Exploring the Impact of Automation Bias and Complacency on Individual Criminal Responsibility for War Crimes

Antonio Coco*

Abstract

With advancing technology, complex decision-making in warfare, including targeting, is increasingly assisted by machines. Although involving humans in decision-making is often seen as a safeguard against machine errors, it does not always prevent them. Machines can make incorrect determinations or delay them when time is critical. In these cases, human operators, influenced by automation bias (excessive trust in machines' determinations, despite the availability of contradicting or different information from other sources) or complacency (excessive trust in machines' determinations, leading to reduced vigilance), may fail to recognize machine errors, potentially resulting in conduct amounting to a war crime. Considering the role of automation bias and complacency in the determination of the criminal responsibility of systems' operators is crucial, especially, for understanding the accountability framework for war crimes involving autonomous weapon systems (AWS). By exploring how automation bias and complacency affect the determination of criminal responsibility for humans who operate AWS, this article offers insights for lawmakers at the national and international levels to understand complexities and effectively shape legislative responses with respect to criminal responsibility. This article also examines automation bias and complacency in psychology, their relevance in military operations employing AWS, and their potential to exonerate human operators from criminal responsibility. It concludes by advocating for

* Senior Lecturer, Essex Law School, University of Essex (United Kingdom). I am grateful to Marta Bo, Paola Gaeta, Florian Jessberger, Henning Lahmann and Meagan Wong for their helpful feedback on earlier versions of this article; and to the Public International Law Cluster at Essex Law School, University of Essex, the Geneva Academy, and the Geneva Graduate Institute for the opportunity to present my work in progress. [antonio.coco@essex.ac.uk]

Journal of International Criminal Justice (2023), 1 of 22 https://doi.org/10.1093/jicj/mqad034

legislative, organizational, and technical measures to counteract automation bias and complacency.

1. Introduction

In 1983, the early warning systems of the Soviet Air Defence Forces detected what appeared to be hostile United States (US) missiles. Stanislav Petrov, lieutenant colonel in charge of the response, did not believe the systems' analysis to be accurate and, correctly identifying the occurrence as a false alarm, decided not to respond with armed force and to simply stay put.[1] Petrov set a well-known example in the history of human–machine interactions. His cautiousness to act upon the machine's input avoided the catastrophic consequences that could have materialized if he had otherwise used force based on an erroneous assessment of facts. Yet, history has known opposite examples: humans placing trust in the accuracy of a machine's determination or input when circumstances would have suggested caution instead. For instance, on 22 March 2003, the US Army's missile system known as 'Patriot' observed an object in the skies and wrongly misidentified it as an Iraqi anti-radiation missile, which would have been a valid target for engagement. The Patriot operators, with about a minute to decide,[2] acted upon the machine's determination to shoot down the object — which turned out to be a British military aircraft, whose two crew members lost their lives.[3] The operators of the Patriot system could have overridden the machine's determination but, due among other things to a lack of proper training, chose to trust it instead.[4] The incident is often recalled as an example of automation bias,[5] i.e. a tendency of humans to rely on, and trust too much, information provided by a machine, even when contradictory or different information provided by other sources is available or could be available if properly searched for.[6] The concept is cognate to that of

1 P. Scharre, *Autonomous Weapons and Operational Risk* (Centre for New American Security, 2016), at 34; R. Crootof, M. Kaminski, and W.N. Price II, 'Humans in the Loop', 76 *Vanderbilt Law Review* (2023) 429, at 500.

2 United Kingdom, Ministry of Defence (UK MoD), *Military Aircraft Accident Summary, Aircraft Accident to Royal Air Force Tornado GR MK4A ZG710*, May 2004, available online at https://www.gov.uk/government/publications/military-aircraft-accident-summary-aircraft-accident-to-raf-tornado-gr-mk4a-zg710 (visited 27 August 2023), § 11.

3 UK MoD, *supra* note 2; Scharre, *supra* note 1, at 30–31; J.K. Hawley, *Patriot Wars: Automation and the Patriot Air and Missile Defense System* (Centre for New American Security, 2017), at 8.

4 UK MoD, *supra* note 2, § 11; M. Cummings, 'Automation Bias in Intelligent Time Critical Decision Support Systems', in *AIAA 1st Intelligent Systems Technical Conference* (2004), at 5.

5 Scharre, *supra* note 1, at 31; Cummings, *supra* note 4, at 5.

6 A. Adensamer, R. Gsenger, and L. Klausner, '"Computer Says No": Algorithmic Decision Support and Organisational Responsibility', 7–8 *Journal of Responsible Technology* (2021) 100014, at 4; Cummings, *supra* note 4, at 2; C. Goddard, A. Roudsari, and J. Wyatt, 'Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators', 19 *Journal of the American Medical Informatics Association* (2012) 121; R. Parasuraman and D. Manzey, 'Complacency and Bias in Human Use of Automation: An Attentional Integration', 52 *Human Factors* (2010) 381; R. Williams, 'Rethinking Administrative Law for Algorithmic

'complacency', which can be observed when humans interacting with machines unwarrantedly monitored them with less vigilance, assuming that the relevant system was operating correctly and failed to check that this was indeed the case.[7]

Various forms of human–machine interaction have been common in warfare for decades. With developments in technology, especially in artificial intelligence (AI) and machine learning, machines are being entrusted with more and more complex determinations, including as they relate to targeting. Debates about the employment of AI in warfare often revolve around the risks of autonomous weapon systems (AWS), which have been defined as 'weapons that select and apply force to targets without human intervention' and that '[a]fter initial activation or launch by a person, ... [self-initiate or trigger] a strike in response to information from the environment received through sensors and on the basis of a generalized ''target profile''.'[8] Concerns arise from the possibility that AWS engage in erroneous target selection.[9] It is often assumed that ethical and legal problems may partially be solved by securing the involvement of humans in the machine's determinations, in what is often called 'meaningful human control'.[10] In fact, 'Almost all AWS are supervised in real time by a human operator that can intervene to authorize, override, veto or deactivate the weapon as needed'.[11] This kind of human–machine interaction can take many configurations, including two well-known ones: a 'human in the loop' setting, in which the machine has identified a target and the human must validate it before the machine initiates engagement — a scenario in which some would deem the weapon system not to be properly 'autonomous'; and a 'human on the loop' setting, in which the machine identifies the target and commences engagement unless the human overrides

Decision Making', 42 *Oxford Journal of Legal Studies* (2022) 468, at 23; J. Zerilli et al., 'Algorithmic Decision-Making and the Control Problem', 29 *Minds and Machines* (2019) 555, at 561. See also M. Bo, 'Autonomous Weapons and the Responsibility Gap in Light of the *Mens Rea* of the War Crime of Attacking Civilians in the ICC Statute', 19 *Journal of International Criminal Justice* (2021) 275, at 296.

7 Goddard, Roudsari and Wyatt, *supra* note 6; Parasuraman and Manzey, *supra* note 6, at 381–382; Williams, *supra* note 6, at 23.

