HARVARD LAW SCHOOL
NATIONAL SECURITY JOURNAL

ONLINE ARTICLE

Countering the "Humans vs. AWS" Narrative and the Inevitable Accountability Gaps for Mistakes in Targeting: A Reply to Kevin Jon Heller[*]

Dr. Marta Bo[†]

[*] This article is part of a symposium on Kevin Jon Heller's "The Concept of "the Human" in the Critique of Autonomous Weapons," published in this journal in 2023. All articles in the symposium can be found in the Harvard National Security Journal Online at https://harvardnsj.org/onlineedition.

[†] Dr. Marta Bo is a Senior Researcher in International Law at the Asser Institute and an Associate Senior Researcher at SIPRI. Her expertise lies in international criminal law, international humanitarian law, and the legal implications of military applications of artificial intelligence. Her research focuses on compliance and accountability frameworks, as well as the role of AI in the investigation and prosecution of international crimes.

Table of Contents

I. INTRODUCTION

In this reply, I challenge the "Humans vs. AWS" narrative, which claims that AWS will achieve unprecedented targeting accuracy compared to humans. By highlighting the flaws in this comparison, I also dispute the idea that there will always be gaps in accountability for mistakes in war, whether they are caused by AI or not.

In his article "The Concept of "the Human" in the Critique of Autonomous Weapons,"[1] Kevin Jon Heller argues in favor of the use of autonomous weapons systems (AWS) and makes several salient points regarding the limitations of humans in targeting. His article raises important concerns about human cognitive biases, negative emotions, and psychological constraints. He argues that AWS would not be subject to these limitations, thus presenting several advantages in terms of compliance with international humanitarian law (IHL). According to Heller, this and other features of AWS would lead to increased accuracy in targeting relative to human targeting — a recurring argument made in the AWS debate.[2]

The objection put forward by Heller to the accountability gap argument is based on a comparison between mistakes resulting from the use of AWS and human mistakes. The AWS-created "accountability gap" refers to the difficulty of attributing criminal responsibility for war crimes to individuals who have no control or reduced control over weapon systems with a certain degree of autonomy in the targeting process. When humans have no intent or knowledge with respect to the development and use of military AI, accidents involving the AI fall outside the scope of war crimes definitions.[3] In Heller's view, the lack of accountability for those instances is no different from the "lack of accountability for '*a human soldier accidentally or mistakenly commit[ing] the* actus reus *of a war crime*'" – an event considered a non-criminal accident in war.[4] AWS-created accountability gaps are a concern only if "larger than the ones that currently exist when states use only human soldiers."[5]

It is undeniable that human decision-making has limitations and an accountability gap already exists. The current gap even extends beyond the conduct of hostilities war crimes that Heller mentions.[6] However, we should not accept this accountability gap as inevitable. Nor should we accept the narrative that the relative accountability gap in human warfare and AWS warfare can be cleanly compared.

Assigning credit or blame to AWS for their accuracy is a fraught exercise. The implementation of military AI on the battlefield involves an increased use of AI-enabled

---

[1] Kevin J. Heller, *The Concept of "the Human" in the Critique of Autonomous Weapons*, 15 HARV. NAT'L SEC. J. 1 (2023).

[2] *See, e.g.*, Jeffrey S. Thurnher, *Means and Methods of the Future: Autonomous Systems*, *in* TARGETING: THE CHALLENGES OF MODERN WARFARE 177, 184–185 (Paul A. L. Ducheine, Michael N. Schmitt & Frans P. B. Osinga eds., 2016); Herwin W. Meerveld et al., *The Irresponsibility of Not Using AI in the Military*, 25 ETHICS AND INFO. TECH. 13, 14–15 (2023).

[3] *See* 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*, 19 J. INT'L CRIM. JUST. 275, 279–298 (2021) [hereinafter Bo, *Autonomous Weapons*].

[4] Heller, *supra* note 1, at 65.

[5] *Id*. at 62.

[6] *See generally* Paola Gaeta, *Serious Violations of the Law on the Conduct of Hostilities: A Neglected Class of War Crimes?*, *in* WAR CRIMES AND THE CONDUCT OF HOSTILITIES: CHALLENGES TO ADJUDICATION AND INVESTIGATION 20, 20–37 (Fausto Pocar et al. eds., 2013).

decision support systems (AI-DSS).[7] This trend suggests the emergence of an interconnected battlefield where manned and unmanned as well as autonomous and non-autonomous platforms work together.[8] In other words, AWS exist as part of complex socio-technical systems. AI in targeting is not standalone and should not be seen in contradistinction to human targeting, but rather as intermediating human decision-making.[9] This should prompt a reconsideration of what accuracy and failure rates in targeting mean, as it is increasingly difficult to separate humans from machines, and human errors from machine errors.

Since AWS errors in war are likely to be caused by a series of linked errors, we should approach the accountability gap as a problem of accountability for the entire socio-technical system's errors — which originate from both humans and machines. Therefore, one of the objectives of this reply is to counter the narratives of "humans vs. AWS" and "human errors vs. AWS errors."

