AI-Augmented Targeting and Reining in the Law of the Horse

Captain Christopher J. Lin*

ABSTRACT

Coined by Judge Frank H. Easterbrook as a rather tongue-in-cheek concept, the "law of the horse" broadly suggests that existing general legal principles can sufficiently govern new technologies. The law of the horse, however, may not be adequate in looking towards the more nuanced area of artificial intelligence as applied to military targeting operations.

This Article endeavors to explore the concept of justifiable human reliance on machine outputs during the targeting cycle, in light of certain factors that are highly likely to incline a commander to rely on machine outputs in making use of force decisions. Since such reliance at this point in the targeting process may be unavoidable, the focus should be on practices at earlier steps in that process that can ensure that this reliance is justified. In addition, even if it is justified, a commander should not automatically defer to machine outputs in every case but should be encouraged to use her own judgment to determine if there is reason to question or reject machine outputs if circumstances indicate that they are inaccurate. These steps can help ensure that the momentous decision to use lethal force is appropriately informed by, but not completely

* Judge Advocate, United States Army. Presently assigned to the United States Army Pacific. LL.M., 2024, Georgetown University Law Center; J.D., 2017, UCLA School of Law; B.A., 2013, University of California, San Diego. Member of the D.C. Bar. The author extends his gratitude to COL John J. Merriam for inspiring this topic, and acknowledges the valuable contributions of David Barnes (BG, USA, Ret.), Professor Kevin Mullaney (CAPT, USN), and Professor Peter "Pooch" Picucci during the initial discussions. Additionally, thanks are owed to Professors Mitt Regan, James Schoettler (COL, USA, Ret.), and Daniel Wilf-Townsend for their generous suggestions and edits throughout the process. Finally, the author wishes to acknowledge the insightful editors at Penn State Law Review, especially Jackson England, Colin Hitt, Katherine Owens, Olivia Painchaud, and Drew Weglarz. The views and opinions presented herein are those of the author and do not necessarily represent the views of the United States Government, the Department of Defense ("DoD"), or its components. Appearance of, or reference to, any commercial products or services does not constitute DoD endorsement of those products or services. The appearance of external hyperlinks does not constitute DoD endorsement of the linked websites, or the information, products, or services therein.

subordinate to the use of artificial intelligence to augment use of force decisions.

Table of Contents

I. INTRODUCTION

Imagine you are on vacation in Honolulu, Hawaii. It is your first time visiting the islands, and you decide to try a plate of loco moco from a nearby restaurant for lunch. Thankfully, your rental vehicle is equipped with a Global Positioning System (GPS) to assist you in navigating as you make the drive over from your hotel in Waikiki. The GPS tells you to continue straight on Ala Moana Boulevard. You do so. It then tells you to turn right on Piikoi Street. You do so. Finally, it tells you that you are arriving at your destination on the right, just as you pass South King Street. True to its word, you see the restaurant and its neon orange sign when you pull into the parking lot. You prepare to embark on your gastronomic experience.

The example above shows an everyday interaction a human has with a machine. Implicit in this interaction is the idea of justifiable reliance at a very basic level. The average person would likely rely on the GPS's directions unless he was familiar with the route or otherwise knew that the GPS navigated him to an incorrect location. Perhaps innocuous in an everyday context, this reliance is potentially more concerning in the context of targeting systems augmented by artificial intelligence (AI). Commanders and operators must decide whether to use force based on information provided by machines and consider how much they should justifiably rely on that information. The concept of automation bias describes a tendency of a human, in some cases, to have uncritical trust in a machine; that is, to automatically defer to machine outputs without exercising any independent judgment. Are there circumstances, however, when it is reasonable to adopt at least a rebuttable presumption of trust in machine accuracy?

This Article examines several factors that are likely to lead commanders making decisions about whether to use force to adopt such a presumption, rebuttable only if there is clear evidence that calls accuracy into question. If these factors are likely to be influential, should we attempt to counteract them, e.g., by providing detailed technical training to the commander about AI technology, to reduce the likelihood of this occurring? Or should we accept that this will occur—and that such a presumption may be reasonable if the AI has been approved after a review process—and focus our efforts on earlier stages in the targeting process to ensure that such reliance by a commander is appropriate?

This Article argues that commanders are likely inclined to rely on machine outputs due to the standards for legal liability, their training and doctrine, and the demanding cognitive load they face. These influences are likely to be very powerful and counteracting them would be extremely difficult. Furthermore, it may be unreasonable and counterproductive to ask a commander to take the time and effort to independently verify the accuracy of machine outputs at this point in the targeting process. This reasoning suggests that the military should focus on earlier phases of the targeting cycle to ensure that such reliance is, in fact, justifiable.

Part II examines the development of automated weapons in the twentieth century and the parallel evolution of reliance on such weapons. Part III explores three factors that increase the likelihood of deferring to machine designation of a lawful target without critically evaluating that recommendation. Finally, Part IV underscores the significance of steps in the targeting process prior to the ultimate decision to use force, with emphasis on Phase 2 of the United States military's targeting cycle, which involves selecting and prioritizing targets based on intelligence gathered.