8 'International Committee of the Red Cross (ICRC) Position on Autonomous Weapon Systems: ICRC Position and Background Paper', 102 *International Review of the Red Cross* (2020) 1335, at 1339. Different definitions have been used by different entities, including states and other organizations: see M. Taddeo and A. Blanchard, 'A Comparative Analysis of the Definitions of Autonomous Weapons Systems', 28 *Science and Engineering Ethics* (2022) 37; A. Seixas-Nunes, *The Legality and Accountability of Autonomous Weapon Systems: A Humanitarian Law Perspective* (Cambridge University Press, 2022), at 67–76.

9 See e.g. Council of Europe, Committee on Legal Affairs and Human Rights, *Emergence of lethal autonomous weapons systems (LAWS) and their necessary apprehension through European human rights law*, Report, Doc. 15683, 9 January 2023, at 4, § 10.

10 In this respect, see M. Bo, 'Criminal Responsibility by Omission for Failures to Stop Autonomous Weapon Systems' and G. Acquaviva, 'Crimes *without* Humanity? Artificial Intelligence, Meaningful Human Control, and International Criminal Law', both in this special issue of the *Journal of International Criminal Justice*.

11 ICRC, *supra* note 8, at 1341.

the machine's determination and halts the engagement.[12] The 'human on the loop' model has been praised not only for its purported ethical benefits, but also for its operational dimension: the machine would observe the environment, interpret it, select a target and engage it autonomously, but the operator would be notified when a target is identified and have the power to override the system, should the circumstances demand it.[13] This is, for example, how the Phalanx defensive system works.[14]

And yet, the involvement of a human does not always constitute an effective safeguard against erroneous targeting, or even reduce the chances of such an erroneous targeting happening.[15] Granted, if the machine always makes the correct determination, automation bias has little to no practical importance.[16] Problems arise, on the contrary, when the machine makes an undesired, unexpected or outright erroneous determination, or when it delays its determination in a situation in which time constraints would impose quicker action.[17] Some of these erroneous determinations may not be ever recognized by any human, if automation bias is widespread enough within the monitoring/supervising human crew, then no one is able to recognize it.[18] In some hypothetical scenario, it could also happen that a machine erroneously targets a protected individual or object, the human operator recognizes the error, but nevertheless, they take advantage of such error and deliberately decide not to correct it. When a serious violation of international humanitarian law ensues from such human decision, it is not difficult to imagine that the operator could be held responsible for a war crime, due to their undoubtedly intentional conduct.[19]

In some instances, instead, it may happen that the machine's operator, affected by automation bias, fails to recognize that a machine's determination is erroneous, when instead they should have been able to recognize such

12 These and other scenarios of human-machine interaction with respect to the employment of lethal AWS in war are presented by G. Acquaviva, 'Autonomous Weapons Systems Controlled by Artificial Intelligence: A Conceptual Roadmap for International Criminal Responsibility', 60 *The Military Law and the Law of War Review* (2022) 89, at 96. See also R. Crootof, 'A Meaningful Floor for Meaningful Human Control Autonomous Legal Reasoning: Legal and Ethical Issues in the Technologies in Conflict', 30 *Temple International & Comparative Law Journal* (2016) 53, at 54; C. McDougall, 'Autonomous Weapon Systems and Accountability: Putting the Cart before the Horse', 20 *Melbourne Journal of International Law* (2019) 1, at 13.

13 H. Scheltema, *Lethal Automated Robotic Systems and Automation Bias, EJIL: Talk!*, 11 June 2015, available online at https://www.ejiltalk.org/lethal-automated-robotic-systems-and-automation-bias/ (visited 6 September 2023).

14 A. Etzioni and O. Etzioni, 'Pros and Cons of Autonomous Weapons Systems', *Military Review* (2017) 72, at 79; United States Navy, *MK 15 - Phalanx Close-In Weapon System (CIWS)*, 20 September 2021, available online at https://www.navy.mil/Resources/Fact-Files/Display-FactFiles/Article/2167831/mk-15-phalanx-close-in-weapon-system-ciws/ (visited 6 September 2023). More on this system *infra*.

15 Scheltema, *supra* note 13.

16 Zerilli et al., *supra* note 6, at 558.

17 Parasuraman and Manzey, *supra* note 6, at 382.

18 S. Chesterman, *We, the Robots? Regulating Artificial Intelligence and the Limits of the Law* (Cambridge University Press, 2021), at 153 — referring to general applications of AI.

19 Acquaviva, *supra* note 12, at 105.

erroneity. It may happen that the operator's failure, adding up to the machine's error, results in conduct which *prima facie* could amount to a war crime. Let us imagine, for instance, that the abovementioned Patriot incident had resulted in the shooting down of a civilian aircraft: would any criminal responsibility have arisen for the system operators, and would their alleged automation bias have played any role in such determination? The answer holds great importance in understanding the benefits and drawbacks of different possible legal avenues for holding individuals accountable for war crimes committed by employing AWS. This article, thus, attempts to shed light upon how automation bias and complacency and — more in general — overtrust exacerbate the general difficulties of ascribing criminal responsibility in AI-based processes in warfare. If the effect of automation bias or complacency, indeed, is that an individual lacks the *mens rea* required for the war crime in question, this could place a further obstacle on the already bumpy road that may lead to accountability. By exploring how automation bias may influence the determination of criminal responsibility for those involved in operating AWS, this article hopes to offer valuable insights to lawmakers at the national level and at the international level, enabling them to better comprehend the complexities at hand and shape legislative responses accordingly, especially when defining relevant crimes.

To pursue these aims, Section 2 starts by examining the meaning of automation bias and the cognate concept of automation complacency as they emerged from studies in psychology. The relevance of these concepts — taken together because of their logical connection, as they both represent manifestations of overtrust in machines — will be contextualized with respect to human–machine interactions in military operations, and to the rules of international humanitarian law. Section 3 recalls the basic tenets of the definition of war crimes related to targeting in the conduct of the hostilities, placing special attention on the mental element, which can be required for such crimes. Section 4, then, considers whether and how automation bias and complacency could have the effect of exonerating a human from criminal responsibility, because they may determine the lack of the mental element required for the relevant crime. Finally, this article concludes that automation bias and complacency could be countered by legislative, organizational and technical measures.