To address the implications for accountability, this article will first address why a comparison between humans and AWS is a simplification that does not account for the complexities of actual and intended integration of AI in military use cases. Second, it will discuss accuracy and errors in human-machine assemblages where agency and control are distributed and mediated by AI. Third, it will address the implications for criminal responsibility for war crimes when these result from the use of AI in targeting. The accountability framework for war crimes should account for the nature of errors in war conducted with AI. To this end, the framework's main pillars need to be rethought to prevent making the accountability gap inevitable and to prevent misplacement of responsibility.

II. A REBUTTAL OF THE "HUMANS VS. AWS" NARRATIVE

One of the main recurrent points supporting the use of AWS in Heller's article revolves around the prospect of unparalleled targeting accuracy, which could result in better compliance with the IHL principles of distinction and proportionality.[10] As Heller puts it, this is "because [of] their non-human nature [and] their lack of dependence on human ability."[11] However, current and foreseen developments in the uses of military AI run counter to this narrative. Military AI use cannot be thought of as a single automated weapons system. Rather, human-machine teaming is the go-to approach to the integration of military AI undertaken by many

[7] *See, e.g.*, Yuval Abraham, '*A Mass Assassination Factory': Inside Israel's Calculated Bombing of Gaza*, +972 MAGAZINE (Nov. 30, 2023), https://www.972mag.com/mass-assassination-factory-israel-calculated-bombing-gaza/ [hereinafter Abraham, *Mass Assassination*]; Yuval Abraham, *'Lavender': The AI Machine Directing Israel's Bombing Spree in Gaza,* +972 MAGAZINE (April 3, 2024), https://www.972mag.com/lavender-ai-israeli-army-gaza/ [hereinafter Abraham, *Lavender*]; Vera Bergengruen, *How Tech Giants Turned Ukraine Into an AI War Lab*, TIME (Feb. 8, 2024), https://time.com/6691662/ai-ukraine-war-palantir/ [hereinafter Bergengruen, *War Lab*].

[8] *See generally* BRIGADIER GENERAL Y.S, THE HUMAN-MACHINE TEAM: HOW TO CREATE SYNERGY BETWEEN HUMAN AND ARTIFICIAL INTELLIGENCE THAT WILL REVOLUTIONIZE OUR WORLD (2021); Joseph Clark, *Hicks Announces Delivery of Initial CJADC2 Capability*, DOD NEWS (Feb. 21, 2024), https://www.defense.gov/News/News-Stories/Article/Article/3683482/hicks-announces-delivery-of-initial-cjadc2-capability/.

[9] *See* Dimitri van den Meerssche, '*The Time Has Come for International Regulation on Artificial Intelligence*' – *An Interview with Andrew Murray*, OPINIO JURIS (Nov. 15, 2020), http://opiniojuris.org/2020/11/25/the-time-has-come-for-international-regulation-on-artificial-intelligence-an-interview-with-andrew-murray/.

[10] *See* Heller, *supra* note 1, at 20–31.

[11] *Id*. at 11.

states. This has been in practice for several years now[12] and recent state policies confirm that human-machine teaming is the "default approach" to the current and future integration of AI.[13] The real and pertinent question is thus not whether machines are "better" than humans but rather how human-machine teaming is operating currently, how it is likely to be configured in the future, and the consequences of these configurations. Recent developments could give us some answers.

Systems like the "Gospel," the "Lavender," and AIP software[14] demonstrate growing reliance on AI-DSS in the targeting cycle.[15] These systems are increasingly relied upon for the processing of data sources such as human intelligence, drone footage, intercepted communications and, in some cases, data gathered in real-time through other Intelligence Surveillance and Reconnaissance (ISR) capabilities. These systems are used, among other things, as decision aids for the production, identification, and even nomination of targets.[16]

As noted by Turek and Moyes, while there is a structural difference between an AWS and an AI-enabled DSS, they operate in a similar way up to a certain point (i.e., the application of force): "the system is proposing targets based on the automatic matching of sensor inputs against [pre-encoded] target profiles."[17] As such, they display, in principle, a certain degree of human control along with human-machine interaction because they allow the human to play a role before the application of force. However, "[w]hile a human theoretically remains in the loop, there are uncertainties regarding the extent to which humans truly maintain meaningful control or exercise judgment within these military decision-making processes."[18] Therefore, some challenges posed by AI-enabled DSS and AWS are similar.[19] Finally, it is also important to note that DSS sensor-based targeting systems could also be shifted and modified from human-in-the loop systems to out-of-the loop systems (AWS).[20]

---

[12] *See* Paul Scharre, *Centaur Warfighting: The False Choice of Humans vs. Automation*, 30 TEMPLE INT'L & COMP. L. J. 151, 151–159 (2016); Bo, *Autonomous Weapons*, *supra* note 3, at 276, 279–281.

[13] UK MINISTRY OF DEFENCE, DEFENCE ARTIFICIAL INTELLIGENCE STRATEGY 15 (2022), https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1082416/Defence_Artificial_Intelligence_Strategy.pdf. The U.S Department of Defense asserts that "[m]ilitary operations of the future will require collaboration between unmanned systems and humans" and emphasizes the importance of "human-machine teaming." THE US DEPARTMENT OF DEFENSE, UNMANNED SYSTEMS INTEGRATED ROADMAP 2017–2042 (2018).

[14] *See* Abraham, *Mass Assassination*, *supra* note 7; Abraham, *Lavender*, *supra* note 7; Bergengruen, *War Lab*, *supra* note 7.

