April 2024

ELSA LAB
DEFENCE

No. 3

Research paper series

The Conceptual Roots of the Criminal Responsibility Gap in Autonomous Weapon Systems

AUTHOR

Jonathan Kwik

www.elsalabdefence.nl

THE CONCEPTUAL ROOTS OF THE CRIMINAL RESPONSIBILITY GAP IN AUTONOMOUS WEAPON SYSTEMS

JONATHAN KWIK*

One major reason for the controversy around autonomous weapon systems ('AWS') is the concern that no criminal liability is possible for resulting war crimes. This article takes a comprehensive look at one factor, the cognitive element of mens rea, and how and when characteristics specific to artificial intelligence ('AI') can render it more difficult to assign criminal liability to the deploying commander. It takes a multidisciplinary approach, considering both technical characteristics of modern AI and realistic conditions under which AWS are used. The article finds that modern AI primarily induces reduced perceivability through imperfect tracking of human intuition, opacity and generic reliability metrics. It also finds that AWS make it easier to willingly avoid acquiring cognition simply through inaction. Subsequently, it attempts to locate the exact loci of the problem within criminal law's spectrum of intent. This article finds that the epicentre of difficulty lies at the intermediate level of risk-taking, and particularly situations of generic risk: the condition where there is awareness only of a nondescript, indeterminate probability of 'something going wrong'. In contrast, no-gap situations are identified higher up the ladder of intent where there is purpose or virtual certainty, and judicious gaps lower down where we want 'impunity' for justified risk-taking and genuine accidents. Additionally, this article also considers the dangers of manufactured ignorance, where the risk can theoretically be known but in practice was not, due to a prior, separate omission. It ends with recommendations to address these challenges, including reducing opacity, standardising iterative investigations and enforcing technical trainings.

CONTENTS

I INTRODUCTION

The 'responsibility gap' is one of the oldest legal arguments raised against the adoption of autonomous weapon systems ('AWS').[1] The responsibility gap problem ('RGP') contends that the use of AWS is prohibited — or at least highly

* Jonathan Kwik is a postdoctoral researcher at the TMC Asser Institute in The Hague, affiliated with the ELSA Lab Defence project. Email: j.kwik@asser.nl; j.h.c.kwik@gmail.com.

[1] See Robert Sparrow, 'Killer Robots' (2007) 24(1) *Journal of Applied Philosophy* 62, 62.

problematic — because of the difficulties attached to allocating responsibility to persons for any harm that ensues from its use.[2] While it is not the only argument used by opponents of AWS technologies,[3] the RGP is a very popular one, and is often listed as one of the top four most common arguments for banning AWS in international discourse.[4] The RGP is, at its core, composed of two premises. First, that AWS possess characteristics — such as autonomy, complexity or intractability — which make attribution of responsibility to persons impossible; second, that an inability to assign responsibility for AWS use to a person makes their deployment unlawful. Let us call these Premise 1 and Premise 2. Both must be true to arrive at the conclusion desired by its authors, ie, that therefore, AWS should not be developed or used on the battlefield.

Note that the way the RGP was formulated in the above paragraph does not identify the *type* of responsibility in question. This is consistent with how the discussion has developed: authors have debated the RGP from the perspective of moral responsibility, tort and state responsibility, amongst others.[5] Nevertheless, the most common strand of the RGP is unquestionably related to criminal liability. In debates, reference is often made to scenarios of an AWS committing war crimes, followed by impunity because no person in the overall chain (from the programmer up to the commander who deployed the system) can fulfil the necessary requirements for criminal liability.[6]

---

[2] Darren M Stewart, 'New Technology and the Law of Armed Conflict' (2011) 87 *International Law Studies* 271, 289–92.

[3] Many other arguments that have been raised in opposition to AWS, which include questions about their accuracy, their ability to implement international humanitarian law ('IHL') and ethical considerations. For an overview of different arguments, see Michael W Meier, 'Lethal Autonomous Weapons Systems' in Winston S Williams and Christopher M Ford (eds), *Complex Battlespaces* (Oxford University Press, 2019) 289, 289–91; Amanda Musco Eklund, *Meaningful Human Control of Autonomous Weapon Systems: Definitions and Key Elements in the Light of International Humanitarian Law and International Human Rights Law* (Report No. FOI-R-4928-SE, February 2020) 2, 13; Masahiro Kurosaki, 'Toward the Special Computer Law of Targeting' in Claus Kreß and Robert Lawless (eds), *Necessity and Proportionality in International Peace and Security Law* (Oxford University Press, 2020) 409, 409–18.

[4] See, eg, Kenneth Anderson and Matthew C Waxman, 'Law and Ethics for Autonomous Weapon Systems: Why a Ban Won't Work and How the Laws of War Can' (Research Paper, Jean Perkins Task Force on National Security and Law Essay Series, Hoover Institution, Stanford University, 2013) 16–17; Christof Heyns, *Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions*, UN GAOR, 23rd sess, Agenda Item 3, UN Doc A/HRC/23/47 (April 2013) para 80; Carrie McDougall, 'Autonomous Weapon Systems and Accountability: Putting the Cart before the Horse' (2019) 20(1) *Melbourne Journal of International Law* 58, 72.

[5] See, eg, respectively, Isaac Taylor, 'Who Is Responsible for Killer Robots? Autonomous Weapons, Group Agency, and the Military-Industrial Complex' (2021) 38(2) *Journal of Applied Philosophy* 320, 322–3; Charles J Dunlap Jr, 'Accountability and Autonomous Weapons: Much Ado About Nothing?' (2016) 30(1) *Temple International and Comparative Law Journal* 63, 73–5; Rebecca Crootof, 'War Torts: Accountability for Autonomous Weapons' (2016) 164(6) *University of Pennsylvania Law Review* 1347, 1355–8.

[6] See, eg, Ian S Henderson, Patrick Keane and Josh Liddy, 'Remote and Autonomous Warfare Systems: Precautions in Attack and Individual Accountability' in Jens David Ohlin (ed), *Research Handbook on Remote Warfare* (Edward Elgar, 2017) 335, 357; Yoram Dinstein, 'Autonomous Weapons and International Humanitarian Law' in Wolff Heintschel von Heinegg, Robert Frau and Tassilo Singer (eds), *Dehumanization of Warfare: Legal Implications of New Weapon Technologies* (Springer, 2018) 20 [23]; Agnieszka Szpak, 'Legality of Use and Challenges of New Technologies in Warfare: The Use of Autonomous Weapons in Contemporary or Future Wars' (2020) 28(1) *European Review* 118, 126.

As the criminal liability aspect of the RGP is both the most prevalent and arguably the most problematic if left unresolved,[7] this article will exclusively focus on obstacles related to establishing criminal liability. As such, the two RGP premises and their conclusion can be reframed more specifically as:

P1. AWS possess characteristics which make the attribution of criminal liability to persons for harm caused by their deployment impossible.

P2. The inability to impose criminal liability on persons for harm caused by the deployment of an AWS makes them unlawful.

C. AWS are unlawful.

With regard to the second premise, various rationales are proposed to argue why the inability to assign criminal liability would impact a weapon's lawfulness. These include reference to the obligation in international humanitarian law ('IHL') to respect and ensure respect (which includes the duty to repress all serious breaches),[8] that impunity is inherently problematic,[9] that a lack of criminal consequences would damage IHL's ability to modify belligerent behaviour on the battlefield,[10] and that it would damage the right of victims to see justice done.[11] For the purposes of this article, let us grant the general notion that the inability to allocate criminal liability is undesirable without taking an explicit position with regard to Premise 2, and instead focus on Premise 1 — with one exception to be discussed in Part IV(C).

Premise 1, in its current formulation, may appear quite broad or generalising. What characteristics? Do *all* AWS possess these characteristics? Are they only technical characteristics, or do they also include emergent circumstances that arise from how the technology is used by humans? Do all levels of criminal intent become impossible to establish, or just certain ones? Does the jurisdiction matter? Yet, once again, this formulation merely tracks how the RGP is frequently presented in literature and discussions. It is often simply proclaimed that AWS, as a class of weapon, 'could malfunction, kill innocents, and nobody be held responsible',[12] or that 'issues with attributing accountability for war crimes committed by an AWS are insurmountable'.[13] The exact circumstances of why and how such criminal liability are often glossed over or discussed only briefly, which does a disservice to the RGP.

---

7 McDougall (n 4) 76; Thompson Chengeta, 'Accountability Gap: Autonomous Weapon Systems and Modes of Responsibility in International Law' (2016) 45(1) *Denver Journal of International Law and Policy* 1, 49.

8 *Geneva Convention Relative to the Protection of Civilian Persons in Time of War*, opened for signature 12 August 1949, 75 UNTS 287 (entered into force 21 October 1950) arts 1, 146 ('*Geneva Convention IV*'); *Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (Protocol I)*, opened for signature 8 June 1977, 1125 UNTS 3 (entered into force 7 December 1978) art 86(1) ('*API*').

9 M Cherif Bassiouni, 'Accountability for Violations of International Humanitarian Law and Other Serious Violations of Human Rights' in M Cherif Bassiouni (ed), *Post-Conflict Justice* (Transnational Publishers, 2002) 3, 26.

10 Stewart (n 2) 291–2; Chengeta (n 7) 12.

11 Chengeta (n 7) 49.

12 Mark Gubrud, 'Why Should We Ban Autonomous Weapons? To Survive' *IEEE Spectrum* (Web Page, 1 June 2016) <https://spectrum.ieee.org/why-should-we-ban-autonomous-weapons-to-survive>, archived at <https://perma.cc/DVL3-DY8H>.

13 Henderson, Keane and Liddy (n 6) 357.

As such, the purpose of the current article is to further dissect and refine this Premise 1 to arrive at a more focused — and limited — assertion as to the conditions which will result in this inability to attribute criminal liability. I do this by examining in closer detail both technical characteristics of modern AI which underlie AWS technology on one hand, and the more practical dimension of how wars are fought and under what conditions commanders make deployment decisions on the other. I subsequently juxtapose this with criminal law theory on mens rea and its many intricacies to determine the true loci of the gap. On the basis of this methodology, I demonstrate that far from all circumstances are problematic, either because (a) the responsibility gap does not exist, or (b) one does, but is 'judicious' — ie, anything *other* than a responsibility gap in that situation would be undesirable. This allows us to narrow our understanding of Premise 1 to only those circumstances where a *problematic* responsibility gap indeed does exist.