2. Overtrust in Machines in Warfare

Resort to automated and autonomous machines[20] undoubtedly helps humans in many respects: machines can often collect more data, process large amounts

20 As said above, different states interpret these expressions differently: see e.g., Seixas-Nunes, *supra* note 8, at 67–76. In the military context, a possible definition of 'automated weapon' or 'automated weapons system' is that of 'one that is able to function in a self-contained and independent manner although its employment may initially be deployed or directed by a human operator'. ICRC, *International Humanitarian Law and the challenges of contemporary armed*

of it efficiently and provide valuable information in times or manners which would not be feasible to humans. Yet, scholars have warned for a long time that — despite the undeniable benefit of automation in many work settings — the consequences of resorting to it must be always carefully evaluated.[21]

Thus, it is not surprising that the effects of automation bias on human behaviour have been long studied in a sector like aviation, where human–machine interaction is common and the consequences of errors by either the machine or the human can be tragic. Such interactions were studied from a psychological perspective, finding that — when performing the same task — humans aided by a machine perform better than humans operating without machine aid, if the machine works as expected.[22] On the contrary, when the machine makes an error, humans who are aided by the machine performed worse than humans who operate without any machine's aid — even if they had the option to ignore or override the machine's suggestion or determination.[23] Such experiments, in particular, evidenced pilots' tendencies to 'overtrust' automated systems, like those employed to detect engine fires. Such overtrust was displayed even when the pilots had been made aware that the systems were subject to fallacies.[24] Alongside automation bias, experiments displayed pilots also experiencing the cognate situation of automation complacency.[25]

In fact, automation bias and complacency are different examples of the effects that trust in technology can induce on the attitude, alertness and attention levels of humans interacting with an automated or autonomous machine[26] — what has been termed as the 'control problem' in ethical and psychological studies of human–machine interactions.[27] Automation bias and complacency can be both conscious and unconscious,[28] and — in the relevant studies — have been observed in both participants who have the expertise for the task at hand (e.g., airplane pilots) and participants who do not have such expertise,[29] in any domain that requires a human to

conflicts, Official working document of the 31st International Conference of the Red Cross and Red Crescent (28 November – 1 December 2011), available online at https://www.icrc.org/en/doc/resources/documents/report/31-international-conference-ihl-challenges-report-2011-10-31.htm (visited 6 September 2023), at 39. For a possible definition of 'autonomous weapons system', see *supra*, Section 1.

21 L. Skitka, K. Mosier, and M. Burdick, 'Does Automation Bias Decision-Making?', 51 *International Journal of Human-Computer Studies* (1999) 991, at 1002.

22 *Ibid.*

23 *Ibid.*

24 K. Mosier et al., 'Automation Bias: Decision Making and Performance in High-Tech Cockpits', 8 *The International Journal of Aviation Psychology* (1998) 47; Scheltema, *supra* note 13.

25 See *supra*, Introduction, and Goddard, Roudsari, and Wyatt, *supra* note 6; Parasuraman and Manzey, *supra* note 6, at 381–382; Williams, *supra* note 6, at 23.

26 Parasuraman and Manzey, *supra* note 6, at 381.

27 Zerilli et al., *supra* note 6, at 556.

28 Parasuraman and Manzey, *supra* note 6, at 406.

29 A. Deeks, 'The Judicial Demand for Explainable Artificial Intelligence', 119 *Columbia Law Review* (2019) 1829, at 1846, fn 103; Parasuraman and Manzey, *supra* note 6, at 397.

monitor the operation of an automated or autonomous system or subsystem.[30] These psychological conditions can produce undesirable results both by commission and by omission. They can induce a human error 'by commission' when the human follows the machine's erroneous determination and acts erroneously too; and they can induce a human error 'by omission' when the human, in the absence of a machine's prompt or alert, fails to recognize a situation in which instead they should have acted.[31]

Automation bias and complacency may arise out of different reasons. One has to do with the sheer amount of human involvement in a somewhat automated process. It has been found that, no matter whether 'in' or 'on' the loop, the less the human is entrusted to do, the more their tendency to trust the machine and live by its determinations — a manifestation of greater belief in the machine's capabilities than in one's own.[32] This ties into the second factor affecting human attitudes when interacting with a machine: the more reliable is an automated or autonomous system, the more likely is that the human who must interact with it will display automation bias and/or complacency.[33] When the machine has been functioning accurately and as desired for a long time, human operators may fall into a 'false sense of security' which in turn could result in lessened vigilance and lower 'operational prudence'.[34] And this without even touching upon the issue of 'skill fade' or 'skill degradation' which may occur when humans become less and less used to perform certain tasks because of their reliance on machines.[35] On the contrary, less reliable automated and autonomous systems demand more vigilance and human agency and, thus, are less likely to result in a display of those psychological conditions.[36] A third reason for automation bias and complacency is the human's workload: they more commonly occur when the human is expected to carry out manual tasks and monitor automated or autonomous machines at the same time, or when the human is in charge of monitoring multiple automated or autonomous machines at the same time.[37] Technological progress in the field of AI has amplified the problem even further: deep learning algorithms based on artificial neural networks —

30 Cummings, *supra* note 4, at 3; Parasuraman and Manzey, *supra* note 6, at 382; D. Lyell and E. Coiera, 'Automation Bias and Verification Complexity: A Systematic Review', 24 *Journal of the American Medical Informatics Association* (2017) 423.

31 Cummings, *supra* note 4, at 2; Parasuraman and Manzey, *supra* note 6, at 397; Skitka, Mosier, and Burdick, *supra* note 21, at 994.

32 Chesterman, *supra* note 18, at 68; Parasuraman and Manzey, *supra* note 6.

33 Zerilli et al., *supra* note 6, at 561.

34 Both quotes from Hawley, *supra* note 3, at 10.

35 Crootof, Kaminski, and Price II, *supra* note 1, at 469.

36 Parasuraman and Manzey, *supra* note 6, at 384–385 (automation complacency) and 395–396 (automation bias); Zerilli et al., *supra* note 6, at 561.

37 V. Boulanin et al., *Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control* SIPRI and ICRC (2020), available online at https://www.sipri.org/publications/2020/other-publications/limits-autonomy-weapon-systems-identifying-practical-elements-human-control-0 (visited 6 September 2023), at 19; Parasuraman and Manzey, *supra* note 6, at 390.

'information processing paradigm[s] that [are] inspired by the way biological nervous systems, such as the brain, process information'[38] — power more and more complex machines, which are able to make determinations so sophisticated (and at times even unexplainable) that humans effectively lack the ability to verify their correctness. The condition has been described as 'deep automation bias', whereby the human who is supposed to interact with the machine is de facto stripped of its autonomy to contest the machine's determinations.[39]

No wonder, then, why automation bias and complacency are such a concern in the setting of military operations. As demonstrated above with the example of the Patriot incident, when automated or autonomous capabilities are applied to weapons systems and employed to use potentially lethal force, the consequences of human overtrust in the correctness of the machine's determinations can be 'severe to catastrophic'.[40] The factors mentioned above in the context of 'civilian' tasks are all still present in military settings, some exerting even a higher pressure on human operators: in highly hierarchical military structures and knowing the amount of resources which went into the development of AWS, humans may be even more reluctant to exercise 'their own discernment' and contradict the machine.[41] Furthermore, the stress of combat operations may also heighten the risks of automation bias.[42] And, given the high sensitivity of battlefield decisions from an ethical standpoint, relying on the assessment made by an inanimate object may help to increase the perception of having behaved with reasonableness and legitimacy.[43] The danger posed by automation bias and complacency in military settings, as in other domains, is higher when the machine's capabilities are higher, and the tasks entrusted to it are more numerous and incisive. As aptly explained by Chesterman, even when a human oversees the machine's target selection process before lethal force is employed,

> If machines are able to make every choice up to that point — scanning an environment, identifying and selecting a target, proposing an angle of attack — the final decision may be an artificial one. . . .automation bias makes the default choice significantly more likely to be accepted in such circumstances.[44]

38 M. Senthilkumar, 'Use of Artificial Neural Networks (ANNs) in Colour Measurement', in M.L. Gulrajani (ed.), *Colour Measurement* (Woodhead Publishing, 2010) 125, at 125. See also, among others, S. Walczak, and N. Cerpa, 'Artificial Neural Networks', in R.A. Meyers (ed.), *Encyclopedia of Physical Science and Technology* (3rd edn., Academic Press, 2003) 631.