[15] Klonowska defines AI-DSS as "tools that use AI techniques to analyse data, provide actionable recommendations," and "assist decision-makers situated at different levels in the chain of command to solve semi-structured and unstructured decision tasks." Klaudia Klonowska, *Article 36: Review of AI Decision-Support Systems and Other Emerging Technologies of Warfare*, 23 Y.B. OF INT'L HUMANITARIAN L. 123, 124 (2020).

[16] Abraham, *Mass Assassination*, *supra* note 7.

[17] ANNA TUREK & RICHARD MOYES, ARTICLE 36, SENSOR-BASED TARGETING SYSTEMS: AN OPTION FOR REGULATION 2, (Nov. 2021), https://article36.org/wp-content/uploads/2022/01/Sensor-based-targeting.pdf.

[18] On the difficulties of exercising human judgment, see Marta Bo & Jessica Dorsey, *The 'Need' for Speed – The Cost of Unregulated AI Decision-Support Systems to Civilians*, OPINIO JURIS (Apr. 4, 2024), https://opiniojuris.org/2024/04/04/symposium-on-military-ai-and-the-law-of-armed-conflict-the-need-for-speed-the-cost-of-unregulated-ai-decision-support-systems-to-civilians/; *see also AI-Enabled Decision-Support Systems in the Joint Targeting Cycle: Legal Challenges, Risks, and the Human(e) Dimension*, INT'L L. STUD. (forthcoming 2025).

[19] Turek & Moyes, *supra* note 17, at 2.

[20] Beyond the shift from non-autonomous to autonomous modes, it has been noted that AI will serve to enhance current systems rather than operate as a standalone solution. In other words, AI is intended to function as an add-on to existing systems. Jan Maarten Schraagen, *Responsible Use of AI in Military Systems: Prospects and Challenges*, 66 ERGONOMICS 1719, 1720 (2023).

This implies that concerns around military AI should not center around AWS as a single, stand-alone AI application. Rather, they should focus on "complex systems that rely on seamless interaction between multiple types of tightly coupled sensors, algorithms, actuators and *human factors*."[21] In this context, data-driven AI methods (e.g., machine learning and deep learning) and their features (training data, the learning program, and the system architecture) are part of a larger system made of physical, computational, and human components.[22]

One important feature of these combinations of human and machine decision-making is the distribution of agency and control.[23] As noted by Bode and Nadibaidze, "we are likely to encounter a form of *distributed agency* in military decisions that rely on human-machine interaction" that "involves a blurring of the distinction between instances of 'human' and 'AI' agency'" and new networks of interactions and agents.[24] Distributed agency and the complex networks of agents involved in the development and use of military AI also entail that control is distributed in the sense that there are "multiple control points across the lifecycle of AWS,"[25] including by programmers.[26]

But how are agency and control distributed? The network of agents involved in the use of AWS and other AI-enabled systems is complex not only in light of the number of agents involved but also because new webs of interactions are created. On the one hand, human choices shape AI technologies in different phases of the lifecycle of the weapon and of the targeting process.[27] "[P]rogrammers … create the basic algorithmic parameters, workers … prepare the data that training machine learning algorithms requires through a series of iterative micro-tasks often subsumed as 'labelling data'," and people provide "data [that] is used to train such algorithms."[28] On the other hand, 'Lavender', 'Gospel' and other AI-DSS also show how AI is integrated to *influence human choices and targeting decisions*. Arguably, these systems

[21] ARTHUR HOLLAND MICHEL, KNOWN UNKNOWNS: DATA ISSUES AND MILITARY AUTONOMOUS SYSTEMS 8 (2021) (emphasis added).
[22] *See generally* MICHÈLE A. FLOURNOY ET AL., BUILDING TRUST THROUGH TESTING: ADAPTING DOD'S TEST & EVALUATION, VALIDATION & VERIFICATION (TEVV) ENTERPRISE FOR MACHINE LEARNING SYSTEMS, INCLUDING DEEP LEARNING SYSTEMS 9 (2020), https://cset.georgetown.edu/wp-content/uploads/Building-Trust-Through-Testing.pdf.
[23] As noted by Boutin and Woodcock, "[c]urrent approaches to MHC often frame a binary dichotomy between the autonomy of a system and the control of the human operator, failing to recognize that human-machine relationships and military processes are complex, distributed, intermediated, and multidimensional." Bérénice Boutin & Taylor Kate Woodcock, *Aspects of Realizing (Meaningful) Human Control: A Legal Perspective*, *in* RESEARCH HANDBOOK ON WARFARE AND ARTIFICIAL INTELLIGENCE 187 (Robin Geiss and Henning Lahmann eds., 2024).
[24] Ingvild Bode & Anna Nadibaidze, *Human-Machine Interaction in the Military Domain and the Responsible AI Framework*, OPINIO JURIS (Apr. 4, 2024), http://opiniojuris.org/2024/04/04/symposium-on-military-ai-and-the-law-of-armed-conflict-human-machine-interaction-in-the-military-domain-and-the-responsible-ai-framework/.
[25] Lauren Sanders, *Bridging the Legal Gap Between Principles and Standards in Military AI – Assessing Australia's 'System of Control' Approach*, OPINIO JURIS (Apr. 3, 2024), http://opiniojuris.org/2024/04/03/symposium-on-military-ai-and-the-law-of-armed-conflict-bridging-the-legal-gap-between-principles-and-standards-in-military-ai-assessing-australias-system-of-control-appro/.
[26] *See generally* Marta Bo, *Are Programmers In or Out of' Control? The Individual Criminal Responsibility of Programmers of Autonomous Weapons and Self-Driving Cars*, *in* HUMAN-ROBOT INTERACTION IN LAW AND ITS NARRATIVES: LEGAL BLAME, CRIMINAL LAW, AND PROCEDURE 23–47 (Sabine Gless & Helena Whalen-Bridge eds., 2024) [hereinafter Bo, *Criminal Responsibility*].
[27] *See* Ingvild Bode, *Falling Under the Radar: The Problem of Algorithmic Bias and Military Applications of AI*, OPINIO JURIS (Mar. 14, 2024), https://blogs.icrc.org/law-and-policy/2024/03/14/falling-under-the-radar-the-problem-of-algorithmic-bias-and-military-applications-of-ai/.
[28] Bode & Nadibaidze, *supra* note 24.

