


ARTICLE

Military Artificial Intelligence and the Principle of Distinction: A State Responsibility Perspective

Magdalena Pacholska 

Marie Skłodowska-Curie Post-Doctoral Fellow, Asser Institute (The Hague), University of Amsterdam (The Netherlands)

E-mail: m.pacholska@asser.nl.

(First published online 13 December 2022)

Abstract

Military artificial intelligence (AI)-enabled technology might still be in the relatively fledgling stages but the debate on how to regulate its use is already in full swing. Much of the discussion revolves around autonomous weapons systems (AWS) and the ‘responsibility gap’ they would ostensibly produce. This contribution argues that while some military AI technologies may indeed cause a range of conceptual hurdles in the realm of individual responsibility, they do not raise any unique issues under the law of state responsibility. The following analysis considers the latter regime and maps out crucial junctions in applying it to potential violations of the cornerstone of international humanitarian law (IHL) – the principle of distinction – resulting from the use of AI-enabled military technologies. It reveals that any challenges in ascribing responsibility in cases involving AWS would not be caused by the incorporation of AI, but stem from pre-existing systemic shortcomings of IHL and the unclear reverberations of mistakes thereunder. The article reiterates that state responsibility for the effects of AWS deployment is always retained through the commander’s ultimate responsibility to authorise weapon deployment in accordance with IHL. It is proposed, however, that should the so-called fully autonomous weapon systems – that is, machine learning-based lethal systems that are capable of changing their own rules of operation beyond a predetermined framework – ever be fielded, it might be fairer to attribute their conduct to the fielding state, by conceptualising them as state agents, and treat them akin to state organs.

Keywords: artificial intelligence; state responsibility; principle of distinction; mistake of fact; autonomous weapons systems

© The Author(s), 2022. Published by Cambridge University Press in association with the Faculty of Law, the Hebrew University of Jerusalem. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

1. Introduction

The adoption of AI in military practice has been described as the third revolution in military affairs, after gunpowder and nuclear weapons.¹ Despite having been envisioned over four decades ago, military AI caught both the academic and political centres of the international community off guard. Nowhere is this bewilderment more self-evident than in the international law realm, where heated debates continue on whether the existing legal frameworks, in particular international humanitarian law (IHL), are sufficient to account for the drastic change that military AI is expected to bring to warfare. In late 2019, the Group of Governmental Experts (GGE) – established by the state parties to the Convention on Certain Conventional Weapons to work on the challenges raised by lethal autonomous weapons systems (LAWS) – produced a list of 11 tentative ‘Guiding Principles’ but failed to reach agreement on the very definition of LAWS or the concept of ‘autonomy’ with regard to such systems.² While many questions remain unanswered, two aspects are clear. First, there is no turning back on the defence and security potential of AI, already seen in military circles as a pervasive technology.³ Second, and equally importantly, AI-enabled military technology goes beyond LAWS, and is already seen in armed conflicts in the form of, inter alia, risk-assessing predictive algorithms used in a variety of military systems.⁴

The global political landscape suggests that a comprehensive prohibition of either LAWS or AI-enabled military technology is not likely to be adopted in the foreseeable future.⁵ Yet, given the significant technological advances of the last years, a steady increase in the integration of AI in military systems is inevitable.⁶ Sooner or later, such systems – just like all other weaponry – will malfunction and result in, inter alia, injuries to civilians (as system malfunctions are inexorable in complex, coupled systems), bringing to the fore

¹ On the strategic importance of AI, see Rod Thornton and Marina Miron, ‘Towards the “Third Revolution in Military Affairs”: The Russian Military’s Use of AI-Enabled Cyber Warfare’ (2020) 165 *The RUSI Journal* 12.

² Meeting of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects [(entered into force 2 December 1983) 1342 UNTS 137], ‘Guiding Principles affirmed by the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons System’ (13 December 2019), UN Doc CCW/MSP/2019, Annex III: Guiding Principles affirmed by the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons System.

³ North Atlantic Treaty Organization, ‘Summary of the NATO Artificial Intelligence Strategy’, 22 October 2021, https://www.nato.int/cps/en/natohq/official_texts_187617.htm.

⁴ For an overview of the existing military AI, and its future utility across military activities, see Daniel S Hoadley and Kelley M Saylor, ‘Artificial Intelligence and National Security’, Congressional Research Service (CRS), 10 November 2020, R45178, 9–15.

⁵ In mid-2021, the International Committee of the Red Cross (ICRC) put forward a position that at least anti-personnel autonomous weapons system should be banned but, so far, this proposal has not been enthusiastically received by states: ICRC, ‘Position on Autonomous Weapon Systems’, 12 May 2021, 2, <https://www.icrc.org/en/document/icrc-position-autonomous-weapon-systems>.

⁶ Forrest E Morgan and others, ‘Military Applications of Artificial Intelligence: Ethical Concerns in an Uncertain World’ (RAND Corporation 2020) xii.

the question of who is responsible for them.⁷ While it is widely accepted that IHL fully applies to the use of AI-enabled technology,⁸ the issue of accountability for IHL violations resulting from the use of such technology remains highly contentious. In fact, for over a decade now, international legal scholarship has grappled with the ostensible ‘responsibility gap’⁹ that AI-enabled military technology, in general, and LAWS, in particular, would create.¹⁰ A large part of the debate has centred on the challenges of holding individuals responsible for war crimes perpetrated ‘by’ AI.¹¹ Some attention has been devoted to the idea of ‘attributing electronic personhood to robots’¹² but, in the more contemporary literature, holding arms manufacturing corporations accountable seems to be gaining more traction than far-fetched attempts to ascribe blame to machines.¹³ Somewhat surprisingly, state responsibility, in turn, has been subject to rather cursory treatment.¹⁴ A few scholars asserted that the conduct of fully autonomous machines could not be attributed to states

⁷ For international responsibility purposes, a simple ‘mechanical’ malfunction of a weapon should be distinguished from a failure in design. The first is usually considered *force majeure* and, as such, excuses the wrongfulness of resulting conduct. An overlooked failure in design could, under certain conditions, lead to responsibility under Art 36 of the Protocol Additional to the Geneva Conventions of 12 August 1949 and relating to the Protection of Victims of International Armed Conflicts (entered into force 7 December 1978) 1125 UNTS 3 (AP I). For a discussion of Art 36 and a review of military AI applications, see Tobias Vestner and Altea Rossi, ‘Legal Reviews of War Algorithms’ (2021) 97 *International Law Studies* 509. See also the reflections in Section 5.

⁸ GGE Guiding Principles (n 2) Principle a.

⁹ Andreas Matthias, ‘The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata’ (2004) 6 *Ethics and Information Technology* 175.

¹⁰ Human Rights Watch (HRW) and Harvard Law School International Human Rights Clinic, ‘Mind the Gap: The Lack of Accountability for Killer Robots’, 9 April 2015, <https://www.hrw.org/report/2015/04/09/mind-gap/lack-accountability-killer-robots>; Michael N. Schmitt, ‘Autonomous Weapon Systems and International Humanitarian Law: A Reply to the Critics’, *Harvard Law School National Security Journal*, 5 February 2013, <https://harvardnsj.org/2013/02/autonomous-weapon-systems-and-international-humanitarian-law-a-reply-to-the-critics/>; Marco Sassóli, ‘Autonomous Weapons and International Humanitarian Law: Advantages, Open Technical Questions and Legal Issues To Be Clarified’ (2014) 90 *International Law Studies* 308.

¹¹ Marta Bo, ‘Autonomous Weapons and the Responsibility Gap in Light of the *Mens Rea* of the War Crime of Attacking Civilians in the ICC Statute’ (2021) 19 *Journal of International Criminal Justice* 275.

¹² European Parliamentary Research Service (EPRS), ‘The Ethics of Artificial Intelligence: Issues and Initiatives’, PE 634.452, March 2020, 20. Interestingly, granting some aspects of personhood to AI is on the rise: in 2021, South African and Australian courts held that AI machines can be granted patent rights, while the United Kingdom (UK) and United States (US) courts decided against such a possibility; see generally on the DABUS patent battle Paulina M Starostka and Daniel Schwartz, ‘South Africa and Australia Break from U.S. and U.K. To Give DABUS Its First IP Breaks’, *Nixon Peabody Blog*, 10 August 2021, <https://www.nixonpeabody.com/insights/articles/2021/08/10/south-africa-and-australia-break-from-u-s-and-u-k-to-give-dabus-its-first-ip-breaks>.