II. BIT BY BIT: A GALLOP THROUGH THE TECHNOLOGICAL EVOLUTION OF WARFARE

Automation has been increasingly integrated into warfare over the past century. Its advantages include automated systems serving as force multipliers by reducing the number of warfighters needed for a particular mission and reducing casualties by serving as substitutes for humans.[1] If autonomy is a capability of a weapon system, then AI is a design choice to achieve such a capability.[2] In recent years, state militaries have been exploring possible uses of AI to augment their capabilities in targeting, which may range from identification of lawful targets[3] to the potential for fully autonomous systems that can identify and strike targets without human intervention, albeit within parameters established by humans.[4] Crucial to this is an ongoing concern about justifiable reliance, which can be viewed, at a basic level, through the historical progression of avionics beginning with the Korean War and the corresponding transfer of autonomy from humans to machines.[5]

A. Automation in Warfare

Defining autonomy has been subject to much consideration due to its nuances as a concept; autonomy can be understood as existing not only on a spectrum of requisite human input required to perform a task,[6] but also in terms of the specific function or task that is being automated.[7] At its core, autonomy can be understood as the ability of a machine to perform a function or make decisions in furtherance of a task, with varying degrees of human input. An air conditioning system, for example, may have an overall function of cooling an apartment but may be autonomous with respect to turning on once a certain temperature threshold is reached, as

1. *See* Amitai Etzioni & Oren Etzioni, *Pros and Cons of Autonomous Weapons Systems*, MIL. REV. 72, 72 (2017).

2. *See* David M. Tate, Senior Def. Analyst, Inst. Def. Analyses, Guest Lecturer at Georgetown University Law Center Regarding Test and Evaluation Challenges (Apr. 2, 2024).

3. *See* Geoff Brumfiel, *Israel is Using an AI System to Find Targets in Gaza. Experts Say it's Just the Start*, NPR (Dec. 14, 2023, 4:58 AM), https://perma.cc/J327-J2LT.

4. *See* Paul Scharre, *The Perilous Coming Age of AI Warfare*, FOREIGN AFFS. (Feb. 29, 2024), https://perma.cc/6RB6-BFV3. While selection and engagement of targets may occur without human intervention, this autonomy is exercised as part of a larger system in which humans define objectives and parameters in which this autonomy may be exercised. In other words, "autonomous" does not mean wholly independent of any human control.

5. *See* Steven Fino, *Automation in Air Warfare: Lessons for Artificial Intelligence Today*, *in* THE BRAIN AND THE PROCESSOR: UNPACKING THE CHALLENGES OF HUMAN-MACHINE INTERACTION 27 (Andrea Gilli ed., 2019).

6. *See* Paul Scharre & Michael C. Horowitz, An Introduction to Autonomy in Weapon Systems 5 (Feb. 13, 2015) (unpublished manuscript) (on file with the Center for a New American Security).

7. *See id.* at 7.

determined by human input. Similarly, in a military context, there are a wide range of functions and tasks that can be automated within broader systems or operations to achieve outcomes or to perform functions determined by humans. For example, when conducting an air interdiction operation, a pilot can navigate to the weapons envelope, i.e., the area in which the target is within range of the weapons, and the bomb's computer, through an automated function, can suggest the optimal time for weapons release to assist in one aspect of the operation to accomplish the overall objective.[8]

While institutional definitions have varied,[9] this Article will proceed using the Department of Defense's (DoD) definition of autonomy. The most recent iteration of the DoD Directive 3000.09 ("DoDD 3000.09") governing autonomy in weapon systems focuses on the spectrum of human input, defining a semi-autonomous weapon system as "[a] weapon system that, once activated, is intended to only engage individual targets or specific target groups that have been selected by an operator," and an autonomous weapon system as "[a] weapon system that, once activated, can select and engage targets without further intervention by an operator."[10] Importantly, DoDD 3000.09 states, as a matter of policy, that "[a]utonomous and semi-autonomous weapon systems will be designed to allow commanders and operators to exercise *appropriate* levels of human judgment over the use of force," with what is appropriate based on contextual features such as the operational environment and mission necessities.[11]

Implicit in the definitions and policy of DoDD 3000.09—particularly, the *appropriate* level of human judgment—is a discussion about the relationship between the machine and the humans who use it.[12] At an operational level, the stakeholders are the commanders and operators who execute commanders' intent and decisions by using such autonomous systems, comprising a form of human-machine teaming.[13] Commanders

8. *See* Merel A.C. Ekelhof, The Distributed Conduct of War: Reframing Debates on Autonomous Weapons, Human Control and Legal Compliance in Targeting 146–47 (Dec. 20, 2019) (Ph.D. dissertation, Vrije Universiteit Amsterdam) (on file with the VU Research Portal, Vrije Universiteit Amsterdam).