might also hold the potential to inform and shape broader strategic and operational military decision-making. Rather than "centaur warfighting" where "human beings direct … multiple robots," researchers Sparrow and Henschke coined the idea of "Minotaur warfighting *where 'artificial intelligences (AI) will direct the activities of multiple human beings*.'"[29] Woodcock warns that "[t]he use of algorithmic DSS within complex environments can …hamper users' autonomy by shaping their choices."[30] Klonowska pointedly observes that the design and use of algorithmic tools influences targeting decisions and affects the reasonableness of decision-making in targeting.[31] Elsewhere, Jessica Dorsey and I have shown how the integration of AI-DSS into the Joint Targeting Cycle, and consequently, the speed and scale of AI-enabled target production and nomination, might affect compliance with IHL.[32]

Rather than direct our attention to the comparison of "humans vs. machines," we should turn towards understanding how agency and control is distributed and configured within these complex socio-technical systems. We should understand that AI can affect and even replace human decision-making, including targeting choices and broader strategic military decisions. This has implications for how errors in targeting arise and raises questions regarding how accountability frameworks for war crimes account for the role of AI.

III. A REBUTTAL OF THE "AWS *ERROR* VS. HUMAN *ERROR*" NARRATIVE

One of the alleged advantages of AWS and military AI is their superior performance, including their enhanced accuracy.[33] In particular, Heller underscores several limitations of human ability to make accurate targeting decisions, including cognitive biases such as over-trusting bias, anchoring bias, situational constraints, and negative emotions.[34]

However, relying on such an argument has limitations. First, the claimed superior accuracy of AWS is a dubious proposition, as shown by current debates around accuracy rates, as well as by the reality of AI-enabled targeting in current conflicts.[35] Second, increased reliance on complex AI-enabled targeting systems and on human-machine collaboration as the default approach to military AI requires us to look at errors arising from combinations of both humans and machines, rather than the errors of isolated components. I will turn to these two issues in turn.

A. Errors and AWS Accuracy

[29] "Artificial intelligences are arguably already more capable of performing the cognitive tasks most relevant to warfighting than robots are capable of performing the functions of the human body most relevant to warfighting…. For the foreseeable future, then, in many domains, it will be more plausible to substitute machines for humans where humans have executive roles than where humans have roles involving the manipulation of objects or movement through cluttered environments." Robert J. Sparrow & Adam Henschke, *Minotaurs, Not Centaurs: The Future of Manned-Unmanned Teaming*, 53 PARAMETERS 115, 116 (2023).
[30] Taylor Kate Woodcock, *Human/Machine(-Learning) Interactions, Human Agency and the International Humanitarian Law Proportionality Standard*, 38 GLOB. SOC'Y 100, 112 (2024).
[31] Klaudia Klonowska, *Designing For Reasonableness: The Algorithmic Mediation of Reasonableness in Targeting Decisions*, OPINIO JURIS (Feb. 23, 2024), https://lieber.westpoint.edu/designing-reasonableness-algorithmic-mediation-reasonableness-targeting-decisions/.
[32] *See* Bo & Dorsey, *supra* note 18.
[33] *See* Thurnher*, supra note* 2, at 184–185.
[34] Heller, *supra* note 1, at 31–49.
[35] *See* Evan Dyer, *Israel's Gaza Bombing Campaign Is the Most Destructive of This Century, Analysts Say,* CBC NEWS (Dec. 30, 2023 4:00 AM), https://www.cbc.ca/news/politics/israel-gaza-bombing-hamas-civilian-casualties-1.7068647.

First, there is neither agreement nor clear understanding on error-tolerance and the required level of accuracy of an AWS among states and other actors involved in the development and use of AWS. One way to understand the accuracy of AWS is their performance and reliability in object recognition and classification. In this sense, accuracy is often understood as the ability of an autonomous system equipped with computer vision software to correctly identify what it is supposed to target (positive IDs). More specifically, accuracy refers to the ability to correctly predict whether an object or a person belongs to a class or a list of targets that was pre-decided and programmed. The overall performance of the system also depends on the system's ability to also recognize non-targetable persons or objects (negative IDs).[36] Although states' stances might diverge as a matter of policy or due to operational considerations,[37] ensuring the non-occurrence of false positives is arguably equally important as ensuring the non-occurrence of false negatives. This is reflected in the balance between military necessity and humanity that IHL is meant to achieve.[38] AI must be therefore trained and tested for both. However, current realities of warfare show that this is not always the case.[39] Despite progress in AI, many states recognize that difficulty remains in training a system to correctly recognize civilians no longer directly participating in hostilities, wounded, or surrendering.[40] Moreover, a difficulty with testing is that any "accuracy rate" is developed on the basis of a sample of data, which will not necessarily give an indication of how the same model may function in new circumstances in the future. This is particularly problematic given the complex, unpredictable and dynamic nature of armed conflict in the first place.