¹³ EPRS (n 12) 22–26; similarly, Robin Geiß, ‘State Control Over the Use of Autonomous Weapon Systems: Risk Management and State Responsibility’ in Rogier Bartels and others (eds), *Military Operations and the Notion of Control under International Law* (TMC Asser Press 2021) 439, 448.

¹⁴ For a rare exception, see Berenice Boutin, ‘State Responsibility in relation to Military Applications of Artificial Intelligence’, TMC Asser Institute for International & European Law, Asser Research Paper 2022-09, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4214292.

in the absence of direct and effective control over their conduct,¹⁵ while some reached the opposite conclusion,¹⁶ often without a comprehensive examination of relevant modalities.¹⁷

The indifference towards state responsibility as a regime relevant in the context of LAWS seems to be ending, though. In the report of its 2022 session, the GGE on LAWS paid heed to its role by stressing that:¹⁸

every internationally wrongful act of a state, including those potentially involving weapons systems based on emerging technologies in the area of LAWS entails international responsibility of that state, in accordance with international law. ... Humans responsible for the planning and conducting of attacks must comply with international humanitarian law.

This article elaborates on this acute GGE premise and demonstrates how the regime of state responsibility applies to the scenario most feared by the opponents of LAWS – that is, a mistaken attack on civilians committed by a state’s armed forces using AI-enabled military technology. It demonstrates that while some legal aspects of AI in the military context remain to be settled, AI has not been developing in ‘a regulatory vacuum’ as frequently purported in the literature,¹⁹ and while the ‘responsibility gap’ exists, it is not where most of the commentators assume it is. The discussion proceeds as follows. Section 2 explains the concept of military AI and distinguishes between (i) already existing AI-powered weapon systems, referred to in this article simply as AWS, and (ii) future potential fully autonomous weapon systems (FAWS).²⁰ It further sets out an imaginary – albeit modelled on already fielded projects – bifurcated scenario in which both types of system contribute to making civilians the object of an attack in the midst of an armed conflict. The following two sections examine the relevant primary rules (which establish obligations incumbent on states) and secondary rules (which regulate the existence of a breach of an international obligation and its consequences). Section 3 inquires whether mistaken attacks on civilians violate the principle of distinction, and demonstrates that any challenges in holding states accountable for the harm caused by AI-powered systems will stem from pre-existing systemic

¹⁵ J-G Castel and Matthew E Castel, ‘The Road to Artificial Super-Intelligence: Has International Law a Role to Play?’ (2016) 14 *Canadian Journal of Law and Technology* 1, 9.

¹⁶ Rebecca Crootof, ‘War Torts: Accountability for Autonomous Weapons’ (2016) 164 *University of Pennsylvania Law Review* 1347, 1389–93; Jack M Beard, ‘Autonomous Weapons and Human Responsibilities’ (2014) 45 *Georgetown Journal of International Law* 617, 663–78; Geiß (n 13) 448; NATO JAPCC, ‘Future Unmanned System Technologies: Legal and Ethical Implications of Increasing Automation’, 2016, 30.

¹⁷ HRW and Harvard Law School (n 10) 13 (simply asserting that ‘state responsibility for the unlawful acts of fully autonomous weapons could be assigned relatively easily to the user state’).

¹⁸ Report of the 2022 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Systems (29 July 2022), UN Doc CCW/GGE.1/2022/CRP.1/Rev.1, para 19.

¹⁹ Matthew U Scherer, ‘Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, Strategies’ (2016) 29 *Harvard Journal of Law and Technology* 353.

²⁰ See practical reservations concerning the likelihood of that happening in Section 2.

shortcomings of the applicable primary rules, in this case IHL, rather than the incorporation of AI as such. Section 4 examines how the secondary rules of state responsibility apply to wrongdoing caused by both today's AWS and future FAWS, should those ever be fielded. Section 5 concludes by offering some tentative solutions for the identified loopholes and indicating avenues for further research.

A few clarifications are required before venturing into the discussion. First, starting from the premise that individual and state responsibility are complementary and concurrent,²¹ this article steers clear from delving into a discussion of which regime is preferable in relation to the harm resulting from the use of AI in the military context.²² Second, it is not the intention of the article to reopen the controversial debate over the concept of 'international crimes of states' and the criminalisation of state responsibility to which it could arguably lead.²³ The following analysis pertains solely to what has been referred to as a 'plain, "vanilla" violation of IHL',²⁴ – namely, a violation of the principle of distinction as such, not the war crime of intentionally directing an attack against civilians.²⁵ Finally, because of its limited scope, the article focuses on the specific problem of *post facto* attribution of an internationally wrongful act to a state and does not aspire to provide an exhaustive examination of all aspects of state responsibility in relation to AI in the military domain.²⁶ In particular, it does not discuss the pre-deployment obligations of states to ensure the compliance of a new weapon, means or method of warfare with IHL, which it leaves to other commentators.²⁷

²¹ Paola Gaeta and Abhimanyu George Jain, 'Individualisation of IHL Rules through Criminal Responsibility for War Crimes and Some (Un)Intended Consequences', 2 June 2021, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3853333; André Nollkaemper, 'Concurrence between Individual Responsibility and State Responsibility in International Law' (2003) 52 *International and Comparative Law Quarterly* 615.

²² For an overview of the contrasting positions see, eg, Daniel Amoroso, 'Jus in Bello and Jus ad Bellum Arguments against Autonomy in Weapons Systems: A Re-Appraisal' (2017) 43 *Questions of International Law* 5 (favouring individual over collective responsibility), and Daniel N Hammond, 'Autonomous Weapons and the Problem of State Accountability' (2015) 15 *Chicago Journal of International Law* 652 (arguing the opposite).

²³ See, generally, Marina Spinedi, 'Crimes of State: The Legislative History' in Joseph H Weiler, Antonio Cassese and Marina Spinedi (eds), *International Crimes of States: A Critical Analysis of the International Law Commission's Draft Article 19 on State Responsibility* (Walter de Gruyter 1989) 5, 7.

²⁴ The term is borrowed from Marko Milanovic, 'Mistakes of Fact when Using Lethal Force: Part I', *EJIL: Talk!*, 14 January 2020, <https://www.ejiltalk.org/mistakes-of-fact-when-using-lethal-force-in-international-law-part-i>.

²⁵ Bo (n 11) 285–95.

²⁶ This article does not discuss the plethora of scenarios in which a state could be ascribed with derived responsibility in relation to the conduct of other actors, be it on the domestic or international plane. For more on those aspects of military AI see Boutin (n 14) 24–28, and in relation to LAWS see Dan Saxon, *Fighting Machines: Autonomous Weapons and Human Dignity* (University of Pennsylvania Press 2021) 106–22.

²⁷ See, for instance, Vestner and Rossi (n 7); Dustin A Lewis, 'Legal Reviews of Weapons, Means and Methods of Warfare Involving Artificial Intelligence: 16 Elements to Consider', *Humanitarian Law & Policy*, 21 March 2019, <https://blogs.icrc.org/law-and-policy/2019/03/21/legal-reviews-weapons-means-methods-warfare-artificial-intelligence-16-elements-consider>.

2. Overview of military AI

In the absence of a universally accepted definition of AI, many contemporary analyses, position papers and policies on its role in the military domain adopt a simple understanding of AI as ‘the ability of machines to perform tasks that normally require human intelligence – for example, recognizing patterns, learning from experience, drawing conclusions, making predictions, or taking action – whether digitally or as the smart software behind autonomous physical systems’.²⁸ While research into AI arguably had started in the 1940s,²⁹ the ‘AI hype’, which began in the early 2010s and still continues, is often associated with three interlinked developments:

- (a) the increasing availability of ‘big data’ from a variety of sources;
- (b) improved machine learning algorithms and approaches; and
- (c) spiking computer processing power.³⁰

An explosion of interest in the military applications of AI, dubbed already by some an ‘AI arms race’,³¹ started some time in the late 2010s after China’s State Council released a grand strategy to make the country a global AI leader by 2030, and President Vladimir Putin announced Russia’s interest in AI technologies by stating that ‘whoever becomes the leader in this field will rule the world’.³² Unsurprisingly, soon thereafter the United States designated AI as one of the means that will ‘ensure [the US] will be able to fight and win the wars of the future’.³³ Similar sentiment has been echoed among members of the North Atlantic Treaty Alliance.³⁴

With the increased buzz around military AI – fuelled by the Campaign to Stop Killer Robots³⁵ – public debate often overlooks that autonomy or automation³⁶ has been incorporated into various military systems for

²⁸ NATO Science & Technology Organization, ‘Science & Technology Trends 2020–2040: Exploring the S&T Edge’, March 2020, 50, https://www.nato.int/nato_static_fl2014/assets/pdf/2020/4/pdf/190422-ST_Tech_Trends_Report_2020-2040.pdf.

²⁹ See, eg, Warren S McCulloch and Walter H Pitts, ‘A Logical Calculus of the Ideas Immanent in Nervous Activity’ (1943) 5 *Bulletin of Mathematical Biophysics* 115. The inception of the concept of AI is typically associated with the seminal 1950 article by Alan Turing in which he posed the question of whether machines can think, and proposed a litmus test for answering it: Alan M Turing, ‘Computing Machinery and Intelligence’ (1950) 59 *Mind* 433.

³⁰ Hoadley and Sayler (n 4) 2.

³¹ For a recent critique of such a narrative see Paul Scharre, ‘Debunking the AI Arms Race Theory’ (2021) 4 *Texas National Security Review* 121.

³² Hoadley and Sayler (n 4) 1.

³³ US Department of Defense, ‘Summary of the 2018 National Defense Strategy: Sharpening the American Military’s Competitive Edge’ 2018, 3.

³⁴ NATO Science & Technology Organization (n 28) 50.

³⁵ Campaign to Stop Killer Robots is a coalition of non-governmental organisations (NGOs) lobbying internationally for a preventive ban on LAWS; see more at <https://www.stopkillerrobots.org>.

³⁶ As rightly pointed out in ICRC, ‘Autonomy, Artificial Intelligence and Robotics: Technical Aspects of Human Control’, August 2019, 7, there is ‘no clear technical distinction between automated and autonomous systems’. The terms ‘automatic’ or ‘automated’ are often used to refer to rule-based systems that mechanically respond to environmental input; see the discussion in

decades.³⁷ In fact, human-machine teaming has been a component of modern warfare at least since the First World War,³⁸ but '[t]raditionally, humans and automated systems have fulfilled complementary but separated functions within military decision making'.³⁹ What the recent advancements in AI technology facilitate is merely a more synchronised, or even integrated, functioning of humans and technology. This, in turn, allows for AI to be incorporated into both selected components of military planning and operations (such as logistics, maintenance, medical and casualty evacuation) as well as into complex C4ISR systems.⁴⁰ Furthermore, AI is already proving to be particularly useful in intelligence, where the ability to comb through a large amount of data and automate the process of searching for actionable information may translate into immediate tactical advantage on the battlefield. As demonstrated by the 2021 Israeli Operation Guardian of the Walls, considered by some as an 'AI-Enhanced Military Intelligence Warfare Precedent', AI-powered intelligence gathering and analysis may even lead to a new concept of operations.⁴¹

It is against this background that the ongoing debate on LAWS, mentioned in the opening of this article, should be viewed. It is worth noting at the outset

Paul Scharre and Michael C Horowitz, 'An Introduction to Autonomy in Weapon Systems: A Primer', *Center for a New American Security*, February 2015. 'Autonomy', in turn, can be described as 'a capability (or set of capabilities) that enables a particular action of a system to be automatic or, within programmed boundaries, "self-governing"; see more in US Department of Defense, 'Task Force Report: The Role of Autonomy in DoD Systems', July 2012, para 1.1.

³⁷ Executive Office of the President, National Science and Technology Council and Committee on Technology, 'Preparing for the Future of Artificial Intelligence', 12 October 2016, 37, https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf.

³⁸ Jonathan BA Bailey, 'The First World War and the Birth of Modern Warfare' in MacGregor Knox and Williamson Murray (eds), *The Dynamics of Military Revolution: 1300-2050* (Cambridge University Press 2001) 132.

³⁹ Karel van den Bosch and Adelbert Bronkhorst, 'Human-AI Cooperation to Benefit Military Decision Making', NATO S&T Organization, 12 July 2018, STO-MP-IST-160, 1.

⁴⁰ In military parlance, C4ISR stands for Command, Control, Communications, Computers (C4) Intelligence, Surveillance and Reconnaissance (ISR). C4ISR systems are already produced by a number of global defence and security companies, including BAE Systems, Lockheed Martin, Thales. On the incorporation of AI into such decision-support systems, see Arwin Datumaya Wahyudi Sumari, Adang Suwandu Ahmad and Cognitive Artificial Intelligence Research Group (CAIRG), 'The Application of Cognitive Artificial Intelligence within C4ISR Framework for National Resilience', Fourth Asian Conference on Defence Technology – Japan (ACDT), 29 November–1 December 2017, <https://ieeexplore.ieee.org/document/8259600>.

⁴¹ Avi Kalo, 'AI-Enhanced Military Intelligence Warfare Precedent: Lessons from IDF's Operation "Guardian of the Walls"', *Frost Perspectives*, 9 June 2021, <https://www.frost.com/frost-perspectives/ai-enhanced-military-intelligence-warfare-precedent-lessons-from-idfs-operation-guardian-of-the-walls> ('[The IDF's Intelligence Division] established a comprehensive, "one-stop shop" intelligence war machine, gathering all relevant players in intelligence planning and direction, collection, processing and exploitation, analysis and production, and dissemination process (PCPAD) into what it termed "intelligence molecules". During the recent conflict, massive AI machinery for Big Data Analytics provided support at every level—from raw data collection and interception, data research and analysis, right up to strategic planning—with the objective of enhancing and accelerating the entire process, from decision-making about prospective targets to the actual carrying out of attacks by pilots from F-35 cockpits').

that the discussions of LAWS on both political and scholarly fora have been obfuscated by overhyped narratives and misunderstandings of the existing technologies, both of which feed into the lack of a universally accepted definition of LAWS, sometimes also alarmingly called ‘killer robots’.⁴² A closer look at the paper trail of GGE on LAWS shows, however, an emerging realisation of a conceptual (and in the future possibly also normative) distinction between:⁴³

- the existing systems incorporating various degrees of automation; and
- the still-to-be-fielded lethal machine learning-based systems capable of changing its own rules of operation beyond a predetermined framework.

For the sake of conceptual clarity, the first category will be referred to in the following analysis as autonomous weapon systems (AWS), and the second as fully autonomous weapon systems (FAWS).

The two most-often referenced conceptualisations of AWS – that is, by the International Committee of the Red Cross (ICRC) and the United States – both reflect the ongoing fusion of AI into the military targeting cycle, and define AWS as weapons systems that, after being activated by a human operator, can ‘select’ and attack/engage targets without ‘human intervention’ (ICRC),⁴⁴ or ‘further intervention by a human operator’ (US).⁴⁵ Target ‘selection’ is often misunderstood in legal scholarship and perceived as the weapon’s ability to choose targets freely, resulting in ‘the removal of human operators from the targeting decision-making process’;⁴⁶ this is incorrect. Target selection is the process of analysing and evaluating potential threats and targets for engagement (attack). The final decision point before a target is destroyed is known in military parlance as engagement. The existing AWS utilise a variety of automated target recognition (ATR) systems, first developed back in the 1970s, which employ pattern recognition to identify potential threats – comparing emissions (sound, heat, radar, radio-frequency), appearance (shape and height)

⁴² The nomenclature used in state positions, NGO position pieces and academic scholarship is perplexing, as different actors often use a given term or even a definition to refer to systems with different technical specifications. For a strongly voiced criticism of the LAWS debate, see Chris Jenks, ‘False Rubicons, Moral Panic, & Conceptual Cul-de-Sacs: Critiquing & Reframing the Call to Ban Lethal Autonomous Weapons’ (2016) 44 *Pepperdine Law Review* 1.

⁴³ While the need for such a distinction has been alluded to by many states, it was most clearly spelled out by France. For an analysis of the French position in English, see Jean-Baptiste Jeangène Vilmer, ‘A French Opinion on the Ethics of Autonomous Weapons’, *War on the Rocks*, 2 June 2021, <https://warontherocks.com/2021/06/the-french-defense-ethics-committees-opinion-on-autonomous-weapons>.

⁴⁴ ICRC (n 36) 5.

⁴⁵ US Department of Defense, ‘Directive Number 3000.09: Autonomy in Weapon Systems’, 21 November 2012 (incorporating Change 1, 8 May 2017). Note that as of mid-2022, the Directive is being revised again.

⁴⁶ Daniele Amoroso and Guglielmo Tamburrini, ‘Toward a Normative Model of Meaningful Human Control over Weapons Systems’ (2021) 35 *Ethics and International Affairs* 245, 253.

or other characteristics (trajectory, behaviour) against a human-defined library of patterns that correspond to intended targets.⁴⁷ Even the most advanced versions of AWS thus 'select' specific targets from a human pre-defined class or category. One of the most widely employed examples of such technology are close-in defence weapon systems (CIWS), developed to provide defence for military bases, naval ships (like the Dutch Goalkeeper or American Phalanx) or other geographically limited zones (such as the Israeli Iron Dome or David's Sling).⁴⁸ A CIWS identifies incoming threats and determines the optimal firing time to neutralise the threat in a way that maximises protection in situations that require an almost immediate decision, and where threats come at a volume and speed which would overwhelm human capacity.

FAWS, in turn, are more elusive, which seems understandable given that they do not exist. As a prospective feature of weapon systems, 'full autonomy' seems to be conceptualised as a capability to change rules of operation beyond a predetermined framework coupled with the impossibility of terminating target engagement.⁴⁹ For many military experts, FAWS understood as such are pure fantasy,⁵⁰ but as the fear of such systems being fielded persists, this article entertains such a possibility for the sake of analysis.

While rarely addressed explicitly, the true concern of military AI in general, and weapon systems in particular, is using it directly against human targets, which according to some would lead to the 'responsibility gap' – that is, a situation in which no entity could be held responsible for the wrong done.⁵¹ In the context of targeting, the problem can therefore be conceptualised as follows. How does the 'outsourcing' of certain elements of the targeting process⁵² to an AI-powered system affect accountability for potential misidentification of a human target? In essence, the core issue is what happens when military AI contributes to inadvertently making civilians the object of an attack. Any work on anti-personnel CIWS is presumptive and classified, but imagine the following scenario, to exemplify the problem:

⁴⁷ Vincent Boulanin and Maaike Verbruggen, 'Mapping the Development of Autonomy in Weapon Systems', Stockholm International Peace Research Institute (SIPRI), November 2017, 24–25.

⁴⁸ *ibid* 36–37.

⁴⁹ As explicitly set forth by the French; see the details in Vilmer (n 43). Cf the categories set out in the US DoD Directive (n 45) Glossary, and the elements of FAWS proposed by the Chinese, Group of Governmental Experts of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons which May Be Deemed Excessively Injurious or to Have Indiscriminate Effects, 'Position Paper Submitted by China', 11 April 2018, CCW/GGE.1/2018/WP.7, para 3.

⁵⁰ James Kraska, 'Command Accountability for AI Weapon Systems in the Law of Armed Conflict' (2021) 97 *International Law Studies* 407, 408.

⁵¹ Robert Sparrow, 'Killer Robots' (2007) 24 *Journal of Applied Philosophy* 62, 66.

⁵² It is often overlooked in international legal scholarship that contemporary military targeting is a complex multistep process, not a simple kinetic action. For a succinct yet accurate explanation of the importance thereof and elucidation of the steps see Merel AC Ekelhof, 'Lifting the Fog of Targeting: "Autonomous Weapons" and Human Control through the Lens of Military Targeting' (2018) 71 *Naval War College Review* 1.

State Alpha is fighting armed group Beta, which is under the effective control of another state, in a foreign territory. As part of this international armed conflict, Alpha operates a military base near the front line (in the said foreign territory), guarded by a CIWS capable of intercepting both material and human incoming threats. The CIWS is programmed to identify threats based on whether they are carrying a weapon, are within a defined perimeter of the base and exhibit indicators of hostile intent (such as failing to heed warnings or avoiding roads), and is programmed to exclude friendly forces (such as those wearing allied uniforms). One evening, Beta attacks a village located a few kilometres from the base. Many civilians flee from the village and carry machetes for protection against Beta fighters pursuing them. As these civilians approach the base after dark through the fields, the CIWS identifies them as a potential target. An Alpha commander orders the initiation of perimeter defence protocols, including lights and audio warnings. The fleeing villagers do not heed the warnings and continue to proceed towards the base.

This is a crucial junction for the purposes of the ensuing analysis. The scenario is therefore bifurcated from this point onwards:

- The Alpha commander requests optical confirmation of potential threats from human sentries, who confirm the hostile intent of the approaching group. The commander thus relies on the CIWS's suggestion and authorises the engagement of the approaching group (Variant I);
- a fully autonomous CIWS, which does not require separate confirmation engagement, fires at the villagers (Variant II).

In both variants the approaching civilians, who are not directly participating in hostilities, are the direct object of the attack. Can either of the variants be classified as a violation of the principle of distinction? This is examined in the following section.

3. Mistakenly attacking civilians: A violation of the principle of distinction?

Many IHL nuances remain disputed, but few are left un(der)explored. Yet, startling as it may be, 44 years after the Additional Protocols to the Geneva Conventions⁵³ entered into force, the question of whether mistakenly attacking civilians constitutes a violation of the principle of distinction has received surprisingly little attention.⁵⁴ This could be because, in real life, the legal evaluation of situations in which civilians were an object of attack⁵⁵ concentrates on

⁵³ AP I (n 7); Protocol Additional to the Geneva Conventions of 12 August 1949 and relating to the Protection of Victims of Non-International Armed Conflicts (entered into force 7 December 1978) 1125 UNTS 3.

⁵⁴ 'Honest Errors?', Online Conference on Combat Decision Autonomy 75 Years after the *Hostages* Case, 4 June 2021, <https://www.honesterrors.net>, constitutes a rare exception; the post-conference edited volume, forthcoming in 2023, might be the first comprehensive examination of honest mistakes.

⁵⁵ As opposed to being collateral damage of an attack directed at a military objective, the legality of which is evaluated under the rule of proportionality. On the incorporation of AI into systems

the determination of whether all feasible precautions were taken, and especially whether the target was verified before launching the attack. In other words, potential violations of the principle of distinction frequently go hand in hand with *prima facie* violations of the principle of precaution, aptly characterised as ‘the procedural corollary to the general obligation to distinguish civilians and civilian objects from military objectives’.⁵⁶ Commercial airline shoot-down incidents by surface-to-air missiles – such as the 2020 Iranian downing of the Ukraine International Airlines Flight 752, the 2014 Malaysia Airlines Flight 17 shoot-down over Eastern Ukraine, and the 1988 downing of Iran Air Flight 655 by the US⁵⁷ – are probably the most apparent examples, but air-to-surface attacks on civilians, by both manned and unmanned aircraft, happen even more frequently.⁵⁸ As the CENTCOM Investigation Report on the 2015 US air strike on the Médecins Sans Frontières Hospital in Kunduz exemplifies, unintentionally attacking civilians resulting from a failure to take all feasible precautions is commonly considered a violation of both the principle of precaution and the principle of distinction, often accompanied by a breach of the rules of engagement (RoE).⁵⁹

What about so-called ‘honest and reasonable’⁶⁰ mistakes that result in attacks on civilians, much like in Variant I of the illustrative scenario outlined in Section 2? It is plausible for a commander to comply fully with the duty to take all feasible precautions (including verification of the target to the extent

supporting the proportionality assessment, see Tomasz Zurek and others, ‘Computational Modelling of the Proportionality Analysis under International Humanitarian Law for Military Decision-Support Systems’, 14 January 2022, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4008946.

⁵⁶ International Law Association, Study Group on the Conduct of Hostilities in the 21st Century, ‘The Conduct of Hostilities and International Humanitarian Law: Challenges of 21st Century Warfare’ (2017) 93 *International Law Studies* 322, 382.

⁵⁷ For an in-depth examination of the last incident, see Marten Zwanenburg, Hans Boddens Hosang and Niek Wijngaards, ‘Humans, Agents and International Humanitarian Law: Dilemmas in Target Discrimination’, Conference Paper, 2005, https://static.aminer.org/pdf/PDF/000/313/514/analysis_on_negotiation_in_platform_level_armored_force_combat_entity.pdf.

⁵⁸ Among many tragic examples see, eg, the 2015 US air strike on the Kunduz MSF Hospital, which was mistaken for the HQ of the Afghan National Directorate of Security under Taliban control, and a variety of incidents reported during the ongoing conflict in Yemen. See, respectively US Forces-Afghanistan, ‘Investigation Report on the Airstrike on the Médecins Sans Frontières / Doctors Without Borders Trauma Center in Kunduz, Afghanistan on 3 October 2015’, 017-21, 28 April 2016 (Investigation Report on the Kunduz Airstrike 2016); and Human Rights Council, Report of the Group of Eminent International and Regional Experts on Yemen: Situation in Yemen, including Violations and Abuses since September 2014 (10 September 2021) UN Doc A/HRC/48/20 (UN Yemen Report 2021).

⁵⁹ Investigation Report on the Kunduz Airstrike 2016 (n 58) 93 para 114(a), 95 para 114(g)(5). Similarly, see UN Yemen Report 2021 (n 58) paras 22, 87.

⁶⁰ Milanovic (n 24) usefully conceptualises ‘honest and reasonable mistakes’ within the context of targeting as ‘after having taken all feasible precautions and measures to verify the nature of the target, an attacker pursues that target while honestly believing that he is attacking combatants/military objects, but it later transpires that in fact the target was civilian’.

possible, given its urgency and ostensibly hostile intent),⁶¹ follow the RoE to the letter, and yet end up making protected civilians, rather than combatants or civilians directly participating in hostilities, the object of the attack. Is such an attack a violation of the principle of distinction? It has been argued in scholarship that the existing state practice of providing compensation in such cases on an *ex gratia* basis without admitting legal responsibility suggests that honest and reasonable mistakes resulting in attacks on civilians 'are not regarded as violations of IHL'.⁶² This argument seems feeble, as many obvious IHL violations are compensated on exactly the same basis,⁶³ making it impossible to determine, based on the manner of payment, whether an incident was lawful. With state compensation policies fraught with ambiguities, it is worth looking at the issue from a more theoretical perspective, and inquiring whether the principle of distinction includes an embedded subjective element (and then at least some mistakes of fact would preclude the wrongfulness of its violations).⁶⁴ If not, is it a purely objective rule, the breach of which remains wrongful even if mistaken?

Some scholars maintain that the principle of distinction, as opposed to the grave breach of wilfully attacking civilians, is expressed in clearly objective terms.⁶⁵ The explicit inclusion of a *mens rea* requirement in listed grave breaches, so the argument goes, confirms the objective nature of the basic

⁶¹ In RoE of state parties to AP I, immediateness and hostile intent are usually the conditions that estop the presumption of civilian status as set forth by AP I (n 7) art 50(1). Such an interpretation seems a reasonable middle ground between troops protection and presumption of civilian status, as explicitly underlined in United Kingdom, 'Declarations and Reservations Made upon Ratification of the 1977 Additional Protocol I', 28 January 1998, para h (specifying the presumption of civilian character as applicable only 'in cases of substantial doubt still remaining after the assessment [of the information from all sources which is reasonably available to military commanders at the relevant time] has been made, and not as overriding a commander's duty to protect the safety of troops under his command or to preserve his military situation, in conformity with other provisions of [AP I]).

⁶² Milanovic (n 24).

⁶³ The 2015 air strike on Kunduz MSF Hospital, for instance, despite being, according to many, an obvious violation of IHL, was also compensated on an *ex gratia* basis, termed in US military parlance 'condolence payments'. On the payments in this and other cases see Joanna Naples-Mitchell, 'Condolence Payments for Civilian Casualties: Lessons for Applying the New NDAA', *Just Security*, 28 August 2018, <https://www.justsecurity.org/60482/condolence-payments-civilian-casualties-lessons-applying-ndaa>. Analysis of practice reveals that *ex gratia* payments are used, at least by some NATO member states, simply to remedy the harm caused for reasons of political expediency, irrespective of whether it was a violation of international law. For more see Amsterdam International Law Clinic, 'Monetary Payments for Civilian Harm in International and National Practice', October 2013, <https://ailc.uva.nl/binaries/content/assets/subsites/amsterdam-international-law-clinic/reports/monetary-payments.pdf>.

⁶⁴ Even the responsibility regimes that recognise a mistake of fact as a justification or excuse put forward some conditions that such mistakes ought to meet, the most conspicuous example being the Rome Statute of the International Criminal Court (entered into force 1 July 2002) 2187 UNTS 90, art 32(1) ('A mistake of fact shall be a ground for excluding criminal responsibility only if it negates the mental element required by the crime').

⁶⁵ This argument has been put forward in Lawrence Hill-Cawthorne, 'Appealing the High Court's Judgment in the Public Law Challenge against UK Arms Export Licenses to Saudi Arabia', 29 November 2018, <https://www.ejiltalk.org/appealing-the-high-courts-judgment-in-the-public>

principle of distinction, as worded in AP I, Articles 48, 51(2) and 52(2); AP II, Article 13(2), as well as Rule 1 of the ICRC Customary IHL Study.⁶⁶ Others assert that ‘the concept of *directing* attacks implies some level of intent, and that an honest and reasonable mistake of fact could negate that element of intent’.⁶⁷ It is the latter position that finds support, even if only modest, in jurisprudence and other forms of state practice.⁶⁸ Non-criminal case law on the nature of the principle of distinction remains meagre,⁶⁹ but in at least two separate cases the adjudicating bodies held quite unequivocally that the principle of distinction indeed includes a subjective element. In the 2005 *Partial Award on Western and Eastern Fronts*, the Eritrea-Ethiopia Claims Commission (EECC) found that:⁷⁰

[a]lthough there is considerable evidence of the destruction of civilian property by Eritrean shelling, ... the evidence adduced does not suggest an *intention* by Eritrea to target Ethiopian civilians or other unlawful conduct. ... [T]he Commission does not question whether this damage did in fact occur, but rather whether it was the result of unlawful acts by Eritrean forces, such as the *deliberate targeting of civilian objects* or indiscriminate attacks.

This view was echoed in the 2017 UK High Court ruling on the legality of arms exports to Saudi Arabia, in which the Court held that ‘[t]he “Principle of Distinction” prohibits *intentional* attacks against civilians’.⁷¹ While the latter

[law-challenge-against-uk-arms-export-licenses-to-saudi-arabia/#more-16674](https://www.cambridge.org/core/terms/https://doi.org/10.1017/S0021223722000188), and supported by a few scholars in the comments section.

⁶⁶ Jean-Marie Henckaerts and Louise Doswald-Beck (eds), *Customary International Humanitarian Law, Vol I: Rules* (ICRC and Cambridge University Press 2005, revised 2009) (ICRC Study) r 1 (‘The parties to the conflict must at all times distinguish between civilians and combatants. Attacks may only be directed against combatants. Attacks must not be directed against civilians’).

⁶⁷ Marco Milanovic and Sangeeta Shah, ‘Ukraine and the Netherlands v. Russia re MH17, Amicus Curiae on behalf of the Human Rights Law Centre of the University of Nottingham’, 12 February 2021, para 30, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3775402 (emphasis added).

⁶⁸ Note that decisions of national courts can be used to determine the existence of both state practice and *opinio juris*: International Law Commission, Draft Conclusions on Identification of Customary International Law, with Commentaries (2018), UN Doc A/73/10, Conclusions 4 and 10.

⁶⁹ For a reflection on mistakes of fact as to the protected status of the target albeit in the context of criminal responsibility see Rogier Bartels, ‘Discrepancies between International Humanitarian Law on the Battlefield and in the Courtroom: The Challenges of Applying International Humanitarian Law during International Criminal Trials’ in Mariëlle Matthee, Brigit Toebes and Marcel Brus (eds), *Armed Conflict and International Law: In Search of the Human Face - Liber Amicorum in Memory of Avril McDonald* (TMC Asser Press 2013) 339, 350 (asserting that ‘[n]otwithstanding the obligation to take all feasible precautions when launching an attack, certain attacks that have resulted in the death of and/or serious injury to civilians or those hors de combat would only constitute a violation of IHL if the attacker intended to cause the resulting harm to the persons protected by IHL’).

⁷⁰ EECC, *Partial Award: Western and Eastern Fronts - Ethiopia’s Claims 1&3*, 19 December 2005, *Reports of International Arbitral Awards*, vol XXVI, 351, para 74 (emphasis added).

⁷¹ *The Queen (on the application of Campaign Against Arms Trade) v Secretary of the State for International Trade* [2017] EWHC 1726 (QB) [208] (emphasis added).

can arguably be considered *obiter dictum*,⁷² the EECC's pronouncements are particularly pertinent to the issue at hand given the Commission's exceedingly rare mandate to adjudicate state responsibility for IHL violations.⁷³ Furthermore, some state practice reads intent into the principle of distinction as reflected in the official positions of, *inter alia*, Israel,⁷⁴ New Zealand,⁷⁵ and the United States,⁷⁶ and the lack of apparent contrary state practice and *opinio juris*.

Conceptually, not treating honest and reasonable mistakes resulting in targeting civilians as violations of IHL goes hand-in-hand with the so-called '*Rendulic rule*'⁷⁷ (assuming it applies to the determination of a breach in non-

⁷² As specified by the Court of Appeal, the core dispute before the Divisional Court concerned the definition of 'serious violations of IHL' and, in particular, whether it was synonymous with 'war crime', rather than the plain principle of distinction: *The Queen (on the application of Campaign Against Arms Trade) v Secretary of the State for International Trade* [2019] EWCA Civ 1020 [155].

⁷³ Agreement between the Government of the State of Eritrea and the Government of the Federal Democratic Republic of Ethiopia for the Resettlement of Displaced Persons, as well as Rehabilitation and Peacebuilding in Both Countries (entered into force 12 December 2000) 2138 UNTS 94, art 5 ('The mandate of the Commission is to decide through binding arbitration all claims for loss, damage or injury by one Government against the other, and by nationals (including both natural and juridical persons) of one party against the Government of the other party or entities owned or controlled by the other party that are (a) related to the conflict that was the subject of the Framework Agreement, the Modalities for its Implementation and the Cessation of Hostilities Agreement, and (b) result from violations of international humanitarian law, including the 1949 Geneva Conventions, or other violations of international law'). Note, however, that the EECC self-limited its jurisdiction to 'serious violations of IHL' only. For criticism see Gabriella Venturini, 'International Humanitarian Law and the Conduct of Hostilities in the Case Law of the Eritrea-Ethiopia Claims Commission' in Andrea de Guttry, Harry HG Post and Gabriella Venturini (eds), *The 1998-2000 Eritrea-Ethiopia War and Its Aftermath in International Legal Perspective* (Asser and Springer 2021) 345, 356-58.

⁷⁴ Ministry of Foreign Affairs of the State of Israel, 'The Operation in Gaza, 27 December 2008-18 January 2009: Factual and Legal Aspects', July 2009, para 110 (arguing that 'a commander's intent is critical in reviewing the principle of distinction during armed conflict').

⁷⁵ New Zealand, *Manual of Armed Forces Law*, vol 4, para 4.5.2 ('The obligation is dependent upon the information available to the commander at the time an attack is decided upon or launched. The commander's decision will not be unlawful if it transpires that a place honestly believed to be a legitimate military target later turns out to be a civilian object').

⁷⁶ US Department of Defense, *Law of War Manual*, para 5.4.3 ('Persons who plan, authorize, or make other decisions in conducting attacks must make the judgments required by the law of war in good faith and on the basis of information available to them at the time. For example, a commander must, on the basis of available information, determine in good faith that a target is a military objective before authorizing an attack against that target').

⁷⁷ US Military Tribunal at Nuremberg, *Hostages Case, United States v List (Wilhelm) and Others*, Trial Judgment, Case No 7 (1948) 11 TWC 757, 19 February 1948, para 194 ('The course of a military operation by the enemy is loaded with uncertainties, such as the numerical strength of the enemy, the quality of his equipment, his fighting spirit, the efficiency and daring of his commanders, and the uncertainty of his intentions. These things when considered on his own military situation provided the facts or want thereof which furnished the basis for the defendant's decision ... [T]he conditions, as they appeared to the defendant at the time were sufficient upon which he could honestly conclude that urgent military necessity warranted the decision made. This being true, the defendant may have erred in the exercise of his judgement but he was guilty of no criminal act').

criminal contexts), and the ancient legal maxim *impossibilia nulla obligatio est*.⁷⁸ Given all of the above, it is defensible to assert that directing attacks against civilians based on an honest and reasonable mistake of fact is not a violation of the principle of distinction. Assuming, for the sake of analysis, that in Variant I of the illustrative scenario no other precautions could feasibly have been taken, IHL was not breached. Without a breach no responsibility can possibly arise, without any further analytical finesse. Note, however, that the resulting ‘responsibility gap’ is anchored in IHL and is in no way affected by the introduction of the AI element into the equation. Whether the erroneous information as to the protected status of the target and the ensuing mistake can be traced back to the most advanced AI-powered system or human sentries is legally irrelevant; what matters is whether the commander complied with the obligation to verify the target before engagement.

This is the core difference between Variant I and Variant II. In the latter, quite clearly not everything practically feasible was done to verify the target, and a breach of both the principle of precaution and the principle of distinction did take place. Is it, however, attributable to state Alpha given that the attack was executed entirely independently by a FAWS? After briefly introducing the regime of state responsibility, the next section substantiates why the answer is in the affirmative.

4. No intent, no problem: The ABCs of state responsibility

The regime of state responsibility was authoritatively codified by the International Law Commission (ILC) in 2001 in the form of the Draft Articles on Responsibility of States for Internationally Wrongful Acts (ARSIWA).⁷⁹ These articles, widely considered customary in nature, in principle contain only the so-called secondary rules regulating ‘the general conditions under international law for the state to be considered responsible for wrongful actions or omissions’,⁸⁰ and not the primary norms specifying the content of obligations incumbent upon a state.⁸¹ ARSIWA are founded on a fundamental premise that ‘[e]very internationally wrongful act of a State entails the international responsibility of that State’.⁸² An internationally wrongful act –

⁷⁸ Latin for ‘inability excuses the law’. The principle has been relied on by various international courts, including the European Court of Human Rights (ECtHR) and the Court of Justice of the European Union (CJEU); see, respectively, ECtHR, *Béldné Nagy v Hungary*, App no 53080/13, 10 February 2015, para 53; and CJEU, Joint Cases C-622/16 P to C-624/16 P, *Scuola Elementare Maria Montessori Srl and European Commission v European Commission and Pietro Ferracci*, Opinion of Advocate General Wathelet, 11 April 2018, ECLI:EU:C:2018:229, para 105.

⁷⁹ ILC, Articles on the Responsibility of States for Internationally Wrongful Acts with Commentaries (2001), UN Doc A/56/10 (ARSIWA).

⁸⁰ *ibid* 31 para 1.

⁸¹ Among vast scholarship on the distinction between primary and secondary norms see, in particular, Eric David, ‘Primary and Secondary Rules’ in James Crawford, Alain Pellet and Simon Olleson (eds), *The Law of International Responsibility* (Oxford University Press 2010) 27–36 (providing a systematic overview of the dichotomy, its origin and pragmatic advantages in the development of international law).

⁸² ARSIWA (n 79) art 1.

that is, conduct which can consist of either action or omission – has two elements: first, it needs to be attributable to the state under international law; second, it must constitute a breach of an international obligation of that state.⁸³ These seemingly straightforward rules have a variety of consequences. In particular, unlike many domestic liability regimes, international responsibility of states is not premised on causation.⁸⁴ Instead, the crux of the whole regime is in the rules of attribution, conceptualised as ‘a pure result of the law’ pursuant to which ‘a will or an act are attributable to a given subject only because a legal provision says so’.⁸⁵ This very feature of the law of state responsibility makes it an objective regime, where – as opposed to international criminal law (ICL) – the mental state of the acting humans is, in principle, irrelevant.⁸⁶ What the two regimes have in common is the fact that (at least under the plain reading of the law as it stands today) only human conduct can lead to attribution of responsibility.⁸⁷ The so-called ‘attribution rules’ – set out in ARSIWA, Articles 4 to 10 – reflect the general rule pursuant to which ‘the only conduct attributed to the State at the international level is that of its organs of government [Article 4], or of others who have acted under the direction, instigation or control of those organs [Articles 5–10], i.e. as agents of the State’.⁸⁸

How do all these principles apply to Variant II of the scenario described in Section 2 – that is, a breach of the principle of precaution and distinction resulting from conduct executed entirely independently by a FAWS? Is it attributable to state Alpha, making it internationally responsible for the internationally wrongful act of killing the civilians? The answer is simply ‘yes’. Despite (F)AWS being frequently (and erroneously) anthropomorphised, under the plain reading of IHL it is those who ‘plan or decide upon the attack’ who are obliged to comply with the rules relating to conduct of hostilities. As rightly pointed out in recent scholarship, if those who decide upon the attack ‘cannot foresee that an AWS will engage only legal targets, then they cannot meet their obligations under the principle of distinction (API,

⁸³ *ibid* art 2.

⁸⁴ *ibid* Commentary to art 3 para 4 (‘The 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’). For more generally on causation in ARSIWA see León Castellanos-Jankiewicz, ‘Causation and International State Responsibility’, ACIL Research Paper No 2012-07 (*SHARES Series*), 2012.

⁸⁵ Dionisio Anzilotti, *Corso di diritto internazionale* (CEDAM 1955) 222, as translated in Francesco Messineo, ‘Multiple Attribution of Conduct’, *SHARES Research Paper No 2012-11*, 2012, 1, 5, <http://www.sharesproject.nl/wp-content/uploads/2012/10/Messineo-Multiple-Attribution-of-Conduct-2012-111.pdf>.

⁸⁶ ARSIWA (n 79) Commentary to art 2 para 10 (‘In the absence of any specific requirement of a mental element in terms of the primary obligation, it is only the act of a State that matters, independently of any intention’).

⁸⁷ *ibid* para 5 (‘an “act of the State” must involve some action or omission by a human being or group’).

⁸⁸ *ibid* Commentary to chapter II para 2.

article 57(2)(a(i)).⁸⁹ It is the commander who is 'ultimately responsible for accepting risk'⁹⁰ and 'in all cases ... has the responsibility for authorizing weapon release in accordance with IHL'.⁹¹ In Variant II, it was therefore the conduct of the Alpha commander – namely, the employment of FAWS in the first place – that was wrongful. Given that the commander undoubtedly constitutes an organ of state Alpha, their actions are attributable to that state, whether or not they exceeded their authority or contravened instructions.⁹² This suffices for establishing state Alpha's responsibility under ARSIWA; no culpability or causal link (between the physical action and the harm resulting from it) needs to be proved.

The existing international law applicable to combat use of (F)AWS clearly reflects a premise often recalled by the critics of the alleged 'responsibility gap' – namely, that '[n]o matter how independently, automatically, and interactively computer systems of the future behave, they will be the products (direct or indirect) of human behaviour, human social institutions, and human decision'.⁹³ Such an approach is obviously factually correct, but as the technology advances it is worth reflecting whether the commander and their conduct should remain the necessary link between the wrong caused by increasingly autonomous weapons and the responsibility of the state. As long as state responsibility hinges on the wrongful conduct of a commander, the latter would have to face disciplinary or even criminal charges for breaching IHL. In some cases, especially when the deployed FAWS, which passed the Article 36 AP I weapon review obligation, was particularly complex and hence difficult to understand,⁹⁴ and was deployed in a combat situation similar in all relevant respects to those for which it was tested but nonetheless malfunctioned, it seems unfair to have the commander face the military justice system.⁹⁵

The intuitive unfairness of such a burden carried by the commander raises the question of whether wrongful conduct resulting from the FAWS deployment could be attributed to the fielding state in another way, as it is beyond doubt that '[a] State should always be held accountable for what it does,

⁸⁹ Tim McFarland and Jai Galliot, 'Understanding AI and Autonomy: Problematising the Meaningful Human Control Argument against Killer Robots' in Jai Galliot, Duncan MacIntosh and Jens David Ohlin (eds), *Lethal Autonomous Weapons* (Oxford University Press 2021) 52.

⁹⁰ NATO, 'Allied Joint Doctrine for the Conduct of Operations', AJP-3 Edition C Version 1, 2019, 1–37.

⁹¹ US Mission to International Organizations in Geneva, 'Intervention on Appropriate Levels of Human Judgment over the Use of Force', paper presented at Technical Report Convention on Certain Conventional Weapons (CCW), Group of Governmental Experts on Lethal Autonomous Weapon Systems, Geneva, 15 November 2017, <https://geneva.usmission.gov/2017/11/16/u-s-statement-at-ccw-gge-meeting-intervention-on-appropriate-levels-of-human-judgment-over-the-use-of-force>.

⁹² ARSIWA (n 79) art 4, in concert with art 7.

⁹³ Deborah G Johnson, 'Computer Systems: Moral Entities but not Moral Agents' (2006) 8 *Ethics and Information Technology* 195, 197. See also Joanna J Bryson, 'Patience Is Not a Virtue: The Design of Intelligent Systems and Systems of Ethics' (2018) 20 *Ethics and Information Technology* 15, 21.

⁹⁴ The combination of commanders' continuing responsibility for weapons employment and the unfairness of imposing responsibility for systems that cannot be understood, resulting in the need for systems that operate in a transparent or predictable manner, is the reason that US DoD promulgated the traceability/understandability principle for AI systems; see US DoD (n 45) para 4.a.3.

⁹⁵ For a similar conclusion, see Amoroso and Tamburrini (n 46) 255.

especially for the responsible use of weapons which it delegated to the armed forces'.⁹⁶ Can the state simply be responsible for the weapons it fields? Interestingly, there is currently no general framework under international law that regulates state responsibility (or liability) for its inanimate objects; only self-contained, specific regimes exist – applicable, for example, to space objects⁹⁷ or transboundary harm arising out of hazardous activities.⁹⁸ There seems to be, however, a possible alternative avenue under ARSIWA that would allow for attributing wrongful conduct (such as the targeting of civilians) of FAWS to the state fielding it, but it has never been utilised in practice. The following discussion is thus entirely *de lege ferenda*.

A careful reading of ARSIWA and its Commentaries indicates that FAWS could be construed as a state agent.⁹⁹ The category of 'agent', while nowhere to be found in ARSIWA themselves, is mentioned frequently in the Commentaries (albeit without a definition), usually in the phrase 'organs or agents'.¹⁰⁰ The term 'agent', often used in older arbitral awards of the early twentieth century,¹⁰¹ was revived by the International Court of Justice (ICJ) in the *Reparations for Injuries* case in which the Court confirmed the responsibility of the United Nations for the conduct of its organs or agents, and underlined that it:¹⁰²

understands the word 'agent' in the most liberal sense, that is to say, any person who, whether a paid official or not, and whether permanently employed or not, has been charged by an organ of the organization with carrying out, or helping to carry out, one of its functions – in short, any person through whom it acts.

⁹⁶ Poland, 'Meaningful Human Control as a Form of State Control over LAWS', Lecture, The Convention on Certain Conventional Weapons, Informal Meeting of Experts on Lethal Autonomous Weapon Systems, UN Headquarters, Geneva, 13 April 2015.

⁹⁷ Treaty on Principles Governing the Activities of States in the Exploration and Use of Outer Space, including the Moon and Other Celestial Bodies (entered into force 10 October 1967) 610 UNTS 205, art VII ('Each State Party to the Treaty that launches or procures the launching of an object into outer space, including the moon and other celestial bodies, and each State Party from whose territory or facility an object is launched, is internationally liable for damage to another State Party to the Treaty or to its natural or juridical persons by such object or its component parts on the Earth, in air or in outer space, including the moon and other celestial bodies').

⁹⁸ ILC, Draft Principles on the Allocation of Loss in the Case of Transboundary Harm Arising out of Hazardous Activities, with Commentaries (2006), UN Doc A/61/10.

⁹⁹ For a similar reading of ARSIWA, see Boutin (n 14) 18.

¹⁰⁰ ARSIWA (n 79) Commentary to art 2 paras 3, 7; Commentary to art 7 paras 8, 10; Commentary to art 15 paras 5, 6; Commentary to art 31 para 12.

¹⁰¹ Several of which reiterated that 'a universally recognized principle of international law states that the State is responsible for the violations of the law of nations committed by its agents': *Reports of International Arbitral Awards*, vol XV, 399 (*Chiessa* claim), 401 (*Sessarego* claim), 404 (*Sanguinetti* claim), 407 (*Vercelli* claim), 408 (*Queirolo* claim), 409 (*Roggero* claim), and 411 (*Miglia* claim).

¹⁰² ICJ, *Reparation for Injuries Suffered in the Service of the United Nations*, Advisory Opinion [1949] ICJ Rep 174, 177. The same definition is repeated verbatim in ILC, Draft Articles on the Responsibility of International Organizations, with Commentaries (2011) UN Doc A/66/10, art 2(d).

That definition was admittedly created with a human agent in mind, but there is nothing in it – either verbatim or analytically – that would prevent its application to non-human persons, or simply objects, whether powered by AI or not.¹⁰³ Such an interpretation appears to be a relatively safe way forward, for two reasons. First, it does not compromise the integrity and coherence of the fundamental pillars on which ARSIWA are based. Second, it should not be controversial among states, which would simply bear responsibility for the wrongful conduct of their own objects, with the principles of attribution regulating the responsibility of state organs, applicable *mutatis mutandis*. In other words, it is suggested here that Article 4 of ARSIWA could be read to refer to ‘organs or agents’ and, as such, allow for the attribution of wrongful conduct caused by FAWS to the fielding state. Within the context of IHL, such an interpretation can be read in concert with Common Article 1 to the Geneva Conventions (I) to (IV),¹⁰⁴ the internal compliance dimension of which is firmly recognised as customary.

It is crucial to underline that the solution proposed herein to consider extending the category of agents to objects such as FAWS is strictly limited to the law of international responsibility of states for internationally wrongful acts.¹⁰⁵ In other words, viewing FAWS as agents is not meant to imply that they themselves could become subjects of international law, or be awarded some kind of moral agency in the ethical sense.¹⁰⁶

5. Tentative conclusions and way forward

Military AI and the delegation of tasks traditionally performed by humans to self-learning machines creates new challenges in a variety of fields, including on the international law plane. The goal of this analysis was to outline a counter-argument to those who lament the ‘responsibility gap’ that allegedly results from the employment of military AI on the battlefield. As demonstrated in the preceding sections, neither contemporary applications of AI nor their future ‘truly autonomous’ incarnations create any major conceptual hurdles under the law of state responsibility. The minor tweak thereto suggested here is intended merely to start the conversation on whether it is time to

¹⁰³ Compare the analysis in Boulanin and Verbruggen (n 47) 41–42 (alluding to such an interpretation).

¹⁰⁴ 1949 Geneva Convention (I) for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field (entered into force 21 October 1950) 75 UNTS 31, art 1 ([t]he High Contracting Parties undertake to respect and to ensure for the present Convention in all circumstances’).

¹⁰⁵ As such, it should not be conflated with the agency law of common law jurisdictions; nor should common law norms (such as tests to determine the existence of agency) be applied to the regime of international responsibility of states. Among many other understandings of agency, and using common law domestic concepts to explain states’ willingness to be bound by IHL, see, in particular, Eyal Benvenisti and Amichai Cohen, ‘War is Governance: Explaining the Logic of the Laws of War from a Principal-Agent Perspective’ (2014) 112 *Michigan Law Review* 1363.

¹⁰⁶ For an ethical analysis see Carissa Véliz, ‘Moral Zombies: Why Algorithms Are Not Moral Agents’ (2021) 36 *AI and Society* 487.

recognise that, in some cases, states should be internationally responsible for their objects, in a way similar to their responsibility for their organs. None of the above, however, should be read as implying that machines themselves can ever be held accountable. Nor does this article suggest that conceptualising agency in the realm of international responsibility as including objects is straightforward or constitutes a ready-made solution, either in general or for AI-enabled technologies in particular. On the contrary, further research is needed on at least three aspects.

First, it needs to be scrutinised what the *mutatis mutandis* application of attribution principles regulating the responsibility for state organs to non-human agents would entail.

Second, and related, it is important to bear in mind that ARSIWA are residual in nature, and, as such, can be displaced by a *lex specialis* regime, should they emerge.¹⁰⁷ Leaving aside the broader question of the normative desirability of *lex specialis* regimes of attribution,¹⁰⁸ an interesting inquiry could also be launched into whether state responsibility for the wrongdoings of its objects – modelled on either the Latin concept of *qui facit per alium facit per se*¹⁰⁹ or strict liability for damage caused by animals that is present in many domestic jurisdictions¹¹⁰ – could be conceptualised as a general principle of law within the meaning of Article 38(c) of the ICJ Statute.¹¹¹

Third and finally, more comprehensive research into the legal significance of mistakes of fact and fundamental principles of IHL is needed. The recent discourse remains so preoccupied with LAWS and the final stages of lethal targeting that the other uses of AI in military practice are largely overlooked. In particular, the growing incorporation of AI into intelligence, surveillance and reconnaissance (ISR) technologies – crucial for commanders' situational awareness and hence the proper application of all combat of hostilities rules – receives surprisingly little attention in legal scholarship. Faulty or incomplete intelligence is the most frequent cause of incidents resulting in outcomes that IHL was established to prevent. Should incorrect intelligence provided by an AI-powered ISR system be classified as a mistake of fact or a technical error? Or perhaps it is a distinction without difference within IHL and those two categories are legally the same? Harm that results from a technical error

¹⁰⁷ ARSIWA (n 79) art 55 ('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').

¹⁰⁸ For a cogent examination of existing *lex specialis* rules of attribution and a discussion of their desirability see Marko Milanovic, 'Special Rules of Attribution of Conduct in International Law' (2020) 96 *International Law Studies* 295.

¹⁰⁹ Latin for 'he who acts through another, acts himself'.

¹¹⁰ In the European civil context, a survey of relevant domestic law regulations with the view of potentially applying them to AI has already been partly carried out; see EPRS (n 12) 23–24. Interestingly, international law scholarship is increasingly interested in animals, even in the warfare context; see Anne Peters, Robert Kolb and Jérôme de Hemptinne, *Animals in the International Law of Armed Conflict* (Cambridge University Press, 2022 forthcoming).

¹¹¹ Statute of the International Court of Justice (entered into force 24 October 1945) 1 UNTS XVI. On utilising the concept in practice see M Cherif Bassiouni, 'A Functional Approach to "General Principles" of International Law' (1990) 11 *Michigan Journal of International Law* 768.

is traditionally not considered a breach of IHL, but blankly treating all cases of malfunction of military AI systems as ‘technical errors’ could be troublesome from a policy perspective.¹¹²

Alarming as LAWS and the idea of robots killing people can be,¹¹³ it is worth recognising that military AI goes beyond the trigger-pulling phase of the targeting process and raises important issues which have concrete consequences for implementation of IHL. It might therefore be preferable to leave the science fiction of FAWS to H(B)ollywood directors and focus on the still-unsettled core IHL issues.

Acknowledgements. The author would like to thank Katharine Fortin and the 4th Young Researchers Workshop on Terrorism and Belligerency participants for comments and feedback that gave form to the initial draft. Thanks are also owed to the Journal editors, Rogier Bartels, Remy Jorittsma and Marcin Krupa, as well as the Asser Institute DILEMA team for many discussions which helped to turn the raw idea into a polished product. Any mistakes, whether of law or fact, remain attributable solely to the author.

Funding. This research received funding from the European Union Horizon 2020 research and innovation programme under Marie Skłodowska-Curie Grant Agreement No 101031698. An earlier version of the article was presented during the 4th Young Researchers Workshop on Terrorism and Belligerency organised by the Minerva Center for the Rule of Law under Extreme Conditions at the University of Haifa Faculty of Law and the Geography and Environmental Studies Department.

Conflicts of interest. None.

¹¹² As contemporary technologically advanced conflicts, such as the Saudi-led coalition in Yemen aptly demonstrates, this ‘excuse’ is already abused; see UN Yemen Report 2021 (n 58) para 23 (‘The frequency with which JIAT finds “technical error” responsible for civilian losses without this leading to apparent changes in coalition procedures itself raises significant concerns as to the coalition’s commitment to meeting the requirements of international humanitarian law’).

¹¹³ In 2017, the Campaign to Stop Killing Robots released a short movie, *Slaughterbots*, which depicted ‘a dystopian future where terrorists could unleash swarms of tiny drones capable of identifying and killing specific people’: Michael Safi, ‘Are Drone Swarms the Future of Aerial Warfare?’, *The Guardian*, 4 December 2019, <https://www.theguardian.com/news/2019/dec/04/are-drone-swarms-the-future-of-aerial-warfare>.

Cite this article: Magdalena Pacholska, ‘Military Artificial Intelligence and the Principle of Distinction: A State Responsibility Perspective’ (2023) 56 *Israel Law Review* 3–23, <https://doi.org/10.1017/S0021223722000188>