39 S. Strauß, 'Deep Automation Bias: How to Tackle a Wicked Problem of AI?', 5 *Big Data and Cognitive Computing* (2021), at 4.

40 Boulanin et al., *supra* note 37, at 19.

41 Seixas-Nunes, *supra* note 8, at 56.

42 Cummings, *supra* note 4, at 5.

43 ICRC, *Ethics and Autonomous Weapon Systems: An Ethical Basis for Human Control?*, 3 April 2018, available online at https://www.icrc.org/en/document/ethics-and-autonomous-weapon-systems-ethical-basis-human-control (visited 6 September 2023) at 13, referring to M. Cummings, 'Automation and Accountability in Decision Support System Interface Design', 32 *The Journal of Technology Studies* (2006) 23, at 29, on issue of moral buffering.

44 Chesterman, *supra* note 18, at 188. See also N. Milaninia, 'Biases in Machine Learning Models and Big Data Analytics: The International Criminal and Humanitarian Law Implications', 102 *International Review of the Red Cross* (2020) 199, at 216.

Current military practice is full of examples in which a human must oversee or validate machine choices. Examples include the mentioned Patriot surface-to-air missile system, employed by the United States.[45] The Patriot system can function both in a 'human-in-the-loop' configuration, where the operator has to authorize engagement after the system has identified a potential target, and in a 'human-on-the-loop' configuration, where the system will engage the identified target unless the human operator decides to halt it at certain decision leverage points. In both cases, after the missile has been launched and is in flight, the human operator can still abort the engagement.[46] Often, human supervision or validation is foreseen in automated defensive systems, like counter-rocket, artillery and mortar devices ('C-RAM') for which a human must validate or oversee — potentially vetoing by means of some sort of kill-switch — a target autonomously selected by the machine. One example is the SGR-A1 sentry robot, allegedly deployed by South Korea to patrol the Korea Demilitarized Zone.[47] The robot is capable of detecting threats in the area of deployment and engaging with them, though it is not clear whether engagement must be authorized by a human operator or whether instead it can occur autonomously, leaving to the human operator the power to veto it.[48] To make just one more example, the Phalanx system, deployed by the US Navy and other militaries to protect ships from incoming fire, is also able to autonomously search for, detect, evaluate, track and engage targets,[49] and provides for a human operator entrusted with aborting engagement if the case.[50]

Admittedly, these are not instances of 'fully autonomous' weapons systems. Yet, these are the instances in which automation bias and complacency on the part of the human operator, who is tasked with handling and overseeing incredibly complex systems, may affect targeting determinations[51] and potentially result in serious violations of international humanitarian law. Whether such violations may entail individual criminal responsibility for the human operator is the question to which I now turn.

3. Machines' Targeting Determinations and the Human Operator's *Mens Rea*

The starting point to analyse the impact of automation bias and complacency on the potential criminal responsibility of an individual for war crimes committed with an AWS is that the *actus reus* of a war crime is somehow

45 Hawley, *supra* note 3, at 4.

46 *Ibid.*, at 4–5.

47 A. Velez-Green, *The Foreign Policy Essay: The South Korean Sentry—A "Killer Robot" to Prevent War*, Lawfare, 1 March 2015, available online at https://www.lawfareblog.com/foreign-policy-essay-south-korean-sentry%E2%80%94-killer-robot-prevent-war (visited 6 September 2023).

48 *Ibid.*

49 See *supra*, note 14.

50 S. Welsh, *We Need to Keep Humans in the Loop When Robots Fight Wars*, The Conversation, available online at http://theconversation.com/we-need-to-keep-humans-in-the-loop-when-robots-fight-wars-53641 (visited 6 September 2023).

51 Seixas-Nunes, *supra* note 8, at 188.

imputable to a human being who interacted with the machine. There are a number of hypothetical scenarios in which human–machine interaction on the battlefield could lead to the commission of war crimes,[52] for instance, if unlawful targeting parameters have been intentionally coded in the machine.[53] Nonetheless, in a few such scenarios automation bias and complacency — both manifestations of overtrust in the determinations of a machine — are likely to play a role, except one: when a human is in charge of monitoring the operation of an automated or autonomous weapon system.[54] At its core, there are two basic variants to this scenario: a 'human in the loop' setting ('scenario A'); and a 'human on the loop' setting ('scenario B').[55]

Let us imagine, for instance, that an autonomous unmanned aerial vehicle (UAV) is deployed by one of the belligerents to patrol a certain area in the context of an armed conflict. The UAV, in this example, has been programmed to detect and, if the case, engage any 'target of opportunity' it may come across, i.e., 'a target visible to a surface or air sensor or observer, which is within range of available weapons and against which fire has not been scheduled or requested'.[56] Let us now imagine that the UAV does come across one such target — say, an object identified as an enemy military vehicle — and, in scenario A, it requests its human operator for the authorization to open fire, which is conceded; in scenario B, the UAV commences engagement, and the human operator — when given the opportunity to do so — decides not to override the machine's determination. After lethal force is used by the UAV, it turns out that the target was a civilian car.[57] The example could work also with other factual details: for instance, if a civilian person is wrongly identified as a person directly participating in hostilities, or if an enemy combatant who is wounded is wrongly not recognized as being *hors de combat*. In yet another variation to the hypothetical facts, it may well be that the target of opportunity was indeed a lawful military objective, but the engagement against it caused significant 'collateral' loss of civilians' life and damage to civilian objects nearby, which had not been correctly detected by the machine and, thus, not taken into account in the proportionality calculation.

In such examples, one should question whether the human operator's failure to correct the machine's error — possibly due to automation bias and complacency, among other factors — entails their individual criminal

52 McDougall, *supra* note 12, at 11–13.

53 *Ibid.*, at 12.

54 Cummings, *supra* note 4, at 3; Parasuraman and Manzey, *supra* note 6, at 382.

55 For more about these two models of human–machine interaction, see *supra*, Introduction. These and other scenarios of human–machine interaction with respect to the employment of lethal AWS in war are presented by Acquaviva, *supra* note 12, at 96. See also McDougall, *supra* note 12, at 13.

56 'Target of Opportunity', in *The Oxford Essential Dictionary of the U.S. Military* (Oxford University Press, 2002). On considerations about their pursuit by AWS, see Seixas-Nunes, *supra* note 8, at 185–186.