In addition to challenges in testing and training AI targeting systems for correctly identifying positive and negative IDs, there is a wide disagreement around error tolerance. What are the unacceptable error rates that would make an AWS inherently indiscriminate? Does a 50% error rate make an AWS per se indiscriminate, or does a 70% error rate do so? This is a complex question to resolve.[41] If we look at studies in the field of medical errors we see that tolerance for AI errors drops when compared with human errors.[42] However, when it comes to military AI—where secrecy and confidentiality play a significant role—an additional question arises: How can we compare AI errors with human errors in targeting if statistics on

[36] *See* Jonathan Kwik and Tom van Engers, *Performance or Explainability? A Law of Armed Conflict Perspective*, *in* 59 L., GOVERNANCE AND TECH. SERIES 255, 271 (2023).
[37] *See, e.g.*, Lora Saalman, *Fear of False Negatives: AI and China's Nuclear Posture*, THE BULLETIN, (Apr. 24, 2018), https://thebulletin.org/2018/04/fear-of-false-negatives-ai-and-chinas-nuclear-posture/.
[38] "IHL represents a carefully thought out balance between the principles of military necessity and humanity. Every one of its rules constitutes a dialectical compromise between these two opposing forces." Michael N. Schmitt, *Military Necessity and Humanity in International Humanitarian Law: Preserving the Delicate Balance*, 50 VA. J. INT'L L. 795, 798 (2011).
[39] Poor training or lack of training data about negative IDs could result in an AI system's biased training and ultimately produce errors. *See* JEWISH INST. FOR NAT. SEC. OF AM., *Gaza Conflict 2021 Assessment: Observations and Lessons* (Oct. 2021), https://jinsa.org/wp-content/uploads/2021/10/Gaza-Assessment.v8-1.pdf.
[40] *See* LAURA BRUUN, TOWARD A TWO-TIERED REGULATION OF AUTONOMOUS WEAPON SYSTEMS? IDENTIFYING PATHWAYS AND POSSIBLE ELEMENTS 11 (2024).
[41] Michel, for example, claims that "[i]f a military assures you that its human-machine systems get it right 99.9 percent of the time, but can't tell you how they apply the law in those 0.01 cases where they don't, then that's not good enough." Arthur Holland Michel, *The Machine Got it Wrong? Uncertainties, Assumptions, and Biases in Military AI*, JUST SEC. (May 13, 2024), https://www.justsecurity.org/95630/biases-in-military-ai/.
[42] "The accuracy and precision of ML/DL systems is typically a composite effect that arises from a combination of the behaviors of different components, such as the training data, the learning program, and even the learning framework. These components are then embedded in larger systems, so interactions with the physical, computational, and human components of the system will ultimately affect system performance." Flournoy et al*., supra* note 22.

the latter are lacking? Without clear benchmarks for human decision-making, is a meaningful comparison even possible?[43]

The "machine vs. human" argument cannot be resolved just by looking at either side's performance because the benchmarks for the human side and the comparing criterion must also be determined. This points to the limited usefulness of comparing human errors with machine errors, and the limited usefulness of using accuracy rates or performance as benchmarks.[44]

Finally, at a more conceptual level, comparing the ability of human combatants to comply with IHL norms with the ability of AWS to do so is problematic. As noted by Woodcock, "this framing of whether military AI can 'comply better' with IHL than humans erroneously equates the output of algorithms with human judgment."[45] Such a framing indeed obscures that compliance with the principle of distinction and proportionality requires more than object recognition and classification. It requires contextual and evaluative judgements in relation to the identification of targetable individuals and targetable objects.

B. Errors Resulting From AI-Human Interactions

Second, the claimed superior accuracy of AWS should prompt consideration of the reciprocal influence that humans exert on AI and vice versa, along with consideration of the errors potentially resulting from these interactions. To understand these errors and their causes, it is essential to consider the issue of algorithmic bias. This issue shows how humans influence AI technologies and their output, in some cases ultimately resulting in misidentification of targets. Algorithmic bias occurs in ML models "when certain types of data are missing or more represented than others."[46] For example, studies show that "facial recognition software recognize male faces far more accurately than female faces and were generally better at recognizing people with lighter skin tones."[47] As highlighted by Bode, these biases can be exacerbated in the training process of AI technologies. In that process, "[h]uman task workers, programmers, and engineers make several choices … such as annotating/labelling/classifying data samples, feature selection, modelling, model evaluation and post-processing after training."[48]

Additionally, "assumptions play an enormously important role in data-driven warfare."[49] For example, current AI-enabled targeting systems show that the identification of targetable individuals relies on assumptions about particular patterns of behavior. Regularly changing one's phone number, combined with other indicators, can classify one as a "combatan[t]."[50]