9. *See* Neil Davison, *A Legal Perspective: Autonomous Weapons Systems Under International Humanitarian Law*, *in* PERSPECTIVES ON LETHAL AUTONOMOUS WEAPON SYSTEMS 5 (2017) (defining, for example, an autonomous weapon system as "[a]ny weapon system with autonomy in its critical functions—that is, a weapon system that can select (search for, detect, identify, track or select) and attack (use force against, neutralize, damage or destroy) targets without human intervention").

10. U.S. DEP'T OF DEF., DIRECTION 3000.09, AUTONOMY IN WEAPON SYSTEMS § G.2 (2023).

11. *See id.* § 1.1 (emphasis added).

12. *See* DEF. SCI. BD., U.S. DEP'T OF DEF., THE ROLE OF AUTONOMY IN DOD SYSTEMS 27 (2012), https://perma.cc/5AUL-SPHK.

13. *See id.* at 23.

must determine the integration and use of autonomous technology in combat operations, with corresponding tradeoffs, such as accuracy versus efficiency.[14] This determination can encompass the degree of human input, which can be conceptualized in three broad categories: human-in-the-loop, human-on-the-loop, and human-out-of-the-loop.[15]

Looking to targeting operations to illustrate the applications of these three categories, a human-in-the-loop scenario is one in which a machine selects potential targets that a human then decides whether to strike.[16] In a human-on-the-loop situation, a machine is capable of both selecting and striking a target, but a human can intervene to take control over the decision of whether to do the latter.[17] Finally, a human-out-of-the-loop scenario occurs when an autonomous weapon system selects and strikes targets without any human intervention at that step.[18]

B. The Beginnings of Automation in Warfare in the Early Twentieth Century

The concept of autonomy began with efforts to employ self-guided bombs in World War II[19] and eventually moved to the integration of increasingly fully autonomous functions in larger weapon systems following the Vietnam War.[20] The Mark 24 torpedo, fondly nicknamed Fido for its capacity to "sniff out" enemy submarines, was one of the first autonomous weapons.[21] It was initially developed by the National Defense Research Committee, tasked with pioneering innovative methods to combat enemy submarines during World War II.[22] Fido was designed with four hydrophones around its casing to detect the sound of an enemy submarine within 1,500 yards underwater, and if no sounds were detected, then Fido would begin a circular search at a predetermined depth for ten to fifteen minutes.[23] By the 1980s, the United States Navy took a step closer to fully autonomous weapon systems with installing the Mark 15

14. *See id.*; *see also* WYATT HOFFMAN & HEEU M. KIM, REDUCING THE RISKS OF ARTIFICIAL INTELLIGENCE FOR MILITARY DECISION ADVANTAGE 21 (2023), https://perma.cc/Q2JV-K93C (noting that "decision makers want to use AI to reduce uncertainty . . . [b]ut the potential unexpected behaviors or failures of AI systems create another source of uncertainty that can lead to misperception and miscalculation").

15. *See* Scharre & Horowitz, *supra* note 6, at 8.

16. *See id.*

17. *See id.*

18. *See id.*

19. *See* ROBERT O. WORK, A SHORT HISTORY OF WEAPON SYSTEMS WITH AUTONOMOUS FUNCTIONALITIES 5 (2021).

20. *See* BONNIE DOCHERTY, LOSING HUMANITY: THE CASE AGAINST KILLER ROBOTS 9 (2012).

21. *See* Thomas Wildenberg, *A Sub-Hunting Bloodhound*, NAVAL HIST. MAG., Oct. 2017.

22. *See id.*

23. *See id.*

Phalanx Close-In Weapon System (Phalanx) on the U.S.S. Coral Sea.[24] The Phalanx "is the only deployed close-in weapon system capable of autonomously performing its own search, detect, evaluation, track, engage and kill assessment functions" for incoming air threats.[25] This progression in autonomous functions in weapon systems shows that, at least on the defensive side, technology has allowed full autonomy for certain weapon systems, i.e., one without a human in the loop.

Concurrent with the integration of autonomous weapon systems into combat operations was the discourse involving the concept of meaningful human control and what it meant to rely on such systems. Reliance on autonomous weapon systems increased from early iterations of automated technologies to more modern versions that had higher levels of efficiency and accuracy.[26] Acknowledging the vast arsenal of autonomous weapon systems, a look towards the evaluation of the relationships between pilots and their avionics helps illustrate the point on the evolving confidence in the reliability of autonomous systems.

As one of the first United States Air Force aircrafts equipped with a radar sight—the A-1CM—to assist with aerial gunnery, the F-86E[27] was described as an aircraft that would enhance ease of engaging targets, in which "the pilot simply keeps the target inside a circular pattern of light or reticule[,] [and] [t]o fire machine guns or rockets he pushes a button when the target is centered."[28] This was meant to remedy the problem that "a pilot [had] very little time to figure the angle between his line of sight and the bore of the guns, the allowance for wind drift, the size and distance of the target."[29] The F-86E pilots during the Korean War, however, provided mixed reviews of the A-1CM.[30] Early usage estimates by pilots placed the A-1CM to help with "only . . . the last 10 percent of the mission, and many thought the new gunsight actually degraded their ability to successfully accomplish the other 90 percent."[31]

24. *See* DOCHERTY, *supra* note 20, at 9; *MK 15 – Phalanx Close-In Weapon System (CIWS)*, U.S. NAVY (Sept. 20, 2021), https://perma.cc/65W2-ZRNT.