57 This example is a slightly modified version of one offered by T. McFarland, *Autonomous Weapon Systems and the Law of Armed Conflict: Compatibility with International Humanitarian Law* (Cambridge University Press, 2020), at 142.

responsibility for a war crime. Importantly, any finding of criminal responsibility will depend on the relevant crime definition, and on the required material and especially mental element. The mental element requirement will vary depending on the war crime in question, and variations may also occur between different national and international jurisdictions.

As an example, let us consider war crimes as defined in the Statute of the International Criminal Court (ICC). As it is well-known, in the ICC system, 'Unless otherwise provided, a person shall be criminally responsible and liable for punishment for a crime within the jurisdiction of the Court only if the material elements are committed with intent and knowledge'.[58] The mental element requirement of 'intent' is met when 'In relation to conduct, that person means to engage in the conduct' and 'In relation to a consequence, that person means to cause that consequence or is aware that it will occur in the ordinary course of events.'[59] For material elements of the crime amounting to a circumstance or a consequence, the ICC Statute requires 'knowledge', defined as 'awareness that a circumstance exists or a consequence will occur in the ordinary course of events'.[60] In addition, it is worth noting that the 'Existence of intent and knowledge can be inferred from relevant facts and circumstances.'[61]

Certain war crimes provisions in the ICC Statute, complemented by the corresponding Elements of Crimes, may add a volitional component to the said mental element requirement.[62] Thus, for instance, with respect to the war crime of intentionally directing attacks against civilians (Article 8(2)(b)(i) ICC Statute), the Elements of Crimes require that 'The perpetrator directed an attack', that 'The object of the attack was a civilian population as such or individual civilians not taking direct part in hostilities', and that 'The perpetrator intended the civilian population as such or individual civilians not taking direct part in hostilities to be the object of the attack'.[63] While several ICC judgments have interpreted this requirement flexibly, by conceding that the indiscriminate nature of an attack bears evidentiary weight when assessing the existence of the required intent to direct an attack against the civilian

58 Art. 30(1) ICCSt.

59 Art. 30(2) ICCSt. Importantly, the *Bemba* Confirmation Decision held that this standard of knowledge must be close to certainty, excluding the possibility of so-called *dolus eventualis*, i.e., the *mens rea* standard — known in some legal systems of civil law tradition — by which a person foresees that a proscribed result may occur as a result of their conduct and, while not intending that result to occur, still decides to engage in the conduct. See Decision Pursuant to Article 61(7)(a) and (b) of the Rome Statute on the Charges of the Prosecutor Against Jean-Pierre Bemba Gombo, *Bemba* (ICC-01/05-01/08-424), Pre-Trial Chamber II, 15 June 2009, §§ 357–363. See also K. Ambos, *Treatise on International Criminal Law - Volume I: Foundations and General Part* (2nd edn., Oxford University Press, 2021), at 374–377.

60 Art. 30(3) ICCSt.

61 ICC Elements of Crimes, General Introduction, para. 3.

62 Ambos, *supra* note 59, at 404; G. Werle and F. Jessberger, *Principles of International Criminal Law* (Oxford University Press, 2020), at 534, mn 1416.

63 ICC Elements of Crimes, Art. 8(2)(b)(i), paras 1–3. See also Art. 8(2)(e)(i) ICCSt., and the corresponding elements of the crime.

population, the fact remains nonetheless that the defendant must have — at the very least — been aware of the presence of civilians in the targeted area.[64]

If one accepts that a human operator in charge of validating or supervising the targeting choices of an autonomous weapon system can be said to be 'directing an attack' (potentially amounting to the *actus reus*), their failure to recognize the machine's erroneous target selection will likely negate knowledge of the civilian status of the victims of the attack and, in turn, intent to target them. For criminal responsibility to arise, the human would need to be aware of the factual circumstances establishing the civilian status of the target and, equipped with that knowledge, they would need to mean to engage in the conduct of attack against such target. It is unlikely that the operator's automation bias or complacency may have an impact on satisfying such stringent requirements because the erroneous machine's determination would in most cases negate the existence of such 'actual' knowledge.

To partially change how the law could be applied in such scenarios, McFarland has proposed that the requirement of knowledge as defined in Article 30(3) ICC Statute — which would apply also to knowledge of the target's civilian status in war crimes like that defined in Article 8(2)(b)(i) ICC Statute — could be read more broadly, allowing consideration of a defendant's 'wilful blindness'[65] towards information which would cast doubt on the correctness of the machine's determination.[66] And yet, even in the case in which the operator is found to be so unjustifiably complacent that, *de facto*, their state of mind can be equated to a sort of 'wilfully blind' knowledge of the factual circumstances establishing the civilian status of the target, intent for civilians to be the object of the attack could still be lacking.

A war crime for which automation bias and complacency could have potentially played a bigger role in excluding liability, in the ICC system, is the 'War crime of excessive incidental death, injury, or damage' listed in Article 8(2)(b)(iv) of the Statute.[67] The human operator's failure to recognize an

64 Judgment pursuant to Article 74 of the Statute, *Katanga* (ICC-01/04-01/07-3436-tENG), Trial Chamber II, 7 March 2014, § 802; Judgment, *Ntaganda* (ICC-01/04-02/06-2359), Trial Chamber VI, 8 July 2019, § 921. Both are discussed in Bo, *supra* note 6, at 283–284. In this sense also H. Olásolo, *Unlawful Attacks in Combat Situations: From the ICTY's Case Law to the Rome Statute* (Brill, 2008), at 218. Olásolo bases his reasoning on a reading of Art. 30 ICCSt., which would encompass '*dolus directus* in the second degree, where the perpetrator does not wish to cause the forbidden result but accepts its occurrence as a necessary consequence of the achievement of his main purpose'. *Contra*, Ambos argued that the attack must be aimed at harming civilians, requiring a purpose-based intent, i.e., *dolus directus* in the second degree. See K. Ambos, *Treatise on International Criminal Law - Volume II: The Crimes and Sentencing* (2nd edn., Oxford University Press, 2022), at 209–210.

65 That is, a situation in which 'the accused realised the high probability that the circumstance existed, but purposely refrained from obtaining the final confirmation because he or she wanted to be able to deny knowledge'. S. Finnin, 'Mental Elements under Article 30 of the Rome Statute of the International Criminal Court: A Comparative Analysis', 61 *International & Comparative Law Quarterly* (2012) 325, at 350.

66 McFarland, *supra* note 57, at 149. No case law at the ICC has endorsed this reading yet.

67 The name of this crime is taken from the rubric of the ICC Elements of Crimes for Art. 8(2)(b)(iv).

erroneous targeting determination by an autonomous weapon system would not negate, arguably, the material requirements that 'The perpetrator launched an attack' and that 'The attack was such that it would cause incidental death or injury to civilians or damage to civilian objects or widespread, long-term and severe damage to the natural environment and that such death, injury or damage would be of such an extent as to be clearly excessive in relation to the concrete and direct overall military advantage anticipated'.[68] What would be in play, especially in the most extreme cases of automation bias or complacency, is whether — based on inferences from the relevant facts and circumstances — the perpetrator could be said to have actually been aware that the attack would cause such clearly excessive damage.[69] Nonetheless, footnote 37 to the Elements of Crimes effectively precludes such scenario of constructive knowledge, by specifying that the element of awareness, in this case, requires the perpetrator to actually make the value judgment of excessiveness of the damage with respect to the military advantage.[70] Automation bias and complacency will likely mean that the defendant cannot be said to have made such value judgment.

As said, the mental element required for targeting-related war crimes may vary across legal systems. For instance, the International Committee of the Red Cross's (ICRC) commentary to Additional Protocol I affirms that, for the grave breach of 'making the civilian population or individual civilians the object of attack',[71] the required mental element of wilfulness includes 'recklessness', i.e. 'the attitude of an agent who, without being certain of a particular result, accepts the possibility of it happening.'[72] Notably, the International Criminal Tribunal for the former Yugoslavia (ICTY) has followed such broader definition when adjudicating this war crime.[73] Such definition would still exclude a finding of responsibility for human operators who had no reason to doubt the accuracy of the machine's targeting determination and, thus, could not foresee a wrongful target engagement. However, it could implicate a finding of responsibility for those who had, instead, foreseen the possibility of a machine's error and would have had reasons to be more cautious.[74] In this light, Ohlin

68 ICC Elements of Crimes, Art. 8(2)(b)(iv), paras 1–2.

69 ICC Elements of Crimes, Art. 8(2)(b)(iv), para 3.

70 A departure from the general rule set in ICC Elements of Crimes, General Introduction, para. 4.

71 Art. 85(3)(a) AP I.

72 Y. Sandoz, C. Swinarski, and B. Zimmermann, *Commentary on the Additional Protocols of 8 June 1977 to the Geneva Conventions of 12 August 1949* (ICRC, 1987), at § 3474. On the other side, the Commentary clarifies that 'ordinary negligence or lack of foresight is not covered, i.e., when a man acts without having his mind on the act or its consequences'. See also J.D. Ohlin, 'The Combatant's Stance: Autonomous Weapons on the Battlefield', 92 *International Law Studies* (2016) 1, at 23.

73 Judgement, *Galić* (IT-98-29-A), Appeals Chamber, 30 November 2006, § 140, Judgement and Opinion, *Galić* (IT-98-29-T), Trial Chamber I, 5 December 2003, § 54; Judgement, *Strugar* (IT-01-42-A), Appeals Chamber, 17 July 2008, § 270.

74 The real 'responsibility gap' within the ICC system, per A. Jain, 'Autonomous Cyber Capabilities and Individual Criminal Responsibility for War Crimes', in R. Liivoja and A. Väljataga (eds), *Autonomous Cyber Capabilities under International Law* (NATO CCDCOE Publications, 2021) 291, at 299–300.

has proposed that a way to ensure accountability for crimes committed when deploying AWS could be to introduce a new criminal offence of 'recklessly perpetrating an international crime'.[75]

National legislation on war crimes may provide for even different definitions of the mental element required for targeting-related war crimes. A state may decide, for instance, to draft the relevant offence with a requirement to prove not the human operator's intent or recklessness, but negligence — for which automation bias and complacency could be the root causes. The concept of negligence as a basis for criminal responsibility is defined in different ways across different legal systems. In English criminal law, for instance, negligence designates the situation of a defendant who behaved below the standard expected of a reasonable person in the circumstances by taking an unjustifiable risk,[76] irrespective of whether they were aware of that risk[77] — including in cases in which they had 'constructive notice' of it.[78] The United States' Model Penal Code (MPC), at section 2.02(2)(d), proposes a different definition:

> A person acts negligently with respect to a material element of an offense when he should be aware of a substantial and unjustifiable risk that the material element exists or will result from his conduct. The risk must be of such a nature and degree that the actor's failure to perceive it, considering the nature and purpose of his conduct and the circumstances known to him, involves a gross deviation from the standard of care that a reasonable person would observe in the actor's situation.[79]

While the MPC's proposed definition introduces a gravity threshold in requiring a 'gross deviation' from a reasonable person's standard of care, it otherwise mirrors the definition adopted in English criminal law in requiring the defendant's failure to perceive the risk in question. In fact, the concept of negligence is tied to that of *culpa*, which has been theorized as 'a culpable failure to be aware of the unreasonable risk entailed in one's conduct'.[80] The concept of negligence — defined along the mentioned lines — may be helpful in capturing the criminal responsibility of a human operator who failed to recognize a machine's targeting error in the course of an armed conflict, even if that failure is due to automation bias or complacency. The human operator could be found to be negligent if — in their interaction with the machine — they took an unjustifiable risk or grossly deviated from a reasonable person's standard of care.

75 Ohlin, *supra* note 72, at 28.

76 J. Herring, *Criminal Law: Text, Cases, and Materials* (10th edn., Oxford University Press, 2022), at 148; D. Ormerod and K. Laird, *Smith, Hogan, and Ormerod's Criminal Law* (Oxford University Press, 2021), at 115–116 and 136–137.

77 Herring, *supra* note 76, at 148; J. Horder, *Ashworth's Principles of Criminal Law* (10th edn., Oxford University Press, 2022), at 227; Ormerod and Laird, *supra* note 76, at 115–116.

78 *Flintshire County Council v Reynolds* [2006] EWHC 195 (Admin), at [17] in Ormerod and Laird, *supra* note 76, at 138.

79 American Law Institute, *Model Penal Code: Official Draft and Explanatory Notes*, adopted at the 1962 annual meeting of the American Law Institute at Washington, D.C., 24 May 1962, s. 2.02(2)(d).

80 J. Blomsma, *Mens Rea and Defences in European Criminal Law* (Intersentia, 2012), at 166.

Of note, some scholars have considered whether the notion of command responsibility — for instance, as defined in Article 28 of the ICC Statute — could be helpful to establish criminal responsibility for commanders who failed to prevent breaches of international humanitarian law caused by AWS they deployed,[81] possibly due to the commander's automation bias. The orthodox view on the law of command responsibility would preclude such reading: a commander would only be responsible when crimes are committed by 'imputable' subordinates who are natural persons. Yet, an alternative and currently minoritarian approach has been proposed by Buchan and Tsagourias.[82] Their proposal stems from the assumption that the notion of crime can be 'decoupled from the notion of culpability' and simply means 'a proscribed act', regardless of its attribution to a moral agent.[83] In fact, they note, a finding of responsibility for the commander does not necessitate that they actually knew the specific details of how the crime in question was committed, or even the identity of their subordinates who perpetrated the crime.[84] Additionally, the duty of the commander to act with a view to preventing ongoing crimes arises even when the crime is unfolding and its elements (including the subordinates' *mens rea*) have not fully materialized,[85] and concerns also inchoate crimes and crimes committed under exculpatory circumstances.[86] While this approach presents some appeal to ensure individual accountability — and, inter alia, avoid that automation bias and complacency may increase the likelihood of a lack of such accountability — it has not yet received support in relevant practice and case law. In addition, such reading rests on the interpretation of superior responsibility not as a special form of liability accessory to the crime committed by a principal perpetrator (the subordinate), but as a criminal offence 'per se', a form of dereliction of duty — an interpretation which is, in itself, open to question.[87]

Perhaps, a more promising avenue to employ superior responsibility concerns not the superiors' own automation bias and complacency, but their failure to exercise control properly over forces under their control who may

81 For instance, Ohlin, *supra* note 72, at 26; Seixas-Nunes, *supra* note 8, at 222–223. See also A. Spadaro, 'A Weapon Is No Subordinate: Autonomous Weapon Systems and the Scope of Superior Responsibility', in this special issue of the *Journal of International Criminal Justice*.

82 R. Buchan and N. Tsagourias, 'Autonomous Cyber Weapons and Command Responsibility', in Liivoja and Väljataga (eds), *supra* note 74, 321.

83 *Ibid.*, at 345.

84 Judgement, *Orić* (IT-03-68-A), Appeals Chamber, 3 July 2008, § 35, affirming however that the subordinates' existence must be demonstrated. See also Judgment Pursuant to Article 74 of the Statute, *Bemba* (ICC-01/05-01/08-3343), Trial Chamber III, 21 March 2016, § 194.

85 Buchan and Tsagourias, *supra* note 82, at 345, citing (and perhaps reading too much into) Judgement, *Hadžihasanović and Kubura* (IT-01-47-T), Trial Chamber, 15 March 2006, § 852.

86 *Ibid.*, at 345–346.

87 For a recent discussion of this possibility, see R. Arnold and M. Jackson, 'Article 28: Responsibility of Commanders and Other Superiors', in K. Ambos (ed.), *Rome Statute of the International Criminal Court: Article-by-Article Commentary* (Beck/Hart/Nomos, 2022) 1280, at 1285–1289, §§ 6–16, and at 1304–1306, §§ 48–51. See also M. Jackson, 'Causation and the Legal Character of Command Responsibility after *Bemba* at the International Criminal Court', 20 *Journal of International Criminal Justice* (2022) 437.

display automation bias and complacency when operating or supervising AWS. Even such an option, however, would require that the subordinates (in this example, AWS operators) can be said to have 'committed crimes'. Where the subordinates do not meet the required mental element (as explained above), it is doubtful that such conclusion can be reached.

4. Reliance on the Machine's Determination as a Defence of Mistake of Fact

As already mentioned, the human operator of an AWS may fail to perceive that the machine's targeting-related determination is erroneous. It could be said that the operator, in such circumstances, has incurred a mistake of fact — possibly due to automation bias or complacency — about the correctness of the machine's determination. Commentators have hypothesized that such mistake of fact could be successfully invoked at trial, to be exonerated from criminal responsibility.[88] Thus, it is important to understand how this result could be brought about.

A mistake of fact, in ordinary language, could be defined as a false or erroneous representation of a fact, as a result of which the actor wrongly assumes to be behaving lawfully. In the lead example for the present article, the fact would be the existence of a lawful military target, which the human operator would erroneously represent to be correct based on a machine-derived information or determination. A mistake of fact, thus, consists of the missed recognition of an existing fact or of the erroneous assumption that a non-existing fact actually exists.[89] Usually, mistakes of fact operate as a '*mens rea*-negating' defence — in essence, precluding a finding that the required mental element for the offence is present.[90] Therefore, in the presence of such a mistake, no crime can be said to have been committed.[91] Indeed, this defensive argument is also at times labelled as a 'failure-of-proof' defence, because the prosecutor would fail to prove all the elements of the crime beyond reasonable doubt. This rule on the defensive relevance of mistakes is reflected in Article 32 of the ICC Statute and commonly envisaged both in countries of common law tradition[92] and of civil law tradition.[93]

As mentioned in the previous section, it can often be the case that a human operator's mistake on the correctness of a machine's targeting determination

88 See e.g., Bo, *supra* note 6, at 297–298; McDougall, *supra* note 12, at 11.

89 A. Eser, 'Mental Elements – Mistake of Fact and Mistake of Law', in A. Cassese, P. Gaeta, and J.R.W.D. Jones (eds), *The Rome Statute of the International Criminal Court: A Commentary*, vol. I (Oxford University Press, 2002) 889, at 936.

90 E. Keedy, 'Ignorance and Mistake in the Criminal Law', 22 *Harvard Law Review* (1908) 75, at 84–85.

91 G. Fletcher, 'The Right and the Reasonable', 98 *Harvard Law Review* (1985) 949, at 962.

92 See A. Simester et al., *Simester and Sullivan's Criminal Law: Theory and Doctrine* (6th edn., Hart, 2016), at 705-706. See also the US Model Penal Code, Section 2.04.

93 See e.g. Art. 47, Italian Criminal Code; § 16, German Criminal Code; Art. 13, Swiss Criminal Code.

negates the presence of the required mental element — which obviously will vary depending on the crime definition which is being applied in international or national law. It is difficult to unqualifiedly say 'this mistake negates intent' or 'this mistake negates knowledge', because these expressions may have a different meaning in different legal systems. Likely, however, a mistake due to automation bias or complacency will negate intent or knowledge as understood in the ICC Statute. The factual circumstances establishing, for instance, the protected status of a civilian are examples of those labelled in Article 30(3) ICC Statute as 'circumstances' of which the perpetrator must be aware: a mistake of fact negating such awareness would mean that the required *mens rea* is not present.[94]

The mistake is assessed subjectively and, thus, does not need to be 'reasonable' in order to produce an exonerating effect: no matter how unreasonable the human operator's understanding of the factual circumstances, it will still deny that they possess the knowledge and intent required in the ICC system.[95] However, at the very least, the human operator's belief in the correctness of the machine's determination must be honestly held, and not falsely raised as a pretext to be exonerated.[96] When the mistake looks objectively unreasonable, this could be used to infer a lack of honesty.[97] Of note, not every mistake of fact on a crime element negates the *mens rea*, but only mistakes on facts that are not equivalent in terms of the crime element. For instance, erroneously attacking a particular civilian while believing that the machine is targeting a different civilian does not negate the intent to direct the attack against a civilian.[98] On the contrary, erroneously attacking a civilian instead of an enemy combatant negates the presence of the required intent.[99]

In addition, and importantly, a mistake of fact induced by automation bias and complacency may also preclude the human operator from being aware of the risk that the machine's targeting determination was incorrect in the

94 E. van Sliedregt, *Individual Criminal Responsibility in International Law* (Oxford University Press, 2012), at 271. It has been persuasively noted that this reading makes Art. 32(1) ICCSt. redundant, as it would merely apply to a specific case (a mistake of fact) the general and logical consideration that a defendant cannot be convicted if the mental element required for the crime has not been satisfied. See *Ibid.*, at 282; Eser, *supra* note 90, at 934.

95 A. Coco, *The Defence of Mistake of Law in International Criminal Law: A Study on Ignorance and Blame* (Oxford University Press, 2022), at 140–142; K.J. Heller, 'Mistake of Legal Element, the Common Law, and Article 32 of the Rome Statute: A Critical Analysis', 6 *Journal of International Criminal Justice* (2008) 419, at 440; M. Milanovic, 'Mistakes of Fact When Using Lethal Force in International Law: Part I', EJIL: Talk!, 14 January 2020, available online at https://www.ejiltalk.org/mistakes-of-fact-when-using-lethal-force-in-international-law-part-i/ (visited 6 September 2023).

96 Coco, *supra* note 95, at 140; Heller, *supra* note 95, at 443.

97 Milanovic, *supra* note 95.

98 On the issues of *aberratio ictus* and 'transferred intent' (holding the criminal responsibility of a defendant who intended to harm an individual but inadvertently harms a different individual instead), see A. Greipl, 'Data-driven Learning Systems and the Commission of International Crimes: Concerns for Criminal Responsibility?', in this special issue of the *Journal of International Criminal Justice*, at Section 4(B)(2).

99 Eser, *supra* note 89, at 938.

circumstances. This would negate the presence of the required mental element even when more broadly defined — for instance in the terms of 'recklessness' used above by the ICRC Commentary and the ICTY case law. Unless one interprets 'recklessness' so broadly as to include cases of 'inadvertent' risk taking, the mistake in question would exonerate the human operator from criminal responsibility.[100] On the contrary, a mistake induced by automation bias or complacency may not negate negligence as defined above, i.e., as inadvertent taking of an unjustifiable risk, in deviation from the standard of conduct expected of a reasonable person in the circumstances — if the automation bias or complacency determine such a deviation.

5. The Way Ahead

Automation bias and complacency exacerbate the factual and legal difficulties with establishing individual criminal responsibility for targeting-related war crimes committed by employing AWS. These two manifestations of overtrust, which I have examined in this article, act on the state of mind and attitude of the human interacting with the machine, further limiting the possibility that they meet the mental element requirement for targeting-related war crimes. A finding of criminal responsibility, when the human has validated or failed to correct an erroneous targeting determination by the machine, will be confined to cases in which the applicable law establishes that a particular conduct amounts to an offence with a requirement of inadvertent recklessness or negligence as described above — assuming that the presence of the relevant material elements can also be proven.

Resort to automated and autonomous machines is intended to aid humans and reduce workload and errors.[101] Yet, when the machines' errors — even if rare and exceptional — are not identified as such, the ensuing human error is harder to accept.[102] It is worth for designers and developers of automated and autonomous weapon systems to keep focusing on how to avoid automation bias and complacency, by striving to make machines' errors more and more easily detectable by human operators, for instance by improving the machines' interface design. More procedural safeguards can be put in place by way of legislation or regulation, for instance by requiring human operators to provide a full explanation of their reasoning for deciding to validate or not to veto a targeting determination made by a machine.[103]

100 In this sense also Milanovic, *supra* note 95.

101 Cummings, *supra* note 4, at 5.

102 For a study on how people react to undesirable outcomes caused by AI-made decisions, see S.M. Jones-Jang and Y.J. Park, 'How Do People React to AI Failure? Automation Bias, Algorithmic Aversion, and Perceived Controllability', 28 *Journal of Computer-Mediated Communication* (2023) 1.

103 One of the proposals made in Fair Trials, *Automating Injustice: The Use of Artificial Intelligence & Automated Decision-Making Systems in Criminal Justice in Europe* (2021), available online at https://www.fairtrials.org/articles/publications/automating-injustice/ (visited 6 September 2023), at 26.

Of course, the first and foremost measure to be adopted is adequate training of human operators, to ensure that they are aware of the risks of machine errors and are equipped with the tools to recognize and respond to such errors. Training humans to 'undertrust' the machine — i.e., not to rely too much on the machine's determination — has been hailed as an important measure,[104] but comes with some dangers of its own. It could lead to inconsistency of practice and generate different types of incidents, whereby the human would not trust what is instead a correct machine assessment.[105] Undertrust in a computer's assessment of a factual situation, for instance, has been reported as one of the causes of the shooting down of the Iran Air Flight 655 in 1988, by means of the Aegis Combat System on the US Navy warship USS Vincennes.[106]

Proper training to recognize machines' errors can mitigate the danger of both overtrust and undertrust.[107] Yet, training alone cannot counter such danger completely, if not accompanied by organizational and technical measures aimed at facilitating the work of human operators and helping them to reach better decisions.[108]

Furthermore, it may be worth considering, for lawmakers at the national or international level, whether to introduce a lower mental element requirement (like negligence as defined above) or to set no mental element requirement at all (strict liability) for crimes committed by employing automated or autonomous weapon systems. Such legislative measures could include lower sentences than those foreseen for crimes committed intentionally, reflecting a different level of blameworthiness. Such changes may contribute to incentivizing human operators to increase standards of care. Early studies on the topic, in fact, suggested that when human operators are held accountable for the incorrectness of their decisions in validating or supervising machines' determination, this has led to a decrease in the occurrences of automation bias[109] — though this conclusion is not uncontroversial.[110] But the suggestion remains appealing, especially if one considers that a contributing factor for an individual's automation bias (or any reduction of their own effort on the workplace, really) is the feeling that responsibility for a choice rests not with them, but with others — the machine, in this case.[111]

104 As discussed in Crootof, Kaminski, and Price II, *supra* note 1, at 500–501.

105 *Ibid.*, at 500–501.

106 D. Linnan, 'Iran Air Flight 655 and Beyond: Free Passage, Mistaken Self-Defense, and State Responsibility', 16 *Yale Journal of International Law* (1991) 245, at 252–254.

107 Goddard, Roudsari, and Wyatt, *supra* note 6, at 125; Mosier et al., *supra* note 24; Parasuraman and Manzey, *supra* note 6, at 387.

108 In this sense Zerilli et al., *supra* note 6, at 575.

109 L. Skitka, K. Mosier, and M. Burdick, 'Accountability and Automation Bias', 52 *International Journal of Human-Computer Studies* (2000) 701, at 701; Parasuraman and Manzey, *supra* note 6, at 396; Zerilli et al., *supra* note 6, at 574.

110 Goddard, Roudsari, and Wyatt, *supra* note 6, at 125; Parasuraman, and Manzey, *supra* note 6, at 396. Adopting a lower *mens rea* or a strict liability standard may open a debate concerning respect for the principle of individual culpability, but this is beyond the scope of this article.

If achieving 'meaningful human control' is a goal in the current debates on the employment of artificial intelligence on the battlefield, then it is fundamental to work not only on the 'machine' side of things, but also on the 'human' side — and keep humans accountable for errors made in circumstances in which they possessed the tools to exercise that so much coveted control.

111 Adensamer, Gsenger, and Klausner, *supra* note 6, at 4; Parasuraman and Manzey, *supra* note 6, at 392.