AI-enabled targeting systems highlight that humans are not out of the picture but that their choices and assumptions are embedded in different phases of the targeting process. Contrary

---

[43] I wish to thank Jonathan Kwik for pointing this out.
[44] *See* Paul Ohm, *Throttling Machine Learning*, *in* LIFE AND THE LAW IN THE ERA OF DATA-DRIVEN AGENCY 214–29. (Mireille Hildebrandt and Kieron O'Hara eds., 2020).
[45] Woodcock, *supra* note 30, at 107.
[46] Bode, *supra* note 27.
[47] Joy Buolamwini and Timnit Gebru examined three kinds of facial recognition software and found that all three recognize male faces far more accurately than female faces and were generally better at recognizing people with lighter skin tones. *See generally* Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 PROC. OF MACH. LEARNING RSCH. 1 (2018).
[48] Bode, *supra* note 27.
[49] Michel, *supra* note 41.
[50] *Id.*

to what Heller argues, human biases and human choices will always remain and shape AI technologies in different phases of the lifecycle of the weapon and of the targeting process.[51]

As noted above, the contrary is also true. AI technologies shape human choices and decision-making. In some cases, they even show signs of replacing human decision-making. The use cases of The Gospel and Lavender show that the "AI is doing the cognitive work"[52] to nominate and prioritize targets. The cognitive role of humans is therefore very limited, as shown by a recent +972 report on the use of Lavender:

> "One source stated that human personnel often served only as a 'rubber stamp' for the machine's decisions, adding that, normally, they would personally devote only about '20 seconds' to each target before authorizing a bombing — just to make sure the Lavender-marked target is male."[53]

The potential for errors in these situations is even greater than with AWS because of the speed, scale and overall climate of risk around this type of AI-enabled targeting. How can errors stemming from these interactions between humans and machines be addressed when they lead to civilian harm? What does it mean for the criminal responsibility framework when complex sociotechnical infrastructures for targeting are in place and war crimes are committed? I will turn to these issues in the following section.

IV. ACCOUNTABILITY FOR WAR CRIMES COMMITTED BY SOCIO-TECHNICAL SYSTEMS

Concerns around the accountability gap point to situations where a war crime has been committed but no one can be held responsible because of the autonomy attained by the weapon system being used.[54] To object to these concerns, one of Heller's key arguments asserts that there already exists an accountability gap for accidents caused by technical failures of weapons (termed "non-criminal accidents").[55] The same goes for soldiers making mistakes of fact about the characteristics of a target – for example, a soldier who "attacks a vehicle genuinely believing it to be an enemy tank."[56] In Heller's view, given the advantages of AWS technology present in target recognition, it is expected to comparatively result in fewer instances of target misidentification compared to humans, thus weakening the objection regarding the accountability gap.[57]

In this section, I will first counter the argument that there is an inevitable accountability gap for mistakes in war, and second problematize the comparison between human error and AI error. Too often, the AWS-created accountability gap argument is being dismissed on the basis of an already existing lack of accountability for mistakes in war and the inevitable existence of non-criminal accidents in war. It is indeed true that there is a scarcity of criminal law cases concerning violations of the rules of the conduct of hostilities[58] and "[t]he law of armed conflict

---

[51] *See* Bode and Nadibaidze, *supra* note 24.
[52] Sparrow and Henschke, *supra* note 29, at 124.
[53] Abraham, *Lavender*, supra note 7.
[54] According to Heller, "[d]uring a protracted conflict, for example, it is likely that at some point a normally reliable cruise missile or precision-guided munition will malfunction and kill civilians. As long as the soldier who operated the weapon did not suspect that it might malfunction, he is not criminally responsible for the civilian deaths." Heller, *supra* note 1, at 66.
[55] *Id*.
[56] *Id*.
[57] *See id*. at 67.
[58] Gaeta, s*upra* note 6, at 22–30.

has a built-in accountability gap."[59] Many reasons underlie this state of affairs, including the legal framework, which is primarily directed at criminalizing *intended* harm, the lack of political will to address instances of unintended harm to civilians, the belief that harm to civilians is inevitable, and the belief that a certain number of mistakes are inevitable. The very nature of the targeting process as a multi-layered decision-making process leads to situations where "a strike can kill civilians while everyone involved can tell themselves they are not responsible."[60] As Oona Hathaway and Azmat Khan describe:

> [T]he targeteer might feel like the commander forced them to provide an unrealistic estimate of expected casualties. The commander can believe that they authorized a strike that was within acceptable thresholds and acted on what they believed was good information. Everyone can persuade themselves that any "mistakes" were caused by flaws in the targeting package or intelligence shortfalls that failed to identify the presence of secondaries. Civilians die and everyone can persuade themselves, and others, that they are not to blame.[61]

The debate around the AWS-created accountability gap should prompt us to question whether this state of affairs is inevitable. Is the term "mistake" too often invoked to conceal systemic mistakes, "minimize the failures involved" in the targeting process, and avoid accountability for civilian harm?[62] This is a pressing question if one thinks that situations such as the ones described above can only be exacerbated with the introduction of technologies including AI, which add other layers of decision-making and invite potential mistakes.