25. *Id.* at 10.

26. *See e.g.*, Fino, *supra* note 5, at 27.

27. *North American F-86E Sabre*, PIMA AIR & SPACE MUSEUM, https://perma.cc/TXT5-WE6W (last visited Nov. 4, 2024) ("Initial design work on the F-86 began in May 1945 and resulted in the first prototype which flew in August 1947.").

28. *New Radar Sight Guides Jets' Guns*, N.Y. TIMES, Apr. 3, 1950, at 25, https://perma.cc/TVN3-ZVN6.

29. *Id.* at 25.

30. *See* Fino, *supra* note 5, at 31–32.

31. *Id.* at 31. A secondary concern at play was the fact that, since World War I, pilots attained the title of ace after five aerial kills. If aerial kills were achieved through automation, however, the question remained as to whether the kills were attributable to the pilot or the machine. *See id.*

F-86E pilots had a range of complaints about the A-1CM gunsight, from unreliability to a concern that it was too complex for use; one ace pilot, Francis Gabreski, expressed a preference for using chewing gum on a windshield as a sight, which was representative of the older pilots' preference for using proven technologies or practices.[32] Though the United States Air Force ultimately decided to continue developing advanced systems for fire control, the decision was against a backdrop marked by mistrust in the equipment.[33]

In the latter half of the Vietnam War, the relationship between the pilot and the autonomous weapon systems in the aircraft underwent another shift.[34] The F-15 Eagle, taking flight in the mid-1970s, was transformative in allowing for the pilot to communicate directly with the aircraft's weapon systems via a central computer and its radar could assume the task of distinguishing between actual targets and false signals.[35] The response of the F-15 Eagle pilots to their aircraft and its avionics was distinctly different from those of the F-86E pilots.[36] F-15 Eagle pilots trusted their machines and began "telling their adversaries, '[i]f you come straight down the snot locker today, I will shoot two Sparrows at you and call you dead. If I am out of Sparrows, I will rip your lips off with a Lima before you can get to the merge . . . .'"[37] Advancements in the technology resulted in an undeniable advantage that F-15 Eagle pilots could achieve with respect to targeting military objectives.[38] This advantage allowed them to rapidly progress through the United States military's targeting cycle, which entails the OODA loop that involves four key stages: observing, orienting, deciding, and acting. In essence, pilots could observe what was happening, understand the situation, make a decision based on that understanding, and then act on that decision more quickly and effectively than before.[39] Consequently, pilots relied on machines, and aerial warfare developed into who had the "best head" for information integration, in contrast to dogfighters in

32. *See* KENNETH P. WERRELL, SABRES OVER MIG ALLEY: THE F-86 AND THE BATTLE FOR AIR SUPERIORITY IN KOREA 25 (2005).

33. *See* WORK, *supra* note 19, at 6.

34. *See* Fino, *supra* note 5, at 37.

35. *See id.* at 38.

36. *See* C. R. ANDEREGG, SIERRA HOTEL: FLYING AIR FORCE FIGHTERS IN THE DECADE AFTER VIETNAM 163 (2001).

37. *Id.* In this context, "Sparrow" and "Lima" likely refer to different types of air-to-air missiles, with the Sparrow being the AIM-7 Sparrow and the Lima being the AIM-9L Sidewinder missile. *AIM-7 Sparrow*, U.S. AIR FORCE, https://perma.cc/TP8P-3D5B (last visited Oct. 22, 2024); *AIM-9 Sidewinder*, U.S. AIR FORCE, https://perma.cc/GDL6-ET5D (last visited Oct. 22, 2024).

38. *See* ANDEREGG, *supra* note 36, at 164.

39. *See id.*

previous eras that prioritized manual dexterity.[40] This suggests that predictable reliability may substitute for complete explicability in some cases, if the user understands the limitations of the technology. As discussed in Part III, we can see early on that this is one of the factors that may push toward overreliance on machine outputs.

C. Automation in Warfare in the Targeting Process

While older automated weapon systems had rules-based software in which humans crafted specific parameters for operation that the systems were bound to follow,[41] the introduction of AI, and particularly machine learning as a subset of AI, meant that certain weapon systems could now be provided with a dataset and "generate[] the rules such that it can receive input x and provide correct output y" in accordance with a human-created algorithm.[42] For example, Project Maven, established in 2017, is designed to "automate the processing, exploitation, and dissemination of massive amounts of full-motion video collected by intelligence, surveillance, and reconnaissance (ISR) assets in operational areas around the globe," with "[s]pecially trained algorithms [that] could search for, identify, and categorize objects of interest in massive volumes of data and flag items of interest."[43] Project Maven used machine learning to "autonomously extract[] objects of interest from moving or still imagery"[44] that would then provide warfighters with real-time intelligence on potential targets, which had another important effect of assisting analysts with processing and disseminating the massive amounts of data collected.[45]

Focusing on the modern targeting cycle, integrating AI offers numerous potential applications to some—or even all—of its phases.[46] Broadly, targeting "is the process of selecting and prioritizing targets and matching the appropriate response to them, considering operational requirements and capabilities."[47] The targeting process achieves this end

40. *See id.*

41. *See* GREG ALLEN, UNDERSTANDING AI TECHNOLOGY 3 (2020) ("[R]ules-based software . . . codify subject matter knowledge of human experts into a long series of programmed 'if given x input, then provide y output' rules.").

42. *Id.* at 7.

43. Richard H. Schultz & General Richard D. Clarke, *Big Data at War: Special Operations Forces, Project Maven, and Twenty-First-Century Warfare*, MOD. WAR INST. (Aug. 25, 2020), https://perma.cc/U5BB-5BJ9.

44. Cheryl Pellerin, *Project Maven to Deploy Computer Algorithms to War Zone by Year's End*, U.S. DEPT. OF DEF. (July 21, 2017), https://perma.cc/RUR4-RJB4.

45. *See* Schultz & Clarke, *supra* note 43.

46. *See* Peter "Pooch" Picucci, PhD, Adjunct Professor of Law, Georgetown University Law Center, Targeting: The Applications of AI in Use of Force Decisions 5–10 (Feb. 13, 2024) (on file with author).

47. JOINT CHIEFS OF STAFF, JOINT TARGETING, JP 3-60, at I-1 (2013) [hereinafter JOINT TARGETING]. Note that while the current version of Joint Publication 3-60 is not publicly available, the principles remain constant. Brian L. Cox, *2023 DoD Manual*

goal through six phases: (1) end state and commander's objectives, (2) target development and prioritization, (3) capabilities analysis, (4) commander's decision and force assignment, (5) mission planning and force execution, and (6) assessment.[48]

As a demonstrative of AI applications, Phase 2 can be examined more closely as "the analysis, assessment, and documentation processes to identify and characterize potential targets that, when successfully engaged, support the achievement of the commander's objectives."[49] Here, AI can augment Phase 2 by functions such as generating potential targets that may have otherwise been overlooked, identifying the best attack vector (e.g., striking a weaker side of a building, which would make the attack more viable), or understanding target significance (e.g., identifying a location at which enemy forces congregate with some frequency).[50]

AI nevertheless has its challenges with respect to use.[51] First, with respect to Phase 2, even if the AI in use generates a potential target, "[d]eep learning, as a technique, may be effective in establishing correlation but unable to yield or articulate a causal mechanism."[52] In other words, commanders would run into the issue of being able to explain the decisions of AI as more than merely correlative. In the example above regarding target significance, the location at which enemy forces congregate may simply be a hot dog shop, not of military significance.[53] Second, AI may hallucinate and generate incorrect or misleading results, including false positives or negatives.[54] Third, a persistent concern is AI explicability.[55] AI models that include machine learning algorithms have been construed as black-boxes, due to their complexity, which results in an inability for users to interpret and understand how the machine reached its conclusion or output.[56] For targeting, this poses a significant problem, as commanders

Revision – Practical Concerns Related to the Presumption of Civilian Status – Part II, ARTICLES OF WAR (Aug. 16, 2023), https://perma.cc/X9CW-GSUW.

48. *See* JOINT TARGETING, *supra* note 47, at II-3.

49. *See id.* at II-5.

50. *See* Picucci, *supra* note 46, at 6.

51. *See id.*

52. Ryan Calo, *Artificial Intelligence Policy: A Primer and Roadmap*, 51 U.C. DAVIS L. REV. 399, 414 (2017).

53. *See*, *e.g.*, Steven D. Smith, *Pentagon Center Courtyard Icon, Cold War Legend, to Be Torn Down*, U.S. AIR FORCE (Sept. 20, 2006), https://perma.cc/8CHY-X4NY (noting that "[r]eportedly, by using satellite imagery, the Soviets could see groups of U.S. military officers entering and exiting the hot dog stand at about the same time every day [and] concluded that the stand was the entrance to an underground bunker[,]" which was not true).

54. *See* Zachary Davis, *Artificial Intelligence on the Battlefield: Implications for Deterrence and Surprise*, 8 PRISM 114, 121 (2019).

55. *See* Giulia Vilone & Luca Longo, *Notions of Explainability and Evaluation Approaches for Explainable Artificial Intelligence*, 76 INFO. FUSION 89, 89 (2021).

56. *See id.*

may have difficulty articulating how and why a target was selected. In addition, commanders would need to navigate the issue in which accuracy could be different based on the training data, e.g., a higher error rate is possible where training data underrepresents or misrepresents a certain group.[57] The integration of AI into targeting operations as a means of automation is the most recent development in which we must consider human-machine collaboration and justifiable human reliance on machine outputs.

The next Part of this Article discusses several powerful factors that are likely to incline a commander deciding whether to use force to rely on such outputs in making their decision. Some may argue that such an inclination is inappropriate and that a commander should independently verify the accuracy of machine outputs. I argue below, however, that it would be both very difficult and unreasonable to attempt to take steps to prevent this reliance. A commander contemplating the use of force in many cases will have neither the time nor the expertise to conduct an independent verification of machine outputs. Ensuring that a commander's trust in a machine's contribution is justified is crucial, as it relies on earlier steps in the targeting process where actors engage in a critical assessment of machine outputs at each step. It is at these steps that analysts will have greater time and expertise to engage in such assessments. If commander reliance is predictable and close to unavoidable, the military, therefore, must do everything possible to make sure that this reliance is reasonable.

At the same time, even with such analysis in previous steps of the process, a commander's presumption that it is reasonable to rely on machine outputs should be rebuttable if there is a clear indication of machine error. To return to the analogy of a driver using GPS, a driver should trust the machine unless it is clear, based on her independent knowledge of the surroundings, that it is providing incorrect directions. It, therefore, will be important not only to ensure that a commander's reliance is justified but that she does not exhibit automation bias by completely deferring to the machine without any reliance on her own judgment.

III. DON'T LOOK A GIFT HORSE IN THE MOUTH . . . OR AN ALGORITHM IN ITS LAYERS

Three factors, in the aggregate, push the needle towards reliance on machine outputs: (1) the lack of legal liability if the commander adheres to the machine's recommendation, (2) training and doctrine, and (3) behavioral responses in combat scenarios. While Judge Easterbrook famously noted that general legal principles can apply to novel

57. *See* James Manyika et al., *What Do We Do About the Biases in AI?*, HARV. BUS. REV. (Oct. 25, 2019), https://perma.cc/6466-H6PE.

technologies,[58] this Part illustrates how there are likely gaps in existing law that fail to adequately consider challenges unique to AI-augmented targeting systems. For purposes of the analysis below, this Article assumes that the AI system can achieve parity or near-parity with humans in terms of accuracy. This assumption was chosen because decision-making in warfare has traditionally been under the purview of humans.[59] A longstanding concern has been whether machines have the same understanding of the world and battlefield space to appropriately apply the principles of the Law of Armed Conflict (LOAC) during any engagement.[60] Though it is currently difficult to have a quantitative measure of human error versus machine error, the standard for weapons reviews under the LOAC is to measure the actions of the machine against what a human would have done in a similar circumstance.[61] Thus, if the AI system can achieve at least near-parity with respect to errors in comparison to human decision-making, at least relating to a raw percentage of accuracy, then there is a greater likelihood that these tools will be deployed, even if the AI system may make different categories of mistakes while retaining the same level of accuracy.[62]

A. Legal Liability

Given the potential errors with respect to hallucinations and biases described in Part II above, a risk with the use of AI-augmented targeting systems is the possibility of unanticipated death or injury to civilians and friendly forces, as well as destruction or damage to civilian structures. Imposing legal liability on the commanders who authorize the use of systems that cause such harm, however, can be challenging because of legal requirements for liability within both the military justice system and international criminal law. The next two Sections describe these requirements.

1. Uniform Code of Military Justice

While the United States military rarely uses its military justice system to handle fratricides[63] and other incidents that involve catastrophic

58. *See* Frank H. Easterbrook, *Cyberspace and the Law of the Horse*, 1996 U. CHI. LEGAL F. 207.

59. *See* Damian Copeland et al., *The Utility of Weapons Reviews in Addressing Concerns Raised by Autonomous Weapon Systems*, 28 J. CONFLICT & SEC. L. 285, 295–96 (2022).

60. *See id.*

61. *See id.* at 295.

62. *See* Paul Ohm, *Throttling Machine Learning*, *in* LIFE AND THE LAW IN THE ERA OF DATA-DRIVEN AGENCY 214 (Mireille Hildebranddt & Kieron O'Hara eds., 2020).

63. *See* Lieutenant Colonel Michael J. Davidson, *Friendly Fire and the Limits of the Military Justice System*, 64 NAVAL WAR COLL. REV. 122, 123 (2011).

accidents involving technological failures,[64] a failure involving command authorization[65] of AI-augmented targeting can presumably be charged as involuntary manslaughter or failure to obey an order or regulation under the Uniform Code of Military Justice (UMCJ).

Article 119 of the UMCJ governs manslaughter and states that "[a]ny person . . . who, without an intent to kill or inflict great bodily harm, unlawfully kills a human being by culpable negligence . . . is guilty of involuntary manslaughter."[66] The term culpable negligence is defined as "a degree of carelessness greater than simple negligence," with examples that include "negligently conducting target practice so that the bullets go in the direction of an inhabited house within range" or "pointing a pistol in jest at another and pulling the trigger, believing, but without taking reasonable precautions to ascertain, that it would not be dangerous."[67] Indeed, culpable negligence is negligence that contains a "disregard for the foreseeable consequences to others of that act or omission."[68] Article 119 could apply to situations where the commander authorizes AI-augmented targeting that results in civilian death or fratricide.

However, the same commander is unlikely to be found guilty under Article 119 because of the difficulty in establishing culpable negligence. At the outset, an AI-augmented targeting system would have undergone a weapons review,[69] to ensure adherence with the principles of international humanitarian law, which includes an analysis that the weapon must be able to distinguish valid military targets.[70] Accordingly, there would be an implicit understanding that such a system is within the available arsenal of weapons authorized for use and that its recommendations or decisions are aligned with the principles of international humanitarian law. Rather, it may be more difficult for a commander to disregard the use of an approved

64. *See e.g.*, Colum Lynch, *Anatomy of an Accidental Shootdown*, FOREIGN POL'Y (Jan. 17, 2020), https://perma.cc/DA7F-BPS7 (noting that U.S. military personnel were relieved of liability after downing Iran Air Flight 655 in 1987 due to fog of war); *cf*. Geoff Ziezulewicz, *The Navy Dropped a Homicide Charge Against the Former McCain CO and No One's Sure Why*, NAVY TIMES (May 23, 2018), https://perma.cc/38JE-GK79 (stating that "[t]he Navy has quietly dropped its pursuit of negligent homicide charges against the former commanding officer of a warship that collided with a tanker near Singapore").

65. The failure contemplated in this context is the commander's failure to make his own decision and relies on the AI system in targeting, which produces an unlawful result.

66. Uniform Code of Military Justice, 10 U.S.C. § 919(b)(1).

67. U.S. DEP'T OF DEF., MANUAL FOR COURTS-MARTIAL UNITED STATES IV-81 (2024).

68. *Id.*

69. See Section II(C)(3) below for further background and discussion on weapons reviews.

70. *See* Tobias Vestner & Altea Rossi, *Legal Reviews of War Algorithms*, 97 INT'L L. STUD. 509, 526, 530 (2021) (noting that "[w]ith AI systems operating autonomously, the role typically performed by the weapon (i.e., releasing force) and that performed by the human (i.e., decision-making on the use of force merge into one unique system," thereby necessitating a review that accounts for the spectrum of targeting law).

AI-augmented targeting system, particularly if the system could automate the processing, exploitation, and dissemination of vast amounts of intelligence in support of targeting operations, which arguably demonstrates that the commander elevated his level of care and responsibility by relying on the system's outputs.[71]

Article 92 of the UCMJ governs a failure to obey an order or regulation and states that one who "violates or fails to obey any lawful general order or regulation; having knowledge of any other lawful order issued by a member of the armed forces, which it is his duty to obey, fails to obey the order; or is derelict in the performance of his duties shall be punished . . . ."[72] Unlike Article 119, the first two delineated offenses under Article 92 do not specify the requisite *mens rea*. However, the appeals court in *United States v. Gifford* noted that "the Supreme Court has repeatedly inferred a *mens rea* requirement in instances where it was necessary to 'separate wrongful conduct from "otherwise innocent conduct"'–even when the text of a statute was otherwise silent."[73] In *Gifford*, the court inferred the applicability of *mens rea* to criminal statutes that were otherwise silent on *mens rea*.[74] In reaching its holding, the *Gifford* court noted that "recklessness is the lowest '*mens rea*' which is necessary to separate wrongful conduct from 'otherwise innocent conduct,'" [75] and that looking to "the Model Penal Code and state courts across the country . . . recklessness [is] the lowest possible standard that can be read into a statute that does not set out 'the culpability sufficient to establish a material element of an offense.'"[76] Indeed, under a recklessness standard, an accused must have at least been aware of the risk that he was

71. *E.g.*, Richard H. Shultz & General Richard D. Clarke, *Big Data at War: Special Operations Forces, Project Maven, and Twenty-First-Century Warfare*, MOD. WAR INST. (Aug. 25, 2020), https://perma.cc/7NGT-EXAR. Note that there is a broader discussion about the role of the human in this case. Turning towards the pistol pointing example under Article 119, a more apt analogy applied to the use of AI-augmented targeting systems would be a shooter using a pistol equipped with an advanced targeting system that is designed to automatically adjust the aim and fire based on input from sensors and algorithms. The shooter relies solely on the pistol's targeting capabilities without considering potential errors inherent in the system, and as a result, the pistol misidentifies a harmless target as a threat and fires a live round, causing unintended injury to a nearby civilian. In this scenario, it matters whether the pistol was providing the shooter with traceability in its decision-making to allow for meaningful human input and precautions. *Cf.* Bartlett Russell, Deputy Dir., Def. Sci. Off., Def. Advanced Rsch. Projects Agency, Guest Lecturer at Georgetown University Law Center Regarding Human-Machine Interface 7–8 (Mar. 26, 2024) (on file with author).

72. Uniform Code of Military Justice, 10 U.S.C. § 892.

73. United States v. Gifford, 75 M.J. 140, 143 (C.A.A.F. 2016) (quoting Elonis v. United States, 135 S. Ct. 2001, 2010 (2014)).

74. *See id.* at 146.

75. *Id.* at 147 (quoting Elonis v. United States, 135 S.Ct. 2001, 2013 (2014)).

76. *Id.* at 147–48 (quoting Model Penal Code § 2.02(3) (Am. L. Inst. 1962)).

violating a regulation and ignored such risk.[77] The *mens rea* for failure to obey a lawful order or regulation under Article 92, therefore, is recklessness. Dereliction of duty under Article 92, on the other hand, is governed by the *mens rea* of willfulness or through neglect or culpable inefficiency.[78]

Here, the analysis hinges on the presumption that a charged violation of such orders would be tied to failure to comply with rules of engagement forbidding the targeting of civilians or civilian objects such as cultural property. The reasoning as to why a commander would likely not be guilty under Article 92 is similar to Article 119 above. At its core, the applicable *mens rea* for Article 92 focuses, at a minimum, on negligence, defined as "an act or omission of a person who is under a duty to use due care which exhibits a lack of that degree of care which a reasonably prudent person would have exercised under the same or similar circumstances."[79] Due care may simply be to rely on the AI-augmented targeting system that would guarantee a baseline of accuracy in combat operations.

2. Law of Armed Conflict

The Law of Armed Conflict also provides an avenue for criminal prosecution of commanders for war crimes committed by themselves or subordinates under the Rome Statute of the International Criminal Court ("Rome Statute"), which covers a range of infractions including committing acts against individuals who are *hors de combat* or acts that violate laws and customs governing international and non-international armed conflicts.[80] In this context, the *mens rea* element provides criminal liability standards are similar to the UMCJ.

The Rome Statute provides for a *mens rea* of intent and knowledge for the prosecution of war crimes.[81] Intent is defined as where a "person means to engage in the conduct" or where a "person means to cause that consequence or is aware that it will occur in the ordinary course of events."[82] Knowledge "means awareness that a circumstance exists or a consequence will occur in the ordinary course of events."[83] Conversely, where the *mens rea* is not specified, international courts and tribunals have nevertheless imputed a mental element, much like the *Gifford* court did for

77. *See* United States v. Rapert, 75 M.J. 164, 178 (C.A.A.F. 2016) (Stucky, J., dissenting).

78. *See* Uniform Code of Military Justice, 10 U.S.C. § 892.

79. U.S. DEP'T OF DEF., *supra* note 67, at IV-28.

80. *See* Rome Statute of the International Criminal Court art. 8, July 17, 1998, 2187 U.N.T.S. 90.

81. *See id.* at arts. 8, 30.

82. *Id.* at art. 30.

83. *Id.*

UMCJ statutes silent on *mens rea*.[84] Indeed, like the *Gifford* Court, such courts and tribunals have held that recklessness is required as the minimum required *mens rea*, where the foreseeability of possible death is a relevant consideration for the accused's mental state.[85]

As with the analysis for Article 119 and Article 92 of the UCMJ, it remains difficult to meet the recklessness standard for *mens rea* to hold commanders liable under international humanitarian law. Again, if the AI-augmented targeting system can operate at a level of accuracy that achieves near-parity or parity with human decision-making, then perhaps using such a system rebuts the element of recklessness.

The difficulty of imposing criminal liability on commanders holds true even when looking toward more specific provisions governing targeting. When conducting targeting operations, the LOAC applies as an integral component of international law that governs the conduct of hostilities using lethal force in international and non-international armed conflicts.[86] Under the LOAC, targeting must be evaluated on the basis of military necessity, the distinction between combatants and civilians or civilian objects, the proportionality of the expected harm to civilians and civilian objects incidental to such attacks, and humanity, i.e., avoiding unnecessary suffering on the part of the enemy forces.[87] Each principle must be considered and govern the conduct of military personnel during operations.[88]

The Rendulic Rule under the principle of military necessity bears special attention given its particular relevance to AI-augmented targeting.[89] In the spring of 1944, Lothar Rendulic, then a German army commander, ordered a scorched earth policy in Finnmark, given information that the Soviet Union had troops in pursuit.[90] The destruction to civilian property, including villages and communication lines, was described to be as "complete as an efficient army could do it," with "the extent of the devastation . . . discernable to the eye" even three years after the operation.[91] The Nuremberg Tribunal, however, found Rendulic not guilty of a criminal act—specifically the wanton destruction of private and public property—because "the conditions as they appeared to the

84. *See* Rebecca Crootof, *War Torts: Accountability for Autonomous Weapons*, 164 U. PA. L. REV. 1347, 1376 (2016).

85. *See* ANTONIO CASSESE ET AL., INTERNATIONAL CRIMINAL LAW 76 (3rd ed. 2013); *see* Prosecutor v. Delalić, Case No. IT-96-21-T, Judgment, ¶ 437 (Int'l Crim. Trib. for the Former Yugoslavia Nov. 16, 1998).

86. *See* MAJOR ADAM S. REITZ ET AL., OPERATIONAL LAW HANDBOOK 55 (2024).

87. *See id.* at 55–58.

88. *See id.* at 55.

89. *See id.* at 56.

90. *See* 11 INT'L MIL. TRIBS., TRIALS OF WAR CRIMINALS BEFORE THE NUERNBERG MILITARY TRIBUNALS 1288 (U.S. Gov't Printing Off. 1950).

91. *Id.* at 1296.

defendant at the time were sufficient, upon which he could honestly conclude that urgