)(b) ICCSt., includes, for international armed conflicts, the following: attacks on civilians and civilian objects; attacks on humanitarian or peacekeeping personnel or objects; attacks on undefended places; attacks on buildings dedicated to religion, education, art, science, or charitable purposes or of

is differently formulated in the relevant international instruments. In particular, these crimes are formulated as crimes of conduct in Article 8 of the Rome Statute establishing the International Criminal Court,[32] while Article 85(3) of the First Additional Protocol to the Geneva Conventions includes the occurrence of a harmful event (causing death or serious injury to body or health).[33] Leaving aside this important difference,[34] what is common to all descriptions of the *actus reus* of crimes of unlawful attacks is the act itself (directing an attack/making an attack/launching an attack)[35] and the need for certain circumstances to be present. One circumstance common to all war crimes, as is well known, is that the act was committed in the course of an armed conflict and is associated with it (the so-called 'nexus').[36] The other

historic monuments; attacks on hospitals or places where sick and wounded are collected; attacks on persons or objects using distinctive emblems; attacks on persons *hors de combat*; attacks that cause excessive incidental death, injury or damage. The list of war crimes of prohibited attacks in non-international armed conflicts contained in Art. 8(2)(e) ICCSt. is much shorter. For example, it does not include attacks on civilian objects and attacks that cause excessive incidental death, injury or damage.

32 See for example Art. 8(2)(b)(i), which provides that the following crime is a war crime within the jurisdiction of the ICC: 'Intentionally directing attacks against civilians not taking direct part in hostilities'. The *actus reus* of this crime consists only of the act (directing attacks) and the circumstance that it is 'against civilians not taking part in hostilities'. An event, i.e. the death or injury of the civilians attacked, is not required (as in Art. 85(3) of the First Additional Protocol: see below, note 33). In the discussion of the Elements of Crimes, it was debated whether to include the event in the description of the *actus reus*, but it was decided not to do so following what had emerged from the negotiations of Art. 8 of the Statute during the Rome Conference. On this point see K. Dörmann (with contributions by L. Doswald-Beck and R. Kolb), *Elements of War Crimes under the Rome Statute of the International Criminal Court* (Cambridge University Press, 2003), at 130.

33 Art. 85(3) of the First Additional Protocol, in listing war crimes consisting of grave breaches of the First Protocol (which include crimes of prohibited attacks, including attacking civilians) states that: '… the following acts shall be regarded as grave breaches of this Protocol, when committed wilfully, in violation of the relevant provisions of this Protocol, and *causing death or serious injury to body or health* …'. (Emphasis added).

34 The material element of the offence, which requires the event, obviously implies that the causality between the act and the unlawful event be proven. This is not necessary in the case of conduct crimes. As has rightly been observed, whether or not the material element of the war crime requires the occurrence of an event (e.g. killing and wounding of civilians) is not a minor difference, if one takes into account the burden of proof on the investigators. Indeed, 'the need to prove beyond reasonable doubt the death or injury of the civilians targeted in the attacks as a consequence of the attacks … inevitably makes the task of the investigating authorities more complex and, depending on the case, may entail a significant modification of the investigative and prosecutorial strategy' (G. Acquaviva, *La repressione dei crimini di guerra nel diritto internazionale e nel diritto italiano* (Giuffrè, 2014), at 118 (original in Italian, translation is by this author).

35 These formulations of the required act are by no means identical. In particular, while the term 'directing attacks' seems to include all steps of the decision-making process of the attack, the terms 'launching an attack' or 'making an attack' seems limited to the actual use of violence. For these different formulations, see the terms used in Art. 8(2)(b) and (e) ICCSt. and the terms used in Art. 85(3) of the First Additional Protocol.

36 For all, see G. Mettraux, 'Nexus with Armed Conflict', in A. Cassese (ed.), *The Oxford Companion to International Criminal Justice* (Oxford University Press, 2009) 435.

circumstances vary depending on the crime. In most cases, it is required that the objects of the attack are persons that in international humanitarian law enjoy immunity from military attack (such as civilians), in keeping with the principle of distinction. A more complex formulation is the one that requires the attack against a military target to have caused death or casualties among civilians, or damage to the natural environment, disproportionate to the expected military advantage. The latter war crime concerns attacks in violation of the principle of proportionality, which allows attacks against military targets that cause incidental harm among civilians and other persons and property immune from the attack. Military attacks in violation of this principle (i.e. attacks against military targets that cause incidental damage disproportionate to the anticipated military advantage) fall into the category of attacks of an indiscriminate nature.[37]

For criminal law purposes, when describing the material element of the offence, it is important to distinguish the act and the circumstances that may be required for its criminality because the subjective element required for the commission of the offence may differ in those regards.[38] For example, in interpreting the war crime related to the prohibition of attacking civilians (as a serious breach of the First Additional Protocol), the International Criminal Tribunal for the former Yugoslavia (ICTY) seems to have required intentionality only with respect to the act (making an attack). In contrast, it held that recklessness is sufficient with respect to the circumstance (i.e. that civilians were the object of the attack).[39] Other courts have followed this approach,[40] deciding that attacks of an indiscriminate nature other than

37 In this sense, see Art. 51(5)(b) of the First Additional Protocol, which expressly considers — by way of example — attacks that cause collateral damage disproportionate to the anticipated military advantage to be indiscriminate. ('Among others, the following types of attacks are to be considered as indiscriminate: ... b) an attack which may be expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated.') Art. 51(4) of the First Protocol prohibits indiscriminate attacks and defines them as follows: 'Indiscriminate attacks are: (a) those which are not directed at a specific military objective; (b) those which employ a method or means of combat which cannot be directed at a specific military objective; or (c) those which employ a method or means of combat the effects of which cannot be limited as required by this Protocol; and consequently, in each such case, are of a nature to strike military objectives and civilians or civilian objects without distinction.'

38 Keiler, *supra* note 13.

39 See in particular at the ICTY, Judgment, *Galić* (IT-98-29-T), Trial Chamber, 5 December 2003, § 55, which states that it is sufficient that the perpetrator 'should have been aware of the civilian status of the person attacked'. On this point, see among others G. Werle and F. Jeßberger, *Principles of International Criminal Law* (4th edn., Oxford University Press, 2020), at 534–535.

40 See for instance Court of Bosnia and Herzegovina, *Prosecutor's Office of Bosnia and Herzegovina v. Novak Dukić* (Case No. X-KR-07/394), *First Instance Verdict*, 12 June 2009 (published on 14 September 2009, available online at http://www.asser.nl/upload/documents/DomCLIC/Docs/NLP/BiH/Dukic_Novak_First_Instance_Verdict_12-06-2009.pdf (visited 30 December 2023). The International Court of Justice also seems to have expressed itself in the same sense, in its Advisory Opinion of 8 July 1996 on the *Legality of the Threat or Use of Nuclear Weapons*,

those consisting in disproportionate attacks are punishable as war crimes of directing an attack on civilians, and thus charged without intention as to making civilians the primary object of an attack.[41] In contrast, the Elements of Crimes (to be applied by the International Criminal Court) state that intentionality is required with respect to both the act (directing/launching the attack) and the circumstance (the object of the attack, e.g. civilians).[42] It thus appears that the International Criminal Court can prosecute as war crimes without intent to attack civilians (or other persons or obejects immune from attacks) only indiscriminate attacks consisting of a violation of the principle of proportionality, as they are expressly covered by its Statute if committed in the context of an international armed conflict.[43] Other attacks of an indiscriminate nature, such as an attack not specifically directed against a military target, would not seem to be prosecutable as war crimes.[44] They lack an express legal basis and cannot amount to directing an attack against civilian targets (or other persons or property immune from military attack).[45]

For our purposes, it is important to note that it is the one who employs an intelligent AWS for a military attack who performs the act (directing/making/launching an attack), even the identification, selection and engagement of the target is made by means of the algorithm embedded in the system. In the

where it stated. 'States must never make civilians the object of attack and must consequently never use weapons that are incapable of distinguishing between civilian and military targets' (§ 78).

41 Recall that for the grave breaches listed in Art. 85 of the First Additional Protocol, the event of causing death and injury is also required.

42 Dörmann, *supra* note 32, at 130–131. The author reports that this was nevertheless a controversial issue in the Preparatory Committee that drafted the Elements of Crimes.

This interpretation is confirmed by Art. 30(1) ICCSt., which states: 'Unless otherwise provided, a person shall be criminally responsible and responsible for punishment for a crime within the jurisdiction of the Court only *if the material elements* are committed with intent and knowledge.' Paragraph 3, clarifies: 'For the purposes of this article, "knowledge" means *awareness that a circumstance exists …*' (emphasis added). See also footnote 32 of the Elements of Crimes for Art. 8(2)(a)(i), with respect to paragraph 3 (The perpetrator was aware of the factual circumstances that established that protected status): footnote 32 clarifies that this subjective element applies 'to the element in other crimes in article 8(2) concerning the awareness of the factual circumstances that establish the status of persons or property protected under the relevant international law of armed conflict'.

43 As regards to non-international armed conflict, the war crime of disproportionate attack is not enumerated in the list of war crimes contained in Art. 8 ICCSt.

44 Indiscriminate attacks consisting in the use of weapons which are by nature indiscriminate is a war crime under Art. 8(2)(b)(xx) ICCSt., but it is required that such weapons 'are the subject of a comprehensive prohibition and are included in an annex' to the Rome Statute.

45 In this sense see more extensively Bo, *supra* note 12, especially at 292. The author rightly observes that, in the context of the punishment of many forms of indiscriminate attacks under the Rome Statute, 'one faces an either/or situation'. This is because: 'Either one considers that "intentionality" of the war crimes of targeting civilians includes at least the most serious forms of risk-taking behaviours (*dolus eventualis*), thus allowing the punishment of most types of indiscriminate attacks. Alternatively, one contends that the requirement of "intentionality" is intent (first and second degree), thus ruling out the criminality of most instances of indiscriminate attacks under the ICC Statute'.

event that the circumstance necessary for the criminalization of the act materializes, i.e. for the war crime of attacking civilians the fact that the latter are targeted it is of no relevance to the *actus reus* that this is a consequence of a failure of the AWS.[46] This applies to military attacks conducted with any kind of conventional weapon. If, due to a failure of the weapon, the attack does not target the intended military objective, but instead targets civilians or other persons and property protected by the military attack,[47] the existence of the material act of the offence cannot be put in question. The crucial issue will mainly concern the presence of the required subjective element with respect to the materialization of the circumstance required for the criminality of the act[48] and, more generally, the 'culpability' of the user.[49]

3. Use of Intelligent AWS as an Act of the State and an Element of an Internationally Wrongful Act

A. Attribution to the State of Unintended Engagements Caused by Intelligent AWS

Usually, the attribution to the state of unintended engagements resulting from military attacks carried out with intelligent AWS does not raise specific challenges. If members of the armed forces of a state employ a weapon system, the rule codified in Article 4 of the Articles of State Responsibility of the International Law Commission ('ILC Articles on State Responsibility') applies. For the purpose of identifying the existence of a wrongful act of the state, this rule permits attribution to the state of the acts (active or omissive) of persons

46 This is also the case for the act consisting in 'directing an attack', which is not to be interpreted as 'launching an attack with a view to/aimed at'. This is evident in the Elements of Crimes for all war crimes consisting of 'directing an attack' against a prohibited target. For these war crimes, the description of the relevant components of the actus reus is as follows: i) the perpetrator directed an attack; ii) the object of the attack was a civilian population (or other prohibited object). The former describes the act and the latter the necessary circumstance that makes the act criminally relevant. If the term 'directed' were interpreted as 'aimed at', the description in the Elements of Crimes would make little sense.

47 Weapon system malfunctions can of course also occur in autonomous defensive weapon systems. In this connection, one can mention the software problem of the US Patriot missile defence system operating at Dhahran (Saudi Arabia). On 25 February 1991, the software problem led to a system failure of the Patriot that failed to track and intercept an incoming Scud, which hit an army barrack causing the death of 28 Americans. See United States — General Accounting Office, *Report to the Chairman, Subcommittee on Investigations and Oversight, Committee on Science, Space, and Technology, House of Representatives, Patriot Missile Defense Software Problem Led to System Failure at Dhahran, Saudi Arabia*, 4 February 1992 (GAOAMTEC-92-26).

48 On the problems inherent in the identification of the subjective element of attack against civilians (and other similar war crimes of prohibited attacks) in the Rome Statute, which are evidently crucial in the case of the use of intelligent AWS and more generally the use of autonomous technology in targeting decisions, see Bo's in-depth analysis, cited above in footnote 12 and also A. Coco's contribution to this special issue of the *Journal*.

49 See A. Coco's contribution to this special issue of the *Journal*.

or groups of persons who are organs (*de jure* or de facto) of the state.[50] The armed forces of a state are typically *de jure* organs of that state. Thus, whatever the weapon employed by the armed forces, the military attack is attributable to the state. If the military attack is conducted in violation of the relevant rules of international humanitarian law, then there will be an internationally wrongful act of the state.[51] The state will therefore be internationally responsible for the unlawful act committed,[52] unless there exists a circumstance excluding unlawfulness.[53]

50 'The conduct of any State organ shall be considered an act of that State under international law, whether the organ exercises legislative, executive, judicial, or any other functions, whatever position it holds in the organization of the State, and whatever its character as an organ of the central Government or of a territorial unit of the State. An organ includes any person or entity which has that status in accordance with the internal law of the State.'

51 The assessment of whether a state has violated rules on prohibited military attacks is not easy, as it depends on militarily sensitive information that is not disclosed. On this point, see for instance the remarks of M. Sassòli, 'Israel—Hamas 2023 Symposium—Assessing the Conduct of Hostilities in Gaza Difficulties and Possible Solutions', 30 October 2023, posted on the blog, Articles of War, Lieber Institute Westpoint, available online at https://lieber.westpoint.edu/assessing-conduct-hostilities-gaza-difficulties-possible-solutions/ (visited 30 December 2023).

As to the content of the rules on prohibited military attacks, it is necessary to determine whether they require a subjective element. If so, establishing that these rules have been violated would add a further layer of complexity, as 'proof of wrongful intent or negligence is always very difficult'. This is true 'in particular, when this subjective element has to be attributed to the individual or group of individuals who acted or failed to act on behalf of a State, its research becomes uncertain and elusive.' See J. Crawford, *State Responsibility: The General Part* (Cambridge University Press, 2013), at 61. The need for a subjective element in rules on prohibited military attacks would obviously make it even more difficult to prove a violation in the case of the use of intelligent AWS that resulted in 'unintended engagements'.

In the case of unintended engagements that result in attacks against civilians or other protected persons or property, or excessive collateral damage, some argue that 'honest' and 'reasonable' mistakes in the targeting process are implicitly permitted by the targeting rules. In this sense, see for instance A.G. Jain's contribution to this special issue. See also M. Milanović, 'Mistakes of Fact When Using Lethal Force in International Law: Part I', 14 January 2020, posted on EJIL Talk!, available online at https://www.ejiltalk.org/mistakes-of-fact-when-using-lethal-force-in-international-law-part-i/ (visited 30 December 2023).

For the reasons explained in the introduction to this special issue of the *Journal*, this author considers instead that 'honest' and 'reasonable' mistakes in the targeting process do not affect the assessment of the violation of many of the relevant rules, e.g. the prohibition of attacking civilians or other protected persons or property.

52 For the existence of the internationally wrongful act of a state, the act of the state must constitute a breach of an international obligation of the state. See Art. 2, lit b, of the ILC Articles on State Responsibility.

53 In the case of the use of AWS resulting in unintended engagement in violation of the rules on prohibited military attacks, the only potential relevant circumstance precluding wrongfulness is 'force majeure', provided for in Art. 23 of the ILC Articles on State Responsibility. It seems, however, that the stringent conditions for the application of this circumstance precluding wrongfulness could not easily be met. On this point, see F. Albader, 'Exploring the Applicability of Force Majeure for AI Mistakes in Armed Conflict', in Harvard Law School—National Security Journal, 29 January 2023, available online at https://harvardnsj.org/2023/01/29/exporing-the-application-of-force-majeure-for-ai-mistakes-in-armed-conflict/ (visited 30 December 2023).

It should also be considered that the rule of attribution enshrined in Article 4 is 'reinforced' in international humanitarian law, even with regard to the responsibility of the state party to the conflict arising from the violation of rules on the conduct of hostilities. Indeed, Article 91 of the First Additional Protocol states that a party to the conflict 'shall, if the case demands, be responsible to pay compensation' and that '[i]t shall be responsible for *all acts committed by persons forming part of its armed forces*'.[54] This rule, read in conjunction with Article 7 of the ILC Articles on State Responsibility, makes it possible to attribute to the state all acts committed by persons who are members of the armed forces of a state, including acts committed *ultra vires* as an organ of the state. According to some commentators, however, this rule would go even further than Article 7 and would also allow all *ultra vires acts* committed by members of the armed forces acting in their private capacity to be attributed to the state party to the conflict.[55] The reason, as was well explained by Kalshoven, is that 'members of an armed force at war stand a greater chance than do other State organs of becoming entangled in ambiguous situations where it may be unclear whether they are acting in their capacity as an organ of the State'.[56] If one accepts this interpretation, it is clear that Article 91 of the First Additional Protocol broadens the sphere of attribution to the state of acts committed by members of the armed forces in their capacity as private individuals. This is also possible by virtue of the ILC Articles on State Responsibility, which in Article 55 provides for the applicability of *lex specialis* in respect of the responsibility of states.[57]

54 The emphasis is added. Art. 91 of the First Additional Protocol reads in full as follows: 'A Party to the conflict which violates the provisions of the [1949 Geneva] Conventions or of this Protocol shall, if the case demands, be responsible to pay compensation. It shall be responsible for all acts committed by persons forming part of its armed forces.

This article reproduces almost verbatim Art. 3 of Hague Convention IV, which reads as follows: 'A belligerent party which violates the provisions of the said Regulations shall, if the case demands, be responsible to pay compensation. It shall be responsible for all acts committed by persons forming part of its armed forces.' (Art. 3, Convention (IV) respecting the Laws and Customs of War on Land (hereinafter Hague Convention IV) and its Annex: Regulations concerning the Laws and Customs of War on Land (hereinafter Hague Regulations).There are some differences in terms of the obligation to provide compensation established in these two articles. For a very detailed analysis, see F. Kalshoven, 'State Responsibility for Warlike Acts of the Armed Forces: From Article 3 of Hague Convention IV of 1907 to Article 91 of Additional Protocol I of 1977 and beyond', 40 *International Comparative Law Quarterly* (1991) 827.

55 See Kalshoven, *supra* note 54, at 853; as regards to Article 3 of Hague Convention IV, at 837. See also M. Sassòli, 'State Responsibility for Violations of International Humanitarian Law', 846 *International Review of the Red Cross* (2002) 401, at 405–406.

56 Kalshoven, *supra* note 54, at 837, concerning Art. 3 of Hague Convention IV. The same argument is advanced with regard to Art. 91 of the First Additional Protocol, at 853.

57 Art. 55 of the ILC Articles on State Responsibility, which provides: 'These articles do 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.'

James Crawford has rightly observed that the attribution rules formulated in the ILC Articles on State Responsibility 'seem to have no rival of a general character'. Indeed, '[w]hatever the range of state obligation in international law, the ways of identifying the state for

However, in a recent study, Boutin has claimed that the attribution to the state of military attacks carried out with intelligent AWS raises particular challenges.[58] On the assumption that these systems are 'independent and endowed with a degree of autonomous agency', 'the link to any human conduct [would be] too vague and weak to ground attribution of conduct' to the state.[59] Instead, it would be possible, for the purposes of attribution, to conceptualize these systems as operating under the direction and control of the state within the meaning of Article 8 of the ILC Articles on State Responsibility.[60] This would also allow the attribution to the state of the *'ultra vires*' acts of the system within the meaning of Article 7 of the ILC Articles.[61]

The approach followed by Boutin resonates with the writings of other commentators in distinct fields. For instance, regarding the attribution to the state of violations of *jus ad bellum* through cyber-attacks, it has been argued that the 'advent of true autonomous agents could really require new interpretations or new formulations' with respect to the question of 'agency', i.e. the attribution of individual conduct to the state.[62] Accordingly, the development of autonomous agents would exacerbate the attribution issues already inherent in cyberattacks, due to the autonomy of the decisions of the systems employed, which would make command and control by the competent persons and bodies 'hard to achieve'.[63] Therefore, so the argument continues, '[t]heoretically, a true autonomous agent could exceed its assigned tasks and engage in what could legally be defined as "use of force"', and therefore one may wonder whether 'in this case, ... the nation state behind the agent's creation [should] be deemed responsible'.[64] The solution suggested is to consider the possibility of recognizing autonomous agents per se as 'state agents' for the purpose of attribution and targeting (thus adopting criteria to distinguish whether they are 'civil' or 'military': e.g.

the purposes of determining breach appear to be common'. However, there may be special cases in which 'special rules of attribution be devised', as is the case — according to Crawford — with Article 1 of the UN Convention against Torture. Indeed, the definition of torture in this article states that the infliction of pain and suffering constitutes torture when, among other things, 'such pain or suffering is inflicted by or at the instigation of or with the consent or acquiescence of a public official or other person acting in an official capacity'. See J. Crawford, 'The ILC's Articles on Responsibility of States for Internationally Wrongful Acts: A Retrospect', 96 *American Journal of International Law* (2002) 874, at 878.

58 B. Boutin, 'State Responsibility in Relation to Military Applications of Artificial Intelligence', 36 *Leiden Journal of International Law* (2023) 133, at 140.

59 *Ibid*., at 143.

60 *Ibid*.

61 *Ibid*. In fact, this statement is incorrect because Art. 7 operates only with respect to *ultra vires* actions of persons or group of persons whose conduct is attributed to the state under Arts 4–6 of the ILC Articles on State Responsibility.

62 A. Guarino, 'Autonomous Intelligent Agents in Cyber Offence', in K. Podins, J. Stinissen, and M. Maybaum (eds), *5th International Conference on Cyber Conflict* (Tallin: NATO CCD COE Publications, 2013) 377, at 385.

63 *Ibid*.

64 *Ibid*.

for software bots, 'through mandatory signatures or watermarks embedded in their codes').[65]

However, these and similar propositions are not fully convincing. Let me first briefly summarize the view put forward by Boutin. Then I will show that it is based on an incorrect understanding of the notion of an act of the state for the purposes of international responsibility accepted by the ILC Articles on State Responsibility.

B. The Alleged Need for Causality between Human Conduct and Breaches of International Law

The assertion by Boutin that the use of intelligent AWS that results in breaches of international law may render inoperative for international state responsibility the rules of attribution to a state of acts of persons to a state[66] is based on a twofold premise. The first is that the rules of attribution of a wrongful act to a state 'unequivocally [hinge] upon actions or omissions by human beings' and that 'the existence of human conduct is therefore a pre-condition for state responsibility'.[67] The second premise is that, for attribution to occur, there must be 'a causal link between acts or omissions by a human being and the occurrence of a breach of international law'.[68] Given the characteristics of systems that make use of artificial intelligence (autonomy, opacity, unpredictability), so the argument goes, it is therefore necessary to establish what human conduct is relevant for attributing to the state the wrongful act caused by the use of such systems.[69]

Accordingly, in discussing the issue of attribution in relation to the use of artificial intelligence systems in the military field, Boutin presents various scenarios to determine in which cases such a causal link exists. The first scenario is where the system operates under the direct and genuine control of a human operator at a tactical level. In this case, Boutin argues there would be no attribution problem: if the operator's conduct is attributable to the state under one of the rules in the ILC Articles of State Responsibility, the operator's action or omission would directly link the state to the occurrence of a breach of international law.[70] Likewise, for Boutin there would be no attribution problem in a second scenario, namely in the case of an AWS that, once activated, operates autonomously but the operator can 'override' the system's decision. Boutin argues that in this scenario, the human conduct relevant to attribution would not be that of the operator, since the latter would be only to a limited extent able to 'abort' the attack conducted by the system. The causal link with the occurence of a breach should instead be found in the conduct of the political and military decision-makers who

65 *Ibid.*
66 See above in the text, and accompanying notes 58–61.
67 Boutin, *supra* note 58, at 139.
68 *Ibid.*
69 *Ibid.*, at 140.
70 *Ibid.*, at 142.

authorized and established the parameters for the use of systems operating in autonomous mode once activated. This in turn would allow the breach of international law resulting from the use of these almost fully autonomous systems to be attributed to the state.[71] According to Boutin, there would also be no attribution problems in a third scenario, which is when an autonomous system merely aids decision-making by a human operator. In this scenario, the system provides information and/or makes recommendations, but it is the operator who decides whether to act in accordance with the information or recommendations. However, so Boutin claims, it would be difficult — if not impossible — for the operator to decide differently from the information or recommendations acquired, which would imply a tenuous link between their conduct and the occurrence of a breach of international law. Instead, she claims, it would be possible to identify such a causal link to the conduct of those in the chain of command who decided to employ the decision-making support system in question, as these persons are in a position to assess the appropriateness of using the system, the degree of control by the operator necessary in the circumstances, and so on.[72] As previously mentioned, according to Boutin problems of attribution would arise in a fourth scenario: the use of intelligent AWS capable of operating without the possibility of operator intervention. Boutin posits that in this scenario the causal link between the human conduct — evidently consisting only in having deployed the weapon system — and the occurrence of breach of international law, caused by an autonomous activity of the weapon system, would be too tenuous to allow attribution to the state.[73] This is why the solution that she suggests for the purpose of attribution, consists in conceptualizing these systems as operating under the direction and control of the state within the meaning of Article 8 of the ILC Articles on State Responsibility.

Paradoxically, this line of reasoning reveals its weakness precisely because it succeeds in concluding in favour of attribution to the state of breaches of international law caused by the operation of the autonomous systems in all scenarios except the fourth. In particular, the attribution of the relevant human conduct to the state in the second and third scenarios seems to be based on a very broad understanding of causality (the human decision to use the intelligent system has brought about the result). This (implicit) understanding of causality, however, could also be applied to the fourth scenario, and one fails to understand why this should not be the case.[74] However, if causation for the purpose of attribution is so broadly understood, it is an

71 *Ibid.*

72 *Ibid.*, at 142–143.

73 *Ibid.*, at 143.

74 In fact, it is not clear why, according to the author of the study in question, there would, in her opinion, be sufficient causality between the human conduct consisting in deciding to employ and operationalize the use of semi-autonomous weapon systems and the unlawful event caused by an autonomous action of these systems, whereas this causality would, in her opinion, be lacking with respect to the decision to employ and operationalize the use of highly intelligent full-autonomous systems and the unlawful event possibly caused by an autonomous action of these systems.

unnecessary criterion in all cases where there is an activity of a person whose acts are attributable to the state. Rather, the challenge would be to establish causation where a failure to act (omission) results in a violation of international law. Partly because of these challenges, the International Law Commission decided to exclude any requirement of causation between the conduct of persons and a breach of international law for determining the existence of an act of state in the area of state responsibility.[75]

C. The Act of the State in the ILC Articles

The work of the ILC, on the other hand, confirms that the attribution of acts of persons or groups of persons to a state does not require that there be causality between the act and the occurrence of a breach of international law. Indeed, the then Special Rapporteur Roberto Ago, on whose reports the formulation of the attribution rules was based, stated that the attribution of the act of persons or groups of persons to a state is a 'normative' operation, which has 'nothing to do with a link of natural causality or with a link of "material" or "psychological" character'.[76] In support of this assertion, Ago quoted in a footnote to his report the opinion of some scholars, including Anzilotti, according to whom: 'Legal imputation is … clearly distinguishable from causal relationship; an act is legally deemed to be that of a subject of law not because it has been committed or willed by that subject in the physiological or psychological sense of those words, but because it is attributed to him by a rule of law'.[77]

The assumption underlying this theoretical approach is that the state, as a subject of international law, is not 'merely an abstract idea or a figment of the imagination' but is instead a 'real entity, in municipal law as well as in international law'.[78] At the same time, however, the state, 'as a legal person, is not physically capable of conduct', and 'it is obvious that all that can be attributed to a State is the action or omission of an individual or a group of individuals, whatever their composition may be'.[79] Hence 'there are no

75 For a *critique* of the approach adopted by the International Law Commission, see D.M. Pusztai, 'Responsibility without Causation? The Public International Law Experiment', in S. Besson, *International Responsibility: Essays in Law, History and Philosophy* (Schulthess, 2017) 53. The need for a causal link between the act and the occurrence of damage may be required by the content of rule, and therefore be an element to prove for establishing the breach of the rule in question. See in this regard V. Lanovoy, 'Causation in the Law of Responsibility', 90 *The British Yearbook of International Law* (2022) (advance access).

76 R. Ago, Third Report on State Responsibility, Yearbook of the International Law Commission, vol II, Part One, at 218. On the 'normative' approach adopted by the International Law Commission, see L. Condorelli and C. Kreβ, 'The Rules of Attribution: General Considerations', in J. Crawford, A. Pellet, S. Olleson (eds), *The Law of International Responsibility* (Oxford University Press, 2010) 221.

77 *Ibid.*, fn 77, which contains the quote in Italian and the translation in English quoted above prepared by the United Nations Secretariat.

78 *Ibid.*, at 217–218.

79 *Ibid.*, at 217.

activities of the State which can be called "its own" from the point of view of natural causality as distinct from that of legal attribution'.[80]

The position taken by Ago was accepted by the ILC. The commentary to Article 3 of the ILC Articles on State Responsibility establishes in fact that '[t]he attribution of conduct to the State as a subject of international law is based on criteria determined by international law and not on the mere recognition of a link of factual causality'.[81] It goes on to clarify that the attribution rules formulated by the ILC only establish which conducts are to be considered 'acts of the state' for the purposes of its international responsibility, but in themselves have no relevance for establishing the unlawful nature of the conduct.[82]

In essence, the attribution rules formulated by the ILC, following the approach proposed by Roberto Ago, are not based on the assumption that the conduct by a person or group of persons must have caused a breach of international law for the conduct to be considered an act of the state. As one commentator has rightly pointed out, this would in fact introduce a distinction between an 'event contrary to international law' and an 'internationally wrongful act' of which there is no trace in the ILC Articles on State Responsibility.[83] As has been observed in the doctrine, the basis of the attribution rules in the ILC Articles on State Responsibility is in fact 'functional' in nature (and not causal). In other words, the conduct of persons or groups of persons is attributed to the state when there is a connection between the conduct and the functions of the state, since the state — conceived as an organization — can only act through persons or groups of persons.[84] For the purposes of attribution, therefore, it does not matter that the persons caused the breach, whereas it matters that they acted to perform a function of the state.

If one follows this approach, then it is clear that, with respect to the use of any intelligent system (including AWS), the question that some authors ask with respect to the problems that might arise regarding the attribution of actions of such systems to the state appears to be beside the point. If the systems in question are deployed by persons whose conduct is attributable to the state, according to the criteria formulated by the ILC, it is of no

80 *Ibid.*, at 218, fn 78.

81 *Yearbook of the International Law Commission*, 2001, at 38–39.

82 'As a normative operation, attribution must be clearly distinguished from the characterization of conduct as internationally wrongful. Its concern is to establish that there is an act of the State for the purposes of responsibility. To show that conduct is attributable to the State says nothing, as such, about the legality or otherwise of that conduct, and rules of attribution should not be formulated in terms which imply otherwise.' *Ibid.*, at 39.

83 On this point, see the apt remarks of I. Plakokefalos, 'Causation in the Law of State Responsibility and the Problem of Overdetermination: In Search of Clarity', 26 *European Journal of International Law* (2015) 471.

84 The functional basis of the attribution rules in the Articles of Responsibility is ably explained and analysed by S. Fleming, 'Causation, Fault and Function in the Rules of Attribution', in S. Besson, *Theories of International Responsibility Law* (Cambridge University Press, 2022) 229, in particular 243–247.

relevance whether the occurrence of the breach is attributable to a natural causality with an 'autonomous' action of the system.

4. Conclusions

This article has clarified that systems that are enabled by so-called 'strong' artificial intelligence, however capable they may be of operating autonomously from the user and of performing actions that the user cannot foresee, are still only 'tools' of the user. This is particularly true for intelligent AWS. However technologically advanced they may be, under international humanitarian law such weapons are still only means of warfare, namely means that the parties to an armed conflict may use (under certain conditions) to conduct hostilities of war.

With regard to criminal liability for war crimes in the case of 'unintended engagements' resulting from the use of intelligent AWS, we are still far from realistically foreseeing criminal liability for the systems in question at present. If this solution were to be reached in the future, it would not be because of the impossibility of regarding the user of the weapon system as the author of the material act constituting the relevant war crimes. These crimes are in fact those generically referred to as 'unlawful attacks', the material act of which is to direct, make or launch a military attack. Natural persons unquestionably carry out this act, whatever means of warfare they have chosen to use. Crimes of 'unlawful attacks' also invariably require various circumstances for the act to be criminal. The person who directed, made or launched the attack using an intelligent AWS may intend the materialization of such circumstances, or it may be the result of a failure of the weapon system. All this, however, relates to the sphere of the subjective element of the offence, which in some formulations consists of intent and in others includes recklessness. Even in the most demanding formulations of the subjective element on the circumstances, it would be too hasty, however, to claim that an 'unintended engagement' resulting from the use of intelligent AWS is always 'unintended' from a criminal law point of view. For instance, there could be intent on the part of one who conducts a military attack with an intelligent AWS that has previously produced 'unintended engagements'. Finally, yet to be explored, is the possible criminal liability of the programmers and developers of intelligent AWS in light of the obligation to verify the potential unlawfulness of the new means and methods of conducting hostilities enshrined in Article 36 of the First Additional Protocol.

For the purposes of international state responsibility for violation of international humanitarian law, the question of attributing the 'unintended engagement' of the intelligent AWS to the state that used the weapon is a question that simply does not arise. One reaches this conclusion in light of the concept of 'act of the state' as it relates to the international responsibility of the state for internationally wrongful acts. Furthermore, it bears reiterating that under international humanitarian law the rules on prohibited attacks

bind directly the belligerents and parties to the conflict.[85] Belligerents and parties to an armed conflict must therefore comply with these prohibitions in all circumstances, irrespective of the type of weapon used in the course of an armed attack, including where the attack is carried out with an intelligent AWS.[86] Unintended engagements in the course of the attack, whatever their cause, are therefore attributable to the state that directed and launched the attack. To doubt this by asserting that, in light of the specific technology built into a weapon system, the unintended engagement might not qualify as an act of the state that used the weapon for the purposes of its international responsibility is — in my opinion — a dangerous intellectual exercise.

The absence of precise international obligations of transparency and information on how states conduct their military operations is already a matter of serious concern for those who would like to ensure that violations of the applicable rules can be investigated.[87] Adding to the 'black box' resulting from the lack of transparency and information on the conduct of hostilities the 'black box' of systems enabled by strong artificial intelligence[88] would contribute to the further weakening of such possibilities. States that use intelligent AWS could argue that prohibited military attacks are the result of an

85 See the formulation of the rules contained in relevant treaties such as the Regulations respecting the Laws and Customs of War on Land annexed to 1907 Hague Convention IV and the two Additional Protocols to the Geneva Conventions.

Admittedly, Art. 1 of the 1907 Hague Convention IV and the annexed Hague Regulations provides that contracting parties are obliged to 'issue instructions to their armed land forces which shall be in conformity with the Regulations'. This is a specific obligation, which however does not imply that contracting parties are not directly obliged — under the Convention — to comply with the Regulations. This is made clear by Art. 2, which in stating the scope of applicability of the Regulations (the so-called '*si omnes*' clause), states: 'The provisions contained in the Regulations … as well as in the present Convention do not apply except between Contracting Powers, and then only if all the belligerents are parties to the Convention.'

86 This is not a trivial observation, as it may appear at first sight. In certain areas, there are in fact international norms that impose an obligation on states to establish specific conduct on specific persons, rather than an obligation to carry out such conduct directly. This is the case with obligations to rescue persons in distress at sea. Art. 98(1) of the UNCLOS Convention establishes an obligation for States Parties to require the master of a ship flying its flag to render assistance in the presence of the conditions identified in the rule. Similarly, the SOLAS Convention stipulates that '[t]he master of a ship at sea which is in a position to be able to provide assistance, on receiving information from any source that persons are in distress at sea, is bound to proceed with all speed to their assistance'. The wording of these and similar rules seem to imply that they are only applicable when there is a master of ship on board. Consequently, a debate has arisen concerning the inapplicability of these rules when the ship does not have a master of ship on board, as in the case of the future use of Maritime Autonomous Surface Ship by the shipping industry. In this regard, see for instance M.R. Leopardi, 'Autonomous Shipping: Some Reflections on Navigational Rights and Rescue at Sea', in A. Basu Bal et al. (eds), *Regulation of Risk Transport, Trade and Environment in Perspective* (Brill/Njhoff, 2023) 451, at 465–467.

87 See again Sassòli, *supra* note 51.

88 With reference to systems enabled by strong artificial intelligence, the term 'black box' is used to clarify that the inputs and operations of the system are not 'visible' to the operator or programmer. In other words, the system arrives at conclusions or decisions without explaining how they were reached.

operation by intelligent software, which cannot be explained or predicted by the user and therefore cannot be attributed to them. Even the hypothesis of such a possibility is unacceptable.

As is well-known, technological advances — including those in the field of armaments — risk making existing legal regulations obsolete. This is also true in the field of artificial intelligence. In the interpretation and application of existing law, and in devising future regulations, it is necessary to remain anchored in human intelligence.