AI often makes it difficult to discern the single cause of a certain accident. Crucially, a system's performance is ultimately the result of "physical, computational, and human components" interacting with each other.[63] Often human errors and AI errors are linked to each other and forced onto each other.[64] As Crootof affirms: "[n]one of the sources of unintended civilian harm in armed conflict—proportionality assessments, equipment errors, user errors, data errors, and communication errors—operate in isolation; rather, they compound each other, sometimes leading to cascading failures."[65]

AI technologies entail interactions between physical, computational, and human components which are unprecedented (one could think of the human data labellers or those training and testing AI). Because AI systems diffuse control across space and time among different actors, how can accountability frameworks address war crimes committed as a result of these systems? How will criminal responsibility be allocated, and to whom? What does this imply for the fundamentals of criminal responsibility for war crimes: *actus reus*, *mens rea*, and causation?

Answering these questions is a predictive exercise since, to date, no prosecutions concerning war crimes committed with AWS or other AI-enabled targeting systems have yet

---

[59] Rebecca Crootof, *War Torts*, 97 N.Y.U. L. REV. 1063, 1063 (2022).
[60] Oona A. Hathaway & Azmat Khan, *'Mistakes' in War*, 173 U. PA. L. REV. 1, 47 (2024).
[61] *Id.*
[62] *Id*. at 6–7.
[63] *See* PAUL SCHARRE, ARMY OF NONE: AUTONOMOUS WEAPONS AND THE FUTURE OF WAR 151 (2018) (emphasis added).
[64] The concept of forced errors "refer to situations where a human does make an error (which may even directly cause the ultimate failure) but where this error itself was the inevitable result of priors in the overall system." *See* JONATHAN H.C. KWIK, LAWFULLY USING AUTONOMOUS WEAPON TECHNOLOGIES 108 (2024); *see generally* Jay Jennings, *Human Factors Analysis and Classification: Applying the Department of Defense System During Combat Operations In Iraq*, 53 PROF. SAFETY 44 (2008).
[65] Crootof, *supra* note 59, at 1078.

taken place. We could, to a certain extent, draw some analogies with cases where responsibility is attributed for accidents involving aircraft and autonomous vehicles.[66] These cases show a significant mismatch between how responsibility was attributed and how control over the system was actually distributed.[67] In some aviation cases, studies show that "while the control over flight increasingly shifted to automated systems, responsibility for the flight remained focused on the figure of the pilot. While automated systems were being relied on more, the nearest human operators were being blamed for the accidents and shortcomings of the purported 'foolproof' technology."[68] The tendency to consider responsibility as lying primarily with end-users of these systems is also re-affirmed in some cases of accidents involving autonomous vehicles (AVs).[69] These approaches insufficiently account for the distributed agency and distributed control within aviation and AVs' human-computer systems.

It remains to be seen what approach international and national courts will take with respect to war crimes committed with AI-enabled targeting systems. Will they misplace responsibility on operators for having incurred known and unjustifiable risks?[70] Or, in the case of AWS, will they identify the commander as the ultimate bearer of responsibility? This approach is considered by many as a panacea for closing the accountability gap.[71]

Alternatively, will the accountability gap be reaffirmed as an inherent feature of the conduct of hostilities? We should not accept this as an unavoidable reality. Rather, accountability frameworks should adapt to and account for the systemic nature of mistakes in war, and the additional layers of distributed agency and control that AI brings. When humans and machines work together and mutually influence each other, traditional conceptions of responsibility – and in particular, the fundamentals of criminal responsibility – may require reconsideration. With regards to *actus reus*, who is launching an attack when AWS or other AI-enabled systems play a role? As Gaeta notes, given the nature and degree of autonomy of AWS, it is unclear whether prohibited attacks caused by a failure of the AI-enabled systems can be considered as the act of the user or commander for the purpose of establishing the *actus reus* of a war crime.[72] Questions also arise regarding the conditions under which programmers, or eventually, Article 36 reviewers, can be held criminally responsible for having committed the *actus reus* of a war crime, or for having causally contributed to it.[73]

---

[66] It is however important to bear in mind the fundamental differences between AI-enabled targeting in war and AI-enabled transportation. *See* Bo, *Criminal Responsibility*, supra note 26, at 26.
[67] *See* Madeleine Clare Elish, *Who Is Responsible When Autonomous Systems Fail?*, CENTRE FOR INTERNATIONAL GOVERNANCE INNOVATION (June 15, 2020), https://www.cigionline.org/articles/who-responsible-when-autonomous-systems-fail/.
[68] *Id*.
[69] *See* Bo, *Criminal Responsibility*, supra note 26, at 23–24.
[70] *See* Jens David Ohlin, *The Combatant's Stance: Autonomous Weapons on the Battlefield*, 92 INT'L L. STUD. 1, 21–29 (2016); Bo, *Autonomous* Weapons, *supra* note 3, at 294-297; Hathaway & Khan, *supra* note 60, at 75–78.
[71] *See generally* Alessandra Spadaro, *A Weapon is No Subordinate: Autonomous Weapon Systems and the Scope of Superior Responsibility*, 21 J. INT'L CRIM. JUST. 1119 (2023).
[72] *See generally* Paola Gaeta, *Who Acts When Autonomous Weapons Strike? The Act Requirement for Individual Criminal Responsibility and State Responsibility*, 21 J. INT'L CRIM. JUST. 1033 (2023).
[73] "If one conceptualizes the obligation under Article 36 as an obligation that is also addressed to individuals involved in the process of studying, developing, acquiring, or adopting a new weapon, its serious violation could give rise to their criminal responsibility for the failure to determine whether the new weapon is illegal." *Id.* at 1037.

With regard to *mens rea*, a shift from penalizing the active, direct perpetration of war crimes to penalizing omission-based forms of commission should be expected.[74] In fact, given that human-machine collaboration is likely to remain the key paradigm for AI integration in targeting, the role of humans is to be confined to supervision or oversight, thus making omissive conduct (lack of supervision or oversight) the main cause of attacks leading to civilian harm.

Mens rea for conduct of hostilities war crimes, which require direct intent,[75] is likely to shift to knowledge-based and risk-based *mens rea*.[76] In other words, the assessment of *mens rea* will boil down to the foreseeability and understandability of the risks of unintended harm to protected individuals or objects when using an AWS.

Finally, causation theories will have to deal with problems related to temporal and spatial gaps between *actus reus* and a war crime. These are often dealt with using theories of intervening and superseding causes. With an AWS, what becomes important are again tests of foreseeability (see the "proximate cause" test) — i.e., what is foreseeable to the developer and user of the AI.[77] However, depending on the AI method used, and its associated 'black box', a causation test may be impossible to undertake.[78]

In conclusion, the debate over accountability gaps in the context of AWS must move beyond comparisons with existing shortcomings in the enforcement of IHL. While it is true that accountability for mistakes in war has long been elusive, the introduction of AI-enabled targeting systems introduces new layers of complexity that challenge the applicability of traditional criminal responsibility frameworks. Rather than accepting the current limitations as a given, the debate around AWS should prompt a reexamination of how responsibility is conceptualized and allocated. We must critically assess whether existing notions of *actus reus*, *mens rea*, and causation are sufficient in a context characterized by distributed agency, complex human-machine interactions, and opaque decision-making processes. The risk is not merely that accountability becomes more difficult, but that it becomes so diffused and obscured that it is effectively erased. As new technologies reshape the conduct of hostilities, legal frameworks must evolve in parallel, ensuring that the integration of AI does not widen, but instead helps close, the accountability gap for civilian harm in armed conflict.

V. CONCLUSION

In this piece, I drew attention to the importance of accurately and fairly locating responsibility for war crimes when agency and control over an attack are distributed throughout an AI system, mediated by AI, or even replaced by AI. To this end, I also discussed how the fundamentals of criminal responsibility for war crimes must be reconsidered, in order to prevent the reinforcement of an inevitable "accountability gap" for war crimes in the conduct of hostilities.

Assigning due responsibility for unintended harm to civilians resulting from the use of AI in war is a matter more pressing than ever given recent reports on the use of AI in Gaza and

[74] *See generally* Marta Bo, *Criminal Responsibility by Omission for Failures to Stop Autonomous Weapon Systems* 21 J. INT'L CRIM. JUST. 1057 (2023).
[75] Decision on the Confirmation of Charges, *Katanga and Ngudjolo Chui*, ICC-01/04-01/07 (Sept. 30. 2008), § 271.
[76] *See* Ohlin, s*upra* note 70, at 21; Bo, *Autonomous Weapons*, *supra* note 3, at 295–297.
[77] *See* Bo, *Criminal Responsibility*, *supra* note 26, at 41.
[78] *See* Yavar Bathaee, *The Artificial Intelligence Black Box and the Failure of Intent and Causation*, 31 HARV. J. L. & TECH. 889, 922–928 (2018).

Ukraine leading in some instances to widespread civilian harm.[79] A prevailing argument in the AWS debate is that keeping a "human-in-the-loop" assures the exercise of human judgment. But, as noted above, keeping a human in the loop does not automatically translate into human judgements being exercised in a meaningful way. AI shapes human decisions, and existing "human-in-the-loop" systems demonstrate that the actual cognitive work in relation to target nomination is done by AI. Reciprocally, human choices, human bias, and human assumptions frame and shape AI outputs. At the stage of development and programming, when selecting and coding target profiles, humans establish the frameworks for target recognition, identification and possibly engagement. Thus, in practice, the dynamics of distributed agency and control across human and computer systems are more complicated than the "humans vs. AI" narrative.

In order to answer the important question of who should be held responsible, these dynamics deserve more understanding. It is not "only" a problem of "many hands," which refers to the difficulty in ascribing moral or legal responsibility when multiple actors are involved in causing a harmful outcome. The accountability problem requires a deeper understanding of how humans iteratively shape AI technology and how AI technology shapes human decision-making. This analysis is crucial in order to prevent the accountability gap for mistakes in war from becoming unavoidable and to prevent misplaced responsibility for war crimes.

[79] *See, e.g.*, HUMAN RIGHTS WATCH, *Gaza: Israeli Military's Digital Tools Risk Civilian Harm New Technologies Raise Grave Laws-of-War, Privacy, Personal Data Concerns* (Sept. 10, 2024), https://www.hrw.org/news/2024/09/10/gaza-israeli-militarys-digital-tools-risk-civilian-harm; Chris Panella, *Artificial intelligence is going to make drone wars much more deadly. It's already started*, BUSINESS INSIDER (Mar. 7, 2025), https://www.yahoo.com/news/artificial-intelligence-going-drone-wars-103702128.html.