This exercise is to the benefit of both proponents and opponents of AWS technology. For opponents, who authored the RGP, a more limited but more realistic Premise 1 would strengthen their argument by not overstepping the argument's reach: precisely pinpointing how and why specific AWS technologies or usage would prevent criminal liability from being allocated provides a much stronger argument than broadly asserting that any and all AWS should be banned because they prevent accountability allocation, which is much less likely to convince detractors. For AWS proponents who are more commonly characterised as 'regulators' — ie, they wish to resolve potential problems through legislation, legal interpretation or new policy instead of prohibiting the technology outright — added clarity as to the particular points of challenge is equally valuable, as this will allow the adoption of more pointed and effective measures which specifically address these issues.[14]

This article proceeds as follows. First, in Part II, I briefly discuss the term 'AWS' and illustrate what type of scenario we have in mind when we speak of the RGP. In Part III, I then consider characteristics of modern artificial intelligence ('AI') used in AWS that risk generating a criminal responsibility gap. This section also considers how such AWS are likely to be used in a military setting and whether this can aggravate these problems. Establishing *cognition* is identified here as the primary threat. This is relevant as cognition is a core element of mens rea, requiring that the accused knew (or, for negligence, should have known) of a harmful consequence or the risk thereof. Subsequently, in Part IV, I apply these findings to the criminal law framework. For this section, I adopt a systematic approach based on the spectrum of intent levels available in criminal law, moving from the highest levels of deliberation to unforeseeable accidents. From this analysis, I argue that the focal point of the problem lies in scenarios related to *generic intermediate risk* — those situations where the accused is aware of *some* level of risk from their decision to deploy an AWS, but which is neither very high nor very low, nor specific. In contrast, for situations of high deliberation or accidents, I argue that they fall either within a 'no-gap' situation (where Premise 1 fails) or a 'judicious gap' situation (which attacks Premise 2) respectively. I also argue that there is a threat of *manufactured gaps,* where actors are incentivised

[14] For a comparison of the two approaches in the debate, see Eklund (n 3) 13; McDougall (n 4) 60–1.

'not to know'. The article concludes with some thoughts for future research and recommendations for legislation and policy.

II STAGE SETTING

The definition of AWS is a contentious subject and there is no universally accepted conception of the characteristics that make a weapon system fall under this label.[15] An entire body of literature is dedicated to debating what level of autonomy a weapon system must have, how complex the AI must be, and what task the weapon must perform to be considered an AWS.[16] For the purposes of this article, I will avoid these discussions on semantics,[17] and instead define an AWS as any kind of weapon which involves the delegation of a substantive targeting task(s)[18] to software installed in a weapon system, which carries out the task without further involvement of the deployer.

This definition is roughly meant to capture the following general scenario: A commander receives an AWS for use, analyses the upcoming military operation, and decides to activate the AWS to fulfil a specific goal (say, identify and neutralise enemy tanks in a specific district), after which the AWS is no longer under the commander's supervision.[19] Implicit in this deployment decision is the delegation of specific tasks related to targeting principles to the AWS, such as distinction and verification (the AWS should not attack civilian cars) and precautions (the AWS should delay a strike if the tank happens to drive by a very crowded market, and the missile's blast radius would hit the marketgoers). Note that 'delegation' here is not meant to reduce the commander's *legal* responsibility to ensure that such obligations are fulfilled: the commander is simply implementing these legal obligations *through* the AWS they deployed.[20]

---

15 Robin Geiß and Henning Lahmann, 'Autonomous Weapons Systems: A Paradigm Shift for the Law of Armed Conflict?' in Jens David Ohlin (ed), *Research Handbook on Remote Warfare* (Edward Elgar, 2017) 371, 375.

16 John Cherry and Durward Johnson, 'Maintaining Command and Control (C2) of Lethal Autonomous Weapon Systems: Legal and Policy Considerations' (2020) 27(1) *Southwestern Journal of International Law* 1, 6–9; Vincent Boulanin, 'Mapping the Development of Autonomy in Weapon Systems: A Primer on Autonomy' (Working Paper, Stockholm International Peace Research Institute, December 2016).

17 The undue focus on semantics has been rightfully criticised as being unproductive and distracting from more important questions: Marcel Dickow, 'Statement by iPRAW during the CCW GGE on LAWS: Human Control' (Speech, International Panel on the Regulation of Autonomous Weapons, 29 March 2019) 1, archived at <https://perma.cc/RQ2T-48VC>.

18 'Substantive targeting task' here refers to the implementation of one of the duties in IHL related to targeting, such as distinction, verification, proportionality or precautions: see *API* (n 8) art 51, 57. This should be distinguished from a targeting task in the technical sense, eg an onboard AI guiding a precision missile to a preselected target point.

19 As this article discusses the RGP, I ignore other arguments for why this might be illegal or unethical.

20 There is general consensus that commanders are ultimately responsible for applying IHL in the field, and that this cannot be ceded to a weapon: see Richard Moyes, 'Article 36 Statement on Human Control to the UN Discussions on Autonomous Weapons' (Speech, Convention on Certain Conventional Weapons — Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, 26 March 2019) <https://article36.org/updates/gge-2019-aws/>, archived at <https://perma.cc/J66H-9HBX>; International Committee of the Red Cross, *Ethics and Autonomous Weapon Systems: An Ethical Basis for Human Control?*, UN Doc CCW/GGE.1/2018/WP.5 (29 March 2018) 11 [32].

Ultimately it remains the case that 'the commander, not the weapon system, makes legal determinations'.[21]

This definition notably does not include conditions where the AWS remains under some form of human control after deployment. This type of situation is described in many different ways in literature, such as 'on-the-loop',[22] 'linked control',[23] 'supervisory control'[24] or 'semi-autonomous',[25] but essentially maintains a human in the decision-making loop as a supervisor to acknowledge or override decisions made by the machine. The methodological reason for this exclusion is practical: such systems are generally not viewed as likely to cause a responsibility gap. Indeed, maintaining such a supervisory structure is sometimes recommended as a *fix* to the RGP. For instance, Thompson Chengeta argues that a 'human should be in control of the system for each individual attack because such control is central to establishing the responsibility of combatants'.[26]

Compared to the circumstances described below in Part III, such a supervisory structure is indeed likely to address the RGP to a certain extent because the link between the machine's decisions and the supervising human's cognition and volition (necessary for criminal liability) is kept relatively narrow. Even if the AI is completely opaque[27] and the commander understands nothing of how the deployed system works for instance, they will still be held liable for wilfully targeting civilians[28] if they witnessed that their AWS was misidentifying a civilian automobile as a tank and failed to override the decision. Some authors rightly point out that such a structure still does not necessarily remove all RGP concerns due to psychological dynamics such as insufficient intervention time,[29] automation bias[30] and a lack of situational awareness.[31] However, for the purposes of the current discussion, this article shall take the assumption that these psychological problems have been properly addressed as part of the AWS's design. I thus exclude such systems from our discussion as they are simply not problematic. In subsequent

---

21 Richard J Sleesman and Todd C Huntley, 'Lethal Autonomous Weapon Systems: An Overview' [2019] (1) *Army Lawyer* 32, 34. Cf John P Sullins, 'When is a Robot a Moral Agent?' (2006) 6 *International Review of Information Ethics* 23. Positions that propose granting legal personality to the system, therefore, are not considered further in this analysis.

22 Oliver Müller, '"An Eye Turned into a Weapon": A Philosophical Investigation of Remote Controlled, Automated, and Autonomous Drone Warfare' (2021) 34(4) *Philosophy and Technology* 875, 886–9.

23 Jonathan Kwik, 'A Practicable Operationalisation of Meaningful Human Control' (2022) 11(3) *Laws* 43:1–21, 13.

24 Mary Cummings, 'Automation Bias in Intelligent Time Critical Decision Support Systems' (Conference Paper, AIAA Intelligent Systems Technical Conference, 20–22 September 2004).

25 Dinstein (n 6) 18 [12].

26 Thompson Chengeta, 'Defining the Emerging Notion of "Meaningful Human Control" in Weapon Systems' (2017) 49 *International Law and Politics* 833, 875.

27 See below Part III(A).

28 *API* (n 8) art 85(3)(a).

29 McDougall (n 4) 69–70.

30 Arthur Holland Michel, UN Institute for Disarmament Research, *Known Unknowns: Data Issues and Military Autonomous Systems* (Report, 17 May 2021) 16 <https://unidir.org/publication/known-unknowns>, archived at <https://perma.cc/77SZ-M86Q>; 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(2) *Journal of International Criminal Justice* 275, 295–8.

31 Bo (n 30) 296.

sections, I only consider the validity of the RGP with respect to systems where such a persistent link is not maintained.

As the topic of the RGP is quite extensive, it is necessary to set some limitations on what aspects of the discussion I will focus on in subsequent sections, to allow a truly exhaustive exploration of the facts. First, this article will take a commander-centric approach, ie, it will primarily examine whether the RGP remains valid vis-a-vis the commander who deployed the AWS. The reason for this focus is that there is some consensus in literature and practice that the deploying commander is primarily responsible for executing precautionary duties under IHL, making a risk analysis of such deployment and bearing the consequences for any harm that results.[32] As stated in the 2019 North Atlantic Treaty Organization Manual: 'The commander is ultimately responsible for accepting risk.'[33] If liability can be established for the deploying commander, then there is a priori no responsibility gap. Only if this is not possible will the attention of prosecutors likely shift to other, less directly involved actors such as programmers, manufacturers, etc. These will be discussed briefly in Part III(D), but they are not the focus of the current discussion.

Second, this article will predominantly discuss obstacles to establishing mens rea. Mens rea is the central pillar of criminal liability and ensures only those with a guilty mind are punished.[34] However, there are other requirements for criminal liability, such as an actus reus and causality, which have also been identified as potentially problematic. I will discuss causality problems briefly in Part III(D), and assume that establishing the actus reus is relatively unproblematic.[35]

III CHARACTERISTICS WHICH CHALLENGE CRIMINAL LIABILITY BUILDING BLOCKS

The RGP as a banning argument implies that there is something innate about AWS that make allocation of criminal liability difficult vis-a-vis older weapons. Indeed, this has been levied as a point of critique against overly broad variants of the RGP: without proper limitations on Premise 1, the same argument structure would lead to the conclusion that older, uncontroversial weapons such as

---

32 Jeffrey S Thurnher, 'Examining Autonomous Weapon Systems from a Law of Armed Conflict Perspective' in Hitoshi Nasu and Robert McLaughlin (eds), *New Technologies and the Law of Armed Conflict* (TMC Asser Press, 2014) 213, 226; Davies et al, *Air Force Operations and The Law* (The Judge Advocate General's School, 3rd ed, 2014) 19; *United Kingdom Expert Paper: The Human Role in Autonomous Warfare*, Agenda item 5, UN Doc CCW/GGE.1/2020/WP.6 (18 November 2020) 4 [10].

33 North Atlantic Treaty Organization, *Allied Joint Doctrine for the Conduct of Operations* (NATO Standardization Office, AJP-3 Edition C Version 1, February 2019) [1.75(a)].

34 Chengeta (n 7) 19.

35 Not all authors agree with this assessment. While the AWS physically carries out the actus reus, some point out that the 'voluntary' component of the actus reus requirement is necessarily not fulfilled in this case, as a machine's action would be more akin to a 'twitch' or 'reflex' than a willed, voluntary act: Pedro Miguel Freitas, Francisco Andrade and Paulo Novais, 'Criminal Liability of Autonomous Agents: From the Unthinkable to the Plausible' in Pompeu Casanovas et al (eds), *AI Approaches to the Complexity of Legal Systems: AICOL 2013 International Workshops* (Springer, 2014) 145, 152; Thomas C King et al, 'Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions' (2020) 26(1) *Science and Engineering Ethics* 89, 95.

landmines[36] and close-in weapon systems would also be unlawful because they would preclude the allocation of criminal liability.[37] Evidently, none of these weapons have triggered RGP controversies. Thus, Premise 1 must be refined by identifying specific characteristics of AWS that generate this responsibility gap, which were not present in older weapons. Let us call this the *retroactive disruption condition*, which states that the characteristics referred to in Premise 1 may not also be present in older, uncontroversial weapons so as to render them retroactively unlawful through the RGP.

In this section, I explore different kinds of characteristics which may prevent criminal liability from being established. As mentioned in Part II, I will focus in particular on factors which defeat the mens rea element, and in particular, its cognition component. It is argued that these come in three forms: characteristics inherent to data-driven AI systems, theoretically preventable factors but which may not always be addressed in practice, and the problem of generic risk. In addition, I will briefly touch upon other challenges not directly related to mens rea for completion, including the control problem and the problem of many hands. The purpose of the current section is to provide a solid foundation for our further exploration in Part IV as to where the loci of the gap lies within criminal law's spectrum of intent.

A *Modern AI and Reduced Perceivability*

When speaking of modern AI, *machine learning* ('ML') invariably comes to mind. ML is usually distinguished from more traditional rule-based AI, which were meticulously handcrafted by programmers.[38] ML, in contrast, uses large amounts of data to allow algorithms to essentially 'program themselves' by finding patterns and optimisations.[39] It has become a ubiquitous technique used in modern robotics and complex decision-making systems,[40] and will practically be unavoidable for the tasks we expect AWS to perform, such as independent navigation in a hostile environment, identification of targets and reacting dynamically to changing circumstances.[41] In this section, I discuss how the nature

---

36 In the sense of the RGP, as there are other reasons why anti-personnel landmines are controversial, such as their inability to distinguish between lawful targets: see especially *Convention on the Prohibition of the Use, Stockpiling, Production and Transfer of Anti-Personnel Mines and on their Destruction*, opened for signature 3 December 1997, 2056 UNTS 211 (entered into force 1 March 1999) Preamble.

37 Christopher P Toscano, '"Friend of Humans": An Argument for Developing Autonomous Weapons Systems' (2015) 8(1) *Journal of National Security Law and Policy* 189, 220.

38 Toshinori Munakata, *Fundamentals of the New Artificial Intelligence* (Springer, 2nd ed, 2008) 2; Yavar Bathaee, 'The Artificial Intelligence Black Box and the Failure of Intent and Causation' (2018) 31(2) *Harvard Journal of Law and Technology* 889, 898–9; Cristoph Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* (Lean Publishing, 2019) 12.

39 Joint Research Centre, European Commission, *AI Watch: Defining Artificial Intelligence* (Technical Report, February 2020) 11 <https://publications.jrc.ec.europa.eu/repository/handle/JRC118163>, archived at <https://perma.cc/8WRM-BH3B>.

40 Marco Tulio Ribeiro, Sameer Singh and Carlos Guestrin, '"Why Should I Trust You?": Explaining the Predictions of Any Classifier' (Conference Paper, ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13 August 2016).

41 Office of the Chief Scientist, United States Air Force, *Autonomous Horizons: System Autonomy in the Air Force — A Path to the Future* (Technical Report No AF/ST TR 15-01, June 2015) 22.

of ML raises challenges to establishing mens rea's cognition element. This is often referred to generally as the *epistemic problem*.[42]

ML techniques — particularly deep neural networks — produce flexible, powerful and well-performing systems,[43] but come with some downsides. The most notable is what I will term *reduced perceivability*. This is an umbrella concept which refers to a general reduction in the ability of humans to understand, foresee and anticipate exactly how the AI will act. This is a well-known limitation of ML systems, and can be contrasted with older, handcrafted symbolic systems, which have 'one major virtue: it is always clear why the machine makes the choice that it does, because its designers set the rules'.[44] In this sense, ML systems can occasionally make decisions which, from a human perspective, could be described as erratic. One famous 2016 example is that of a Tesla vehicle which misclassified the side of a white truck crossing the highway as part of the sky, and rammed its driver straight into the truck, killing him.[45] Could the driver have foreseen this failure? Perhaps he could have been on notice that the system was not meant to be fully automated and required human oversight,[46] but it is inconceivable that he could have known about the *specific* trigger that led to the deadly crash, namely that the AI could not distinguish the white side of a truck from a spring sky. This accident is quite illustrative of how modern AI systems are less perceivable. To obtain a better understanding of this phenomenon, however, let us discuss in a little more depth the various technical, design and production factors that make this the case.

The first factor which causes lowered perceivability is that ML systems *approximate*, but do not exactly *track*, human intuition. A majority of the time, a well-programmed ML system will produce outputs which agree well with our human judgements, and often may even exceed human accuracy levels.[47] The issue arises with those situations where mistakes are made. While both humans and machines make mistakes — stories of soldiers misidentifying civilians as combatants are plentiful — AI do not make mistakes in the same way we do. To return to the Tesla case as an example, we would have understood the failure much better if, for instance, the car was blinded by a frontal sundown, as this is also how human drivers often cause accidents. It is relatively simple to intuit that a frontal sun will reduce visibility and increase the risk of crashes, and that more caution

---

42 Johannes Himmelreich, 'Responsibility for Killer Robots' (2019) 22(3) *Ethical Theory and Moral Practice* 731, 743.

43 Roman V Yampolskiy, 'Unexplainability and Incomprehensibility of Artificial Intelligence' (2020) 7(2) *Journal of Artificial Intelligence and Consciousness* 277.

44 Boer Deng, 'The Robot's Dilemma: Working out How to Build Ethical Robots Is One of the Thorniest Challenges in Artificial Intelligence' (2015) 523 *Nature* 25, 25.

45 Danny Yadron and Dan Tynan, 'Tesla Driver Dies in First Fatal Crash While Using Autopilot Mode', *The Guardian* (online, 1 July 2016) <www.theguardian.com/technology/2016/jun/30/tesla-autopilot-death-self-driving-car-elon-musk>.

46 Which was the case in this accident, as later investigations placed the fault at the driver for not properly paying attention to the road despite the fact that the autopilot was not intended to function entirely without human oversight: Neal E Boudette, 'Tesla's Self-Driving System Cleared in Deadly Crash', *The New York Times* (online, 19 January 2017) <www.nytimes.com/2017/01/19/business/tesla-model-s-autopilot-fatal-crash.html>.

47 Yampolskiy (n 43) 2; Logan Engstrom et al, 'A Rotation and a Translation Suffice: Fooling CNNs with Simple Transformations' (Conference Paper, International Conference on Machine Learning, June 2019) 2.

should be exercised (eg by driving slower or temporarily stopping at a rest area). In comparison, the inability of the AI to distinguish the side of a truck from the sky is baffling to us. The problem is not that ML systems necessarily make more mistakes than humans, but that they make those mistakes *differently*.[48] This 'befuddlement' has evident consequences for a user's ability to predict and anticipate in advance the risks related to employing ML systems — a critical aspect for establishing the cognitive aspect of mens rea. It is conceptually problematic to blame someone for a consequence they did not (or even *could* not) foresee.[49]

Compounding the difficulties caused by this 'different way of thinking' is the fact that it may sometimes not even be possible to determine how it thinks differently, even if one would want to. This may be caused by *opacity*. Opacity, often also referred to as the 'black box phenomenon', refers to the situation where we cannot understand an AI's internal workings nor trace how decisions are made.[50] While inputs and outputs are available, we are unable to see the internal rationale that led to these pairings, or what may cause them to change.[51] This applies even to the system's designers, who 'themselves often do not understand how their systems work'.[52] It is important to emphasise that the operative word is *cannot*: cognition is a technical *impossibility*, instead of something caused by, eg, a commander's negligence or lack of technical knowledge. Yavar Bathaee links opacity directly to intent in a criminal law sense, theorising that it can 'functionally [immunise] the user of the AI from liability'.[53] Justifiably, opacity has been identified by commentators as an obstacle to establishing criminal liability for war crimes caused by AWS.[54] While opacity can be reduced via techniques such as Explainable AI ('XAI'),[55] it is far from a guarantee that all future AWS will incorporate XAI measures.[56]

---

48 See Defense Innovation Board, *AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense* (Report, 31 October 2019) 3; Matthew U Scherer, 'Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies' (2016) 29(2) *Harvard Journal of Law and Technology* 353, 364.

49 Thomas Weigend, 'Subjective Elements of Criminal Liability' in Markus D Dubber and Tatjana Hörnle (eds), *The Oxford Handbook of Criminal Law* (Oxford University Press, 2014) 490, 490.

50 High-Level Expert Group on Artificial Intelligence, European Commission, *A Definition of AI: Main Capabilities and Disciplines* (Report, 18 December 2018) 6.

51 Todd Kulesza et al, 'Principles of Explanatory Debugging to Personalize Interactive Machine Learning' (Conference Paper, International Conference on Intelligent User Interfaces, 18 March 2015) 126, 127.

52 Shane T Mueller et al, 'Explanation in Human-AI Systems: A Literature Meta-Review Synopsis of Key Ideas and Publications and Bibliography for Explainable AI' (Literature Review, DARPA XAI Program, February 2019) 63.

53 Bathaee (n 38) 921.

54 Jonathan Kwik and Tom van Engers, 'Algorithmic Fog of War: When Lack of Transparency Violates the Law of Armed Conflict' (2021) 2(1–2) *Journal of Future Robot Life* 43, 56–7.

55 For a primer, see Amina Adadi and Mohammed Berrada, 'Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)' (2018) 6 *IEEE Access* 52138.

56 This is particularly true as XAI is often inversely related to performance, meaning that designers of AWS will have to sacrifice military efficiency to reduce opacity — a sacrifice not all producers may be willing to make: ibid 52143–4.

B *Preventable Obstacles to Cognition*

A few other factors have been raised in literature as possibly aggravating the lack of cognition. One is *online learning*, which allows the AI to continue optimising its algorithm after deployment to improve performance over time, but at the cost of predictability.[57] Evidently in such a situation, humans may not be able to foresee how the algorithm evolves, generating further epistemic difficulties.[58] For some weapon systems like swarms, the interaction between individual agents may give rise to a higher variation and probability of unexpected failures, which a commander may not be able to predict.[59] Another possible epistemological challenge is the *lack of technical expertise* on the part of the deploying commander. Commanders are not AI experts and it is unreasonable to expect them to be. However, this means that they will be even less adept at recognising the risks of input data issues,[60] algorithmic bias,[61] datashift[62] and incorrect proxies,[63] amongst others, which may cause AI failures. If such a threat was not recognised because of the technical expertise this requires, then it cannot be said that the accused had the requisite cognition necessary for mens rea.

As indicated by this subsection's title, these epistemic problems are more preventable than those in the previous subsection (which are inherently baked into ML technology). The ability to continue learning can be removed by design or toggled on or off by the user (if such a setting is available). Indeed, this is often recommended by commentators in order to avoid the RGP.[64] Similarly, a commander could take more time to familiarise themselves with the AI technology installed in their AWS to better take abovementioned failure triggers into account. The question is, however, *whether the necessary steps are, in practice, taken* to remedy these problems. If such steps are not taken, then cognition cannot be proven by prosecutors. While the prosecution could argue that the commander *should* have known if they had toggled off online learning or taken AI courses,

---

57 Kevin Nelson, George Corbin and Misty Blowers, 'Evaluating Data Distribution and Drift Vulnerabilities of Machine Learning Algorithms in Secure and Adversarial Environments' in Misty Blowers and Jonathan Williams (eds), *Proceedings Volume 9119, Machine Intelligence and Bio-Inspired Computation: Theory and Applications VIII* (2014) 1, 1–2.

58 Taylor (n 5) 325.

59 Swarms are often deemed advantageous for their fault tolerance, since the failure of one unit can be compensated by other units in the swarm: see, eg, Matt Luckcuck et al, 'Formal Specification and Verification of Autonomous Robotic Systems: A Survey' (2019) 52(5) *ACM Computing Surveys* 100:1, 13–14. While this may be advantageous from an efficiency perspective, if this one unit's failure should result in a civilian casualty, it remains a concern of IHL and criminal law.

60 Holland Michel (n 30) 3–4.

61 Alejandro Barredo Arrieta et al, 'Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI' (2020) 58 *Information Fusion* 82, 104.

62 Alex A Freitas, 'Comprehensible Classification Models: A Position Paper' (2014) 15(1) *SIGKDD Explorations Newsletter* 1, 2.

63 See Dario Amodei et al, 'Concrete Problems in AI Safety' (Paper, arXiv, 21 June 2016) <http://arxiv.org/abs/1606.06565>, archived at <https://perma.cc/FAP7-Y86K>.

64 Kwik (n 23) 12; William H Boothby, 'Highly Automated and Autonomous Technologies' in William H Boothby (ed), *New Technologies and the Law of War and Peace* (Cambridge University Press, 2019) 137, 151.

this wording betrays the mens rea mode available for such an argument: it would amount to a negligence claim, not intent.[65]

To illustrate this issue, let us discuss a more flagrant circumstance where the necessary steps are *not* taken to enable cognition: *iteration*. Say that during a prior deployment of a sister model[66] in a different town, an accident occurred: an unexpected edge case surfaced in the form of a minibus painted in bright pink, and the AWS mistakenly identified it as an enemy tank. From this point onward, this edge case is no longer unknown or unexpected: humans can now foresee that the AWS will mislabel pink minibuses as tanks. Armed with this information, a commander who releases a sister AWS into a city, knowing a festival with hipster buses is taking place, could reasonably be characterised as possessing the requisite knowledge for attacking civilians. However, this knowledge is *only* available if iterative steps were taken to investigate the prior accident and communicate the findings to other commanders in possession of the same model.[67] If such measures are not taken, the commander can still not be said to possess the necessary cognition for mens rea, even if the edge case is *theoretically* known.

This finally brings us to a manufactured threat to cognition: the *wilful avoidance* of it. As mentioned, the problems presented in this subsection can be avoided if steps are taken to this effect. However, if taking no action actually reduces the chances of criminal responsibility for harmful results, there is no reason to take these extra steps. To the contrary, actors may be *incentivised* not to know.[68] This was astutely remarked by Rebecca Williams at the House of Lords: the cognition element in mens rea combined with the complexity of AI 'provides a great incentive for human agents to avoid finding out what precisely the ML system is doing, since the less the human agents know, the more they will be able to deny liability for both these reasons'.[69] This suggests external pressures[70] are needed to enable the necessary cognition of commanders, such as requirements for AI training or military policy that AWS accidents be investigated as quickly as possible and communicated to other commanders.

---

65 See below Part IV(D).

66 'Sister model' is meant to indicate that the AI's internals are identical. They would react identically when provided with the same inputs.

67 IHL does mandate states to take appropriate measures to investigate and suppress any violations during operations: *API* (n 8) art 86(1). To do so, states have implemented after-action procedures: see, eg, Department of the Army Headquarters, *Protection of Civilians* (Army Techniques Publication No 3-07.6, 29 October 2015) [5-74]. However, it is not guaranteed that such investigations or analyses are always conducted by all actors using AWS.

68 See, eg, Keith J Hayward and Matthijs M Maas, 'Artificial Intelligence and Crime: A Primer for Criminologists' (2021) 17(2) *Crime, Media, Culture* 209, 217.

69 Evidence to the House of Lords Select Committee on Artificial Intelligence, United Kingdom Parliament, London, 8 September 2017 (Rebecca Williams) <http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/artific ial-intelligence-committee/artificial-intelligence/written/70496.html#_ftn13>, archived at <https://perma.cc/A9XQ-298X>.

70 Notably, IHL does not explicitly require technical training for commanders, even though some authors have derived such a duty from other obligations such as precautions: Marco Longobardo, 'Training and Education of Armed Forces in the Age of High-Tech Hostilities' in Elena Carpanelli and Nicole Lazzerini (eds), *Use and Misuse of New Technologies: Contemporary Challenges in International and European Law* (Springer, 2019) 73, 77–81; Isabel Robinson and Ellen Nohle, 'Proportionality and Precautions in Attack: The Reverberating Effects of Using Explosive Weapons in Populated Areas' (2016) 98(1) *International Review of the Red Cross* 107, 140.

C *The Problem of Generic Risk*

One can argue at this point that establishing mens rea need not necessarily require complete and exact perceivability from the part of the decision-maker, nor expert-level understanding of the weapon. A commander, for instance, can still be held responsible for executing an indiscriminate attack with an airborne virus even if they have no background in pathology,[71] or for deliberately targeting civilians if they attack a small target within a densely populated city with artillery known to be statistically inaccurate.[72] In that sense, ML's reduced perceivability does not prevent humans from obtaining an *overall* idea of how well the system operates on a macro scale. Based on large-scale testing, for instance, producers may be able to present a commander with a relatively reliable indicator of the AI's performance metrics, eg, '99% accuracy rate in sunny weather'. In this respect, these performance metrics can be compared to older ways to describe a weapon's reliability, such as circular error probable ('CEP') for artillery systems.[73] If this is true, then we stumble upon a problem with regard to our retroactive disruption condition. If general reliability standards such as CEPs are enough for courts to hold that the deployer had enough knowledge of the weapon's workings to establish mens rea,[74] is this so different with regard to AWS accuracy statistics? It is argued that a key distinguishing element does exist, which once again relates to the imperfect way AI tracks our intuition and understanding of the world.

To demonstrate this, let us consider a more concrete scenario, building upon our previous example in Part II. Say that there are two commanders A and B, one with a rocket system and another with an AWS, who each hope to disable enemy tanks within a city they wish to capture. Both were given performance indicators in the form of CEP or accuracy metrics and must then decide whether or not to deploy their weapons in light of factual operational conditions (let us assume arguendo that intelligence is complete and accurate, meaning that both commanders have perfect information). One could say that the subsequent decision-making process will be quite comparable. The commanders will consider the probability of errors, what effect this will have on the civilian population and civilian objects around the target tanks,[75] and decide whether this risk is reasonable. Say now that both commanders consider that the risk is acceptable.[76] They order the deployment of the weapons, but something goes wrong.

---

71 Peter Margulies, 'Making Autonomous Weapons Accountable: Command Responsibility for Computer-Guided Lethal Force in Armed Conflicts' in Jens David Ohlin (ed), *Research Handbook on Remote Warfare* (Edward Elgar, 2017) 405, 412–13.

72 See, eg, *Prosecutor v Galić (Judgement and Opinion)* (International Criminal Tribunal for the Former Yugoslavia, Trial Chamber I, Case No IT-98-29-T, 5 December 2003) [64]–[65], [769] ('*Galić Trial Judgement*').

73 See, eg, PAX and Article 36, *Areas of Harm: Understanding Explosive Weapons with Wide Area Effects* (Report, October 2016) 14–18.

74 See, eg, *Prosecutor v Martić (Judgement)* (International Criminal Tribunal for the Former Yugoslavia, Appeals Chamber, Case No IT-95-11-A, 8 October 2008) [256].

75 *API* (n 8) arts 51, 52.

76 We leave aside specific details as to whether this analysis was *objectively* correct from a targeting perspective and focus only on the commander's subjective conviction that the risk is small enough to justify an attack.

- A's initial rocket deviates wildly and hits a market on the edge of the projected CEP range, causing massive civilian casualties. According to the CEP table, this had a 20% chance of occurring.
- B's AWS temporarily glitches because of a reflection of sunlight off a church's stained-glass windows and fires a missile against a market, causing massive civilian casualties. Later analysis suggests this failure forms part of the 20% of edge cases noted in the AWS's manual.

In both scenarios, the commanders 'knew' of the 20% risk of something going wrong in terms of general statistics. However, it is argued that their ability to foresee *what is entailed* by such an error rate is very different. An artillery commander has a very limited set of possibilities to consider as to what could result from the given error rate manifesting. Invariably, the rocket will land *somewhere*. Thus, if the area surrounding the target location is predominantly civilian housing, they 'know' they are taking a 20% risk of hitting said civilian concentrations. In contrast, what will happen when an AWS fails is less clear, even if we know that it is unreliable 20% of the time. B is aware of *a* risk, but not any risk in particular. It might attack a personnel carrier, a bus, a horse-drawn cart, a building, friendly forces — or do nothing at all. Note that in the example above, the attacked target was a market — something that in our human eyes, looks nothing like a tank. Thus, even if B correctly took into consideration the risk of error to civilian vehicles, it is unlikely they could have anticipated this particular counterintuitive misclassification. I will call this the *problem of generic risk*: even if humans obtain general knowledge of the risk of failure, it is more difficult to associate this risk with a particular harmful consequence. The consequences of this problem for establishing mens rea will be discussed in Part IV(B).

D *Characteristics Not Directly Related to Cognition*

There are two other factors commonly raised in literature which, it is often argued, make attribution of criminal liability for AWS harm more challenging. These are the *control problem*[77] and the *problem of many hands*.[78] As this article focuses primarily on the impact of AWS characteristics on mens rea, I will only briefly discuss these two other problems for the sake of completion, as they would still be necessary for an exhaustive formulation of Premise 1.

The *control problem* asserts that criminal liability cannot be established because AWS are not under the accused's control when the deed is done.[79] Control is not formally an element a prosecutor must prove, but is in the view of many a fundamental component of criminal liability: that of only punishing blameworthy conduct.[80] In one of the first major publications discussing the responsibility gap, Andreas Matthias asserts that 'for a person to be rightly held responsible, that is, in accordance with our sense of justice, she must have control over her behaviour

---

77 See, eg, Himmelreich (n 42).

78 Dennis F Thompson, 'Moral Responsibility of Public Officials: The Problem of Many Hands' (1980) 74(4) *American Political Science Review* 905.

79 Himmelreich (n 42) 743.

80 Geiß and Lahmann (n 15) 393–4.

and the resulting consequences'.[81] Indeed, this is one of the reasons why meaningful human control has become such a major focus in the debate on AWS: retaining control is viewed as one way to ensure that criminal liability can still be established.[82]

The second problem, the *problem of many hands*, refers to the proliferation of persons unto whom we can attribute responsibility.[83] It was coined originally to refer to situations where a harmful outcome was 'caused' by multiple decision-makers within a large organisation (or across several of them).[84] The theory was subsequently found to be particularly applicable to software development,[85] which invariably involves many different programmers, designers, labellers, quality control staff, etc. Each can potentially contribute toward the ultimate harmful outcome. With regard to AWS, the pool of persons involved expands even further to include, inter alia, the officials who ordered the weapons, weapon reviewers, and policymakers.[86] The problem of many hands is often associated with causality problems (isolating the proper 'cause' of the harm in such a distributed setting) as well as evidentiary challenges (eg proving such causality),[87] but also touches upon the epistemological problem.[88] Is it reasonable, for instance, to assert that a programmer 'knew' that the subsystem they trained would fail one year later because of an unexpected interaction with another subsystem developed by a different department? This is very hard to establish, and even harder to prove in court. Note that the problem of many hands, by definition, is primarily an obstacle for establishing criminal liability for actors earlier in the weapon's lifecycle; it would be less relevant for the liability of the deploying commander, as they 're-centre' the causal network back unto one decision-making figure who makes the final decision to deploy the weapon.[89]

IV ACROSS THE SPECTRUM OF INTENT: NO-GAP, JUDICIOUS GAP AND TRUE GAP SITUATIONS

In light of the concrete challenges discussed in Part III, let us now attempt to determine the loci of the problem in terms of criminal law proper. As this article

---

81 Andreas Matthias, 'The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata' (2004) 6(3) *Ethics and Information Technology* 175, 175 (emphasis omitted).

82 Kwik (n 23) 15–16.

83 See Alice Giannini and Jonathan Kwik, 'Negligence Failures and Negligence Fixes. A Comparative Analysis of Criminal Regulation of AI and Autonomous Vehicles' (2023) 34 *Criminal Law Forum* 43, 58–59.

84 Thompson (n 78).

85 Helen Nissenbaum, 'Accountability in a Computerized Society' (1996) 2(1) *Science and Engineering Ethics* 25, 28–32.

86 Daniele Amoroso and Benedetta Giordano, 'Who Is to Blame for Autonomous Weapons Systems' Misdoings?' in Elena Carpanelli and Nicole Lazzerini (eds), *Use and Misuse of New Technologies: Contemporary Challenges in International and European Law* (Springer, 2019) 211, 216.

87 Nissenbaum (n 85) 29; Chantal Grut, 'The Challenge of Autonomous Lethal Robotics to International Humanitarian Law' (2013) 18(1) *Journal of Conflict and Security Law* 5, 16.

88 Ibo van de Poel et al, 'The Problem of Many Hands: Climate Change as an Example' (2012) 18(1) *Science and Engineering Ethics* 49, 61.

89 Cf Daniele Amoroso, *Autonomous Weapons Systems and International Law: A Study on Human-Machine Interactions in Ethically and Legally Sensitive Domains* (Nomos Verlagsgesellschaft, 2020) 127.

has focused on mens rea problems, I will do so by examining the spectrum of intent in criminal law, taking us from the highest levels of deliberation down to situations where knowledge is impossible. I submit that there are in actuality three situations that can be distinguished: *no-gap* situations, where Premise 1 simply fails; *judicious gap* situations, where there is technically a responsibility gap in the sense that mens rea cannot properly be established, but where this responsibility gap is justified (which negates Premise 2); and *true gap* situations, which constitute the sole scenarios where both Premises 1 and 2 hold. Future discussions and policy proposals would thus do well to focus on these problematic loci.

A *No-Gap Situations: Purpose or Virtual Certainty*

The best illustration that Premise 1 is limited and does not apply comprehensively to AWS use are situations of direct intent or knowledge. Evidently, like all tools, AWS can be used with the intention to commit war crimes,[90] just like other algorithms have been used to commit cyberattacks, financial fraud, identity theft, forgery, phishing and market manipulation, and make deepfakes.[91] This can be done purposely or knowingly, and in either case, there is no responsibility gap.

In the case where an AWS is used to commit a war crime *purposely*, the actual probability of error is irrelevant. Purpose, or *dolus directus* in continental systems, can be established even if the actual risk of the actus reus manifesting is very low or unknown to the perpetrator:[92] it is the perpetrator's *volition* which is determinative.[93] Thus, if there is evidence that a commander desires a certain ethnic group to be killed by an AWS they deploy, their lack of knowledge about the AI's internals would not bar their responsibility for deliberately willing the ethnic group's extermination. Similarly, the racist programmer who deliberately leaves in an algorithmic bias which always categorises members of a particular ethnic group as combatants will act purposely, even if they may not know exactly when the final products will encounter members of this group.[94] There is little controversy in literature that this is a no-gap situation. 'Criminal culpability is self-evident in the case of intent.'[95]

The inverse situation, where there is no direct intent but very clear awareness of the consequences despite the challenges discussed in Part III, should also not present a gap. In many legal systems, knowingly causing an actus reus (in continental systems: *dolus indirectus* or *dolus directus* in the second degree) also qualifies as intent.[96] 'Knowledge' in this sense 'is a subjective, practical certainty

---

90 Amoroso and Giordano (n 86) 217.

91 Hayward and Maas (n 68) 215–17.

92 Bernard E Gegan, 'More Cases of Depraved Mind Murder: The Problem of Mens Rea' (1990) 64(3) *St. John's Law Review* 429, 447.

93 Weigend (n 49) 495.

94 Amoroso and Giordano presented a similar example, but with a discriminatory arms dealer: Amoroso and Giordano (n 86) 218.

95 Nikolas Stürchler and Michael Siegrist, 'A "Compliance-Based" Approach to Autonomous Weapon Systems', *EJIL Talk* (Blog Post, 1 December 2017) <www.ejiltalk.org/a-compliance-based-approach-to-autonomous-weapon-systems>.

96 Elies van Sliedregt, *Individual Criminal Responsibility in International Law* (Oxford University Press, 2012) 41. But not all; France, for instance, does not recognise *dolus indirectus*.

that a particular result will occur in the ordinary course of events, but without any positive desire to bring it about'.[97] The previous example with the commander who releases an AWS knowing that it mislabels pink buses as tanks, and knowing that pink buses will be present in the operational area, will knowingly target those buses, even though they know nothing about the AI's internals. The fact that the commander may have no ill will toward those civilians at all, and simply is indifferent to their fate, does not reduce their criminal liability under this mode.[98] Like with the *dolus directus* cases, this situation is similarly labelled as an 'easy case' in most commentaries.[99] There is no gap, despite the AWS still acting autonomously and being imperceivable. In addition, many jurisdictions do not require 100% certainty to consider someone to have acted knowingly: often, 'virtual' certainty suffices.[100] Thus, it is likely that a commander deploying an AWS with a 95% failure rate would also be convicted for knowingly causing that failure to happen. However, it is less clear where exactly the lower boundary of the *dolus indirectus* probability range starts.[101] For instance, would releasing an AWS with a 90% failure rate still be considered as acting with *dolus indirectus*? What about 80%? In practice, the result will probably vary by situation, as this lower boundary is 'necessarily fluid and contingent'.[102]

B *True Gap Situations? Unreasonable Risk-Taking*

In the absence of direct intent or virtual certainty, we descend to the realm of risk-taking. In this situation, the actor is aware of a significant (but not certain) risk of the actus reus manifesting, but decides to take this risk anyway.[103] Risk-taking is criminalised under recklessness in common law doctrines and either *dolus eventualis* or conscious negligence in continental systems.[104] It is probably not controversial to state that the commander who releases an AWS with an 80% failure rate would be held liable under recklessness or *dolus eventualis*. They know that there is a high probability that the AWS will mischaracterise civilians as targets, but accept this possibility anyway. Similarly, the commander who knows that the AWS distinguishes targets based on outline size, and deploys the AWS in

---

97 David L Nersessian, 'Whoops, I Committed Genocide! The Anomaly of Constructive Liability for Serious International Crimes' (2006) 30(2) *Fletcher Forum of World Affairs* 81, 83.

98 See, eg, Geiß and Lahmann (n 15) 392.

99 Amoroso (n 89) 133.

100 This is often formulated as 'practically certain', 'virtually certain', or 'will almost inevitably cause': see, eg, *Model Penal Code* art 2 § 2.02(2)(b)(ii) (American Law Institute, 1985); *Prosecutor v Bemba Gombo (Decision Pursuant to Article 61(7)(a) and (b) of the* Rome Statute *on the Charges of the Prosecutor)* (International Criminal Court, Pre-Trial Chamber II, Case No ICC-01/05-01/08, 15 June 2009) [358]–[362].

101 See Alexander F Sarch, 'Willful Ignorance, Culpability, and the Criminal Law' (2014) 88(4) *St. John's Law Review* 1023, 1033–4.

102 See Abhimanyu George Jain, 'Autonomous Cyber Capabilities and Individual Criminal Responsibility for War Crimes' in Rain Liivoja and Ann Väljataga (eds), *Autonomous Cyber Capabilities under International Law* (NATO CCDCOE Publications, 2021) 291, 311.

103 Sliedregt (n 96) 42.

104 Weigend (n 49) 499. In the remainder of this discussion, I will ignore conscious negligence and discuss only *dolus eventualis* for simplicity, as the boundary between them in continental systems is quite fuzzy. Additionally, courts are often prepared to convict a risk-taker on the basis of *dolus eventualis* even if its distinguishing element vis-a-vis conscious negligence (volition to accept the risk) is proven only circumstantially or by inference: at 501–2.

a city with vehicles of comparable size to the target tanks, will also accept the risk that such vehicles are mistakenly targeted, even though they may not be certain that the AWS will encounter such a vehicle nor exactly how the AI was trained to recognise such outlines.[105]

It is submitted that the majority of problematic AWS-related incidents will involve this type of risk-taking. While AWS can deliberately be misused purposely or knowingly, and malicious actors certainly exist, they will likely constitute the exception rather than the rule.[106] The views of states toward AWS technology appear to be far from cynical: indeed, what we mostly see are reaffirmations of states' commitments to develop and use AWS technology, but as IHL-compliantly as possible.[107] As Jens Ohlin argues, the 'most likely scenario is that the military commander was reckless in his deployment of the AWS, in the sense that he was aware of the risk that the AWS would violate a core prohibition in IHL, but the military commander decided to deploy the system anyway'.[108] This is indirectly also reflected in the solutions presented in literature to address the RGP. For instance, one of the most popular proposals, that of reframing command responsibility to include AWS as 'subordinates', is often provided on the explicit rationale of its reduced mens rea requirement of recklessness.[109] Marta Bo discusses interpreting recklessness into the International Criminal Court ('ICC') mens rea standard, motivated by the observation that most AWS operators 'may not have intended to attack civilians, but [have] "only" taken the risk of such occurrence'.[110] A similar rationale also motivates other proposals, such as relying on joint criminal enterprise doctrine ('JCE'),[111] which under JCE III similarly allows reckless mental states.[112] This indicates that these authors also presume (even if this is unstated) that risk-taking will be the dominant mental state that should be addressed.

Thus clearly, many authors view risk-taking as one of the prime reasons why a responsibility gap will materialise. If true that risk-taking will indeed be the dominant degree of mens rea in practice, this is indeed prima facie disconcerting.

---

105 A similar example was given by Taylor of an AWS that relies on heat signatures to make its distinctions: Taylor (n 5) 321.

106 Amoroso (n 89) 138.

107 See, eg, Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, *Draft Report of the 2021 Session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems*, UN Doc CCW/GGE.1/2021/CRP.1 (8 December 2021) [12]; Ministère des Armées, *L'intelligence Artificielle Au Service de La Défense* (Report, September 2019) 3; Defense Science Board, Department of Defense, *The Role Of Autonomy in DoD Systems* (Report, 2012).

108 Jens David Ohlin, 'The Combatant's Stance: Autonomous Weapons on the Battlefield' (2016) 92 *International Law Studies* 1, 26.

109 Neha Jain, 'Autonomous Weapons Systems: New Frameworks for Individual Responsibility' in Nehal Bhuta et al (eds), *Autonomous Weapons Systems: Law, Ethics, Policy* (Cambridge University Press, 2016) 303, 315; Russell Buchan and Nicholas Tsagourias, 'Autonomous Cyber Weapons and Command Responsibility' in Rain Liivoja and Ann Väljataga (eds), *Autonomous Cyber Capabilities under International Law* (NATO CCDCOE Publications, 2021) 321, 324–5.

110 Bo (n 30) 278–95.

111 Amoroso (n 89) 137.

112 This only requires that the actus reus was 'a natural and foreseeable consequence': *Prosecutor v Tadić (Judgement)* (International Criminal Tribunal for the Former Yugoslavia, Appeals Chamber, Case No IT-94-1-A, 15 July 1999) [204].

However, it is argued that this gap — insofar as it exists — is not specifically caused by the characteristics of AI but rather by the jurisdiction many analyses are based on: that of the ICC.[113] It is true that before the ICC, risk-taking is hard to prosecute. The *Rome Statute of the International Criminal Court*'s art 30[114] is usually interpreted as only allowing *dolus directus* and *indirectus*,[115] even though there has been some contrary jurisprudence in the past allowing recklessness.[116] With this jurisdiction as a reference point, authors are indeed right to worry about risk-taking staying unpunished.

However, that war crimes can only be tried under *dolus directus* or *indirectus* is not the default position, and the ICC is not the only forum for war crime prosecutions. The International Criminal Tribunal for the former Yugoslavia ('ICTY'), as an example, has been far more willing to interpret recklessness into its Statute's art 3.[117] With regard to the customary standard, while some disagreement persists, commentators generally agree that recklessness is an accepted intent standard for war crimes, even though this may differ per specific crime.[118] Finally, the grave breaches provision in art 85 of *Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (Protocol I)* ('*API*') related to acts such as deliberately targeting civilians and launching disproportionate attacks — likely the most common types of AWS-caused crimes — requires that the crime was committed 'wilfully', which has over the years been almost categorically interpreted as including risk-taking.[119] The Committee reviewing NATO's Yugoslavia campaign also believed the standard to be 'intention or recklessness'.[120] Thus, the gap seemingly 'caused' by risk-taking in this situation is a matter of *jurisdiction*, rather than a function of specific characteristics of AWS. In courts — domestic or international — which allow recklessness for war crimes,

---

113 See, eg, Jain (n 102); Buchan and Tsagourias (n 109). Cf Amoroso and Giordano (n 86) 221–2.

114 *Rome Statute of the International Criminal Court*, opened for signature 17 July 1998, 2187 UNTS 90 (entered into force 1 July 2002) art 30.

115 *Prosecutor v Katanga (Judgment Pursuant to Article 74 of the Statute)* (International Criminal Court, Trial Chamber II, Case No ICC-01/04-01/07, 7 March 2014) [775]; Sliedregt (n 96) 47.

116 See, eg, *Prosecutor v Lubanga Dyilo (Judgment Pursuant to Article 74 of the Statute)* (International Criminal Court, Trial Chamber I, Case No ICC-01/04-01/06, 14 March 2012) [1012]. Some scholars have also argued that recklessness can be read into the ICC's intent standard: see, eg, Knut Dörmann, *Elements of War Crimes under the Rome Statute of the International Criminal Court: Sources and Commentary* (Cambridge University Press, 2003) 43; Bo (n 30) 284–95.

117 *Galić Trial Judgement* (n 72) [596].

118 Dörmann (n 116) 43; Gerhard Werle and Florian Jessberger, '"Unless Otherwise Provided": Article 30 of the ICC Statute and the Mental Element of Crimes under International Criminal Law' (2005) 3(1) *Journal of International Criminal Justice* 35, 53–4; International Committee of the Red Cross, *Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons* (Report, 2016) 45.

119 Claude Pilloud et al, *Commentary on the Additional Protocols of 8 June 1977 to the Geneva Conventions of 12 August 1949* (Martinus Nijhoff Publishers, 1987) 994 [3474]; *Prosecutor v Perišić (Judgement)* (International Criminal Tribunal for the Former Yugoslavia, Trial Chamber I, Case No IT-04-81-T, 6 September 2011) [100]; Crootof (n 5) 1350.

120 Committee Established to Review the NATO Bombing Campaign against the Federal Republic of Yugoslavia, *Final Report to the Prosecutor by the Committee Established to Review the NATO Bombing Campaign Against the Federal Republic of Yugoslavia* (Report, 2001) [28].

commanders who are aware of an unreasonable risk that their AWS will cause an actus reus can be held criminally liable for their culpable risk-taking.[121]

Instead, the '?' in this subsection's title is meant to refer to a different possible obstacle to responsibility for risk-taking, and one which *does* flow from the specific characteristics of AWS: generic risk. As discussed in Part III(C), this refers to the situation where a person is aware only of a general notion of riskiness without exactly knowing what this risk may entail. Say a commander reads in the system manual that it is 'untested in snowy weather'. Even if this is not expressly prohibited by military policy, a reasonable commander would probably not deploy the weapon if it just snowed the prior night, just to be safe. However, our current commander is frustrated by the defenders' tenacity and decides to launch the weapon anyway, which after half an hour of operation, mistakenly targets a protected historical building and destroys it.

Can the commander be convicted under *API* art 85(4)(d) for wilfully attacking a historic monument? Even granting that 'wilfully' also encompasses recklessness as is commonly interpreted, this is still unclear. Of what risk was the commander aware when they made the decision to launch the AWS? On the one hand, they certainly were aware that a risk *existed*: that the AI was not validated in snowy weather implies that there is a chance it will not perform as well as during standard conditions. However, that the system is untested in condition *X* does not necessarily imply it *will* perform worse in condition *X*. Indeed, a quality ML system usually has some built-in robustness that allows it to maintain performance in novel situations.[122] In addition, even if one (hypothetically) assumes that the commander *knew* that this would drop the AWS's performance by 40%, does this general awareness that the AI might fail more frequently translate toward accepting the risk that a cultural monument would be destroyed? This might depend on the particular jurisdiction's conception of risk, and more importantly, how *specific* it must be. Does the accused need to be aware of the risk of a *specific* consequence ('that historical building might get damaged'), a *category* of harm ('a protected object might get damaged'), or simply a *risk in general* ('something might go wrong')?[123] In this particular case, proving anything but the last type of awareness will be quite hard, especially since the building was not even a type of civilian vehicle.

---

121 See Jens David Ohlin, 'Targeting and the Concept of Intent' (2013) 35(1) *Michigan Journal of International Law* 79, 81–90. Some disagreement exists with regard to the appropriate reading of 'willingly' in *API* (n 8) art 85(3)(a), which discusses deliberately targeting civilians: Ohlin (n 121) 93. Ohlin argues convincingly that for this subparagraph, interpreting 'willingly' as including recklessness would make subparagraph (b) (which criminalises disproportionate attacks) superfluous since any attack that carries the risk of civilian casualties would immediately qualify as falling under subparagraph (a), even though the casualties are acceptable from the perspective of proportionality: at 112–13. If one subscribes to this position, then any (mere) risky misuse of AWS should indeed not be characterised as a direct attack on civilians. Instead, the most viable provision for criminal liability would be subparagraph (b), as an overly high failure rate would likely (but not always) cause excessive collateral damage. Cf Part IV(C) of this paper, where the failure rate is low enough to justify an attack, since even if collateral damage occurs, it is unlikely to be excessive to the military advantage anticipated.

122 Arthur Holland Michel, UN Institute for Disarmament Research, *The Black Box, Unlocked: Predictability and Understandability in Military AI* (Report, 2020) 5. This is often referred to as the AI's ability to 'generalise'.

123 Jain (n 109) 317.

Thus, the problem of generic risk presents a genuine obstacle toward establishing mens rea that intrinsically flows from the characteristics of modern AI, ie, their lack of perceivability. How large this gap is, however, is also a function of the jurisdiction's particular interpretation of risk. If defendants can be convicted for engaging in generic 'risky business',[124] the gap will be narrowed; if not, many situations similar to the example given above will not be prosecutable. The gap can potentially also be narrowed if the jurisdiction in question allows negligence for war crimes: the argument, then, would be that the commander *should* have known not to deploy the AWS in snowy weather, which skirts around the generic risk problem.[125] Note however that importing negligence into criminal law is not an uncontroversial solution, and is often criticised on conceptual grounds.[126] While some domestic systems allow for the prosecution of negligently committed war crimes,[127] it is almost unheard of in international criminal law.[128]

C *Judicious Gap Situations: Justified Risk-Taking and Genuine Accidents*

Military operations inherently involve risk. IHL is aware that belligerents cannot be expected to fight perfect wars and is permissive of mistakes, errors in judgement, incomplete information and even 'deliberate' civilian casualties as long as these are justified by the military advantage accrued.[129] This brings us to two scenarios with harmful consequences where criminal liability is similarly barred due to the lack of mens rea, but where — this article argues — such a limitation is justified. I call these *judicious gaps*.

The first is *justified risk-taking*. Risk is unavoidable as long as we are working with imperfect tools. Just like any other weapon developed by humankind, AWS will not be perfect: it is unlikely that a 100% accuracy rate will ever be achieved.[130] IHL acknowledges this. 'Attacks against lawful targets cannot be risk-free to civilians located in or near them … [IHL] sets its sights lower, trying to minimize — rather than completely avert — such collateral damage.'[131] The aim of the principle of proportionality in IHL is exactly this: it permits attacks where commanders 'know' that civilian casualties will likely ensue due to unavoidable inaccuracies during attacks.[132] Thus, a commander who activates an AWS

---

124 See, eg, Kimberly Kessler Ferzan, 'Opaque Recklessness' (2001) 91(3) *Journal of Criminal Law and Criminology* 597, 597–600.

125 Sliedregt (n 96) 41.

126 See, eg, Crootof (n 5) 1381–6.

127 See ibid 1383.

128 Jack M Beard, 'Autonomous Weapons and Human Responsibilities' (2014) 45(3) *Georgetown Journal of International Law* 617, 644; McDougall (n 4) 67.

129 Laurie R Blank, 'Operational Law Experts Roundtable on the *Gotovina* Judgment: Military Operations, Battlefield Reality and the Judgment's Impact on Effective Implementation and Enforcement of International Humanitarian Law' (Research Paper 12-186, International Humanitarian Law Clinic, Emory University School of Law, 2012) 5; Stefan Oeter, 'Specifying the Proportionality Test and the Standard of Due Precaution: Problems of Prognostic Assessment in Determining the Meaning of "May Be Expected" and "Anticipated"' in Claus Kreß and Robert Lawless (eds), *Necessity and Proportionality in International Peace and Security Law* (Oxford University Press, 2020) 343, 352.

130 William H Boothby, *Weapons and the Law of Armed Conflict* (Oxford University Press, 2nd ed, 2016) 61–2.

131 Yoram Dinstein, *The Conduct of Hostilities under the Law of International Armed Conflict* (Cambridge University Press, 3rd ed, 2016) 155.

132 Blank (n 129) 5, 12.

knowing that there is a 1% chance of failure and that this risk is far outweighed by the military advantage gained by this deployment is taking a justified risk. It is no different from the commander who orders an artillery bombardment, 'knowing' that some deviations might hit surrounding civilian buildings and reconciling themselves with this fact.[133] Such persons would not be prosecutable under current risk-taking doctrines. Recklessness only condemns *un*justified risk-taking,[134] while continental courts rarely accept empirically unlikely events as *dolus eventualis*.[135]

The second is *genuine accidents*. As mentioned above, no tool is 100% reliable. Even with all the due diligence from designers, programmers, reviewers and commanders, accidents will inevitably occur, and they may sometimes have devastating consequences.[136] This is a fundamental principle of normal accident theory: 'An AI designed to do X will eventually fail to do X.'[137] Take for instance the commander who, after multiple efforts to ensure that the operational environment indeed corresponds with the AWS's testing and validation conditions and is reasonably convinced that no known failure triggers are present in the area, releases the system.[138] Unluckily, an unknown edge case emerges where the system is momentarily confused by light reflected off stained glass and causes massive damage to a crowded market. Under such situations, the pressure to find a person to blame is significant, but a responsibility gap will indeed manifest: under no circumstances can it be argued that the commander had the cognition and volition to attack that market. In fact, the defence can even argue that no reasonable person *could* have known.[139]

Thus, strictly speaking, we do have difficulties assigning liability in the two above situations, validating Premise 1. However, perhaps in such cases, the gap is justified. We cannot fight wars without risk, and holding persons liable for accidents no person could foreseeably have predicted or foreseen would be fundamentally unfair and contrary to the very purpose of criminal law. As noted by Thomas Weigend:

---

133 Some will disagree, finding that even a 1% failure rate is unacceptable if it risks harming civilians. This is often argued on the base of deontological reasoning such as the idea that it is never acceptable that machines can 'decide' (even in error) to take a human life: see, eg, Peter Asaro, 'On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-Making' (2012) 94(886) *International Review of the Red Cross* 687, 688–9. While there is much to say about this debate, we shall not explore this further as it strays too far from the accountability leg of the discussion.

134 Weigend (n 49) 494. This is usually a function of the probability and extent of the harm.

135 Jeroen Blomsma and David Roef, 'Forms and Aspects of Mens Rea' in Johannes Keiler and David Roef (eds), *Comparative Concepts of Criminal Law* (Intersentia, 2nd ed, 2015) 103, 183, 186. The ICTY also found that liability for international crimes requires 'an awareness of a *higher likelihood of risk* and a volitional element': *Prosecutor v Blaškić (Judgement)* (International Criminal Tribunal for the Former Yugoslavia, Appeals Chamber, Case No IT-95-14-A, 29 July 2004) [41] (emphasis added).

136 Charles Perrow, *Normal Accidents: Living With High-Risk Technologies* (Basic Books, 1984) 43.

137 Roman V Yampolskiy and MS Spellchecker, 'Artificial Intelligence Safety and Cybersecurity: A Timeline of AI Failures' (Paper, arXiv, 25 October 2016) 5 <http://arxiv.org/abs/1610.07997>, archived at <https://perma.cc/T4HB-E9VV>.

138 Dunlap (n 5) 71.

139 This would bar even negligence charges.

> Criminal punishment imposed on a person expresses blame for what the person did. In a rational system of law, it makes no sense to blame and punish a person for harmful occurrences that he had no possibility to prevent or that she was unable to foresee.[140]

Indeed, Robert McLaughlin takes an even clearer stance: in such cases, responsibility gaps are consciously enforced by IHL to prevent scapegoating.[141] Thus, while the market incident may be deeply emotionally outraging, anything *but* 'impunity' is unjustified from the perspective of criminal justice. Such situations are *judicious gaps*, which indirectly defeat Premise 2 of the RGP. We indeed establish that criminal liability cannot be assigned, but this is a justified result and certainly no reason to prohibit the use of the weapon involved in the accident. Note also that judicious gaps do not fulfil our retroactive disruption condition, as they may occur with any type of weapon and not specifically AWS.

D *Manufactured Gaps: Preferring Not to Know*

Let us now expand our above scenario. Say that our commander has just heard of the market attack and the civilian casualties that resulted. At this point in time, no one truly knows that the failure was caused by the stained-glass edge case: this would be information that can only be determined after an investigation on the ground and a technical inspection of the AWS. The commander only knows that an accident did occur. Above, this article concluded that the commander cannot be held criminally liable for failing to predict this unforeseeable result. What is important, however, is what happens next. A responsible commander would, in this situation, retire the AWS and request an investigation to determine the cause of the previous failure.[142] This would allow them to uncover the failure trigger and consciously avoid sending the AWS to areas with stained-glass panes in future operations. This falls under a commander's duty to suppress in IHL.[143] However, say that our commander is less considerate. Not content with losing a valuable asset while the enemy is gaining territory in the city, they do not prioritise the post-action investigation and once again activate the AWS while the technical analysis is slowly getting underway. As the city is famous for its historic churches, the AWS encounters another reflection and repeats the same error, destroying a protected site. How to classify this situation? Is this just another genuine accident for which we should enforce a responsibility gap? Or something more disconcerting?

Intuitively, we would probably not be satisfied with granting the commander another 'free pass'. After the first incident, the repeat event is both foreseeable and avoidable, and it seems fair to blame the commander for creating the conditions (ie, deploying the AWS) which allowed the second attack to take place. However, note that the commander was not aware of this failure trigger at the moment they made the second deployment decision, and thus not aware of the risk. *We* are, but only as omniscient discussants. At most, one could argue that the previous incident

---

140 Weigend (n 49) 490.

141 Robert McLaughlin, 'Unmanned Naval Vehicles and the Law of Naval Warfare' in Hitoshi Nasu and Robert McLaughlin (eds), *New Technologies and the Law of Armed Conflict* (TMC Asser Press, 2014) 229, 235–6.

142 See, eg, Department of the Army Headquarters (n 67) [5-73].

143 *Geneva Convention IV* (n 8) art 146.

would have given the commander awareness of a generic risk ('perhaps this system is not as robust as foreseen'), which, as discussed in Part IV(B), is already quite difficult to prosecute. At worst, the defence could argue that the commander perceived the market incident as falling within the 0.5% error rate mentioned in the manual and assumed that it would not repeat again. Thus, a gap would result, but this time it is hard to call it a judicious one.

One may argue at this point that the commander clearly made blameworthy decisions. As mentioned above, the choice not to follow up on a civilian casualty event is already a violation of their duty to suppress, and an argument could be made that they more generally violated the constant care obligation of *API* art 57(1).[144] They might even have done this for unscrupulous motivations, such as not wanting to lose an important asset in the case that the AWS is indeed found to be problematic. However, linking this omission of an IHL obligation to criminal liability for the historic site attack is difficult. While perhaps meriting disciplinary measures, failing to investigate an incident does not amount to a war crime in itself. What the omission accomplished was to prevent the commander from gaining knowledge of a (specific) risk — knowledge which is necessary in order to convict them for wilfully attacking a protected monument.

This is a *manufactured gap* — where all components are present to close the gap quite nicely, but an omission causes the cognition component to remain unfulfilled when it could have been fulfilled in theory if the required action was taken. Failure to investigate is the best example of this gap, but it can also occur in more straightforward ways, such as commanders deliberately learning as little as possible of their AWS's algorithm. Like failure to investigate, technical ignorance can arguably be viewed as a violation of IHL (specifically: the duty of precautions),[145] but is not a war crime in itself. Indeed, the fact that these ploys are attempted quite frequently in civil situations — such as the corporate manager who prefers not to know why their affiliate needs millions in cash for 'commissions', or the AI developer who deliberately refuses to study how their bot chooses items to purchase on the darknet[146] — is an indication that it will likely also occur with military systems. If the commander is truly malicious, they might do so with the express aim of arguing later that they lacked awareness of the risk. Such situations are often referred to as wilful blindness in common law systems.[147] While these systems have developed a specific doctrine to address this legal ploy,[148] it comes with its own controversies and is not universally applicable.[149] Thus, the extent that this manufactured gap persists depends on the jurisdiction's ability to manage

---

144 See Chris Jenks and Rain Liivoja, 'Machine Autonomy and the Constant Care Obligation', *Humanitarian Law and Policy* (Blog Post, 11 December 2018) <blogs.icrc.org/law-and-policy/2018/12/11/machine-autonomy-constant-care-obligation>.

145 Longobardo (n 70) 80–1.

146 See, eg, David Luban, 'Contrived Ignorance' (1999) 87(4) *Georgetown Law Journal* 957, 962; Jana Kasperkevic, 'Swiss Police Release Robot That Bought Ecstasy Online', *The Guardian* (online, 23 April 2015) <www.theguardian.com/world/2015/apr/22/swiss-police-release-robot-random-darknet-shopper-ecstasy-deep-web>.

147 Luban (n 146) 959.

148 If wilful blindness is proven, some jurisdictions equate that with knowledge: see, eg, *R v Briscoe* [2010] 1 SCR 411, [21]. Evidently, in such a case, there is no gap.

149 See, eg, Christopher Sherrin, 'Wilful Blindness: A Confused and Unnecessary Basis for Criminal Liability?' (2014) 47(2) *University of British Columbia Law Review* 709, 710; Luban (n 146) 959–61.

wilful blindness cases,[150] and in any event is an unwelcome prospect indeed that creates perverse incentives for belligerents.

V RECOMMENDATIONS AND CONCLUDING REMARKS

In this article, I examined an important leg of the responsibility gap problem, namely the premise that characteristics of AWS cause criminal liability for harmful consequences during wartime to become impossible. However, this premise is only useful if sufficiently demarcated by specifying what particular circumstances are problematic. Enhanced specificity makes the RGP more convincing for prohibitory agendas and more concrete for regulatory agendas by properly identifying what problems future measures must address. In this light, the article focused on a very important requirement for criminal liability, mens rea. The cognition and volition elements of mens rea are meant to limit criminal liability only to truly blameworthy actors,[151] but may sometimes create obstacles for the prosecution of acts we would prefer not to go unpunished.

In the realm of AI in AWS, this article has found that these obstacles flow primarily from *reduced perceivability*, which impacts the accused's cognition of the consequence or risk thereof. In some instances, this reduced perceivability is a direct product of the technology: modern, opaque machine learning. In others it is avoidable but such avoidance necessitates active steps, such as technical expertise and iterative awareness. In addition, the problem of generic risk was identified, where humans can achieve only a general understanding of the risk entailed by the use of an AWS, but which cannot be linked to any particular result.

From these foundations, I argued that three main conditions exist, with their *problematique* relatively scaling in bell curve fashion with degrees of intent in criminal law. Departing from the highest modes of intent, *dolus directus* and *indirectus*, it was found that Premise 1 fails: the lowered perceivability of AI does not prevent criminal liability for purposeful and knowing conduct. Moving toward specific risk-taking, it was found that concerns about gaps are jurisdiction-specific: it is only in forums such as the ICC, where *dolus eventualis* is not allowed, where the problem lies. The primary challenge lies with *generic risk-taking*, where the actor is aware of a general idea that their decision to use an AWS is risky, but where this risk is so diluted across an almost infinite set of possibilities that it is difficult to link with the actus reus for which they are to be charged. Turning toward lower probabilities of risk, these indeed induce a gap in the sense that cognition cannot be proven — however, this is by design. Thus, the fact that 'impunity' persists for justified risk-taking and genuine accidents (even if the actual harm, ex post, shocks the conscience) is judicious. More problematic is failures to know more. If the first accident is unforeseeable, the second need not be. However, criminal law cannot impute liability for consequences which are

---

150 Some authors have proposed a repurposed command responsibility doctrine to address this issue as well, as command responsibility attributes criminal liability *for the subordinate's violation* instead of the prior omission: Buchan and Tsagourias (n 109) 328. In that sense, this repurposed command responsibility would be quite effective as it would address both manufactured gaps and the unavailability of *dolus eventualis*. On the other hand, many have also vehemently rejected this proposal as command responsibility was strictly developed to address human-to-human relationships: Chengeta (n 7) 27. This debate, therefore, is still undecided.

151 Weigend (n 49) 491.

only theoretically knowable, and the omission which led to such ignorance cannot always be transformed into liability for the harmful result.

For regulationists, having identified the loci of the problem, we can also better tailor our recommendations for future legislation and policy. The main issues to be addressed are generic risk-taking and manufactured ignorance. With regard to the former, it may imply that global performance metrics as an indicator of the AWS's reliability may not be sufficient, as they leave too broad a margin between the commander's state of cognition and the ensuing harm. Reducing the systems' opacity via XAI and requiring better commander training in the technical aspects can help transform generic awareness of risk into more specific predictions of consequences. Technical training also helps with manufactured ignorance, as commanders who have some background in AI from mandatory training sessions will find it more difficult to claim that they did not foresee the resulting harm when they made the deployment decision. Technical training is not an explicit obligation under IHL but can be derived from it.[152] Total expertise is not necessary, but some background (or at least, a technical adviser to consult with) is crucial.[153] Finally, post-action reporting and investigation should be mandatory and strictly enforced for all AWS deployments to prevent manufactured ignorance in iterative cases. Any results — particularly of previously unknown failure triggers — should also be communicated immediately to all users of sister systems, rendering these persons incapable of claiming that they were unaware of the risk.

---

152 Marco Sassòli, 'Autonomous Weapons and International Humanitarian Law: Advantages, Open Technical Questions and Legal Issues to Be Clarified' (2014) 90 *International Law Studies* 308, 339.

153 The same conclusion was reached with respect to cyberweapons, which also require some technical expertise: Michael N Schmitt (ed), *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (Cambridge University Press, 2017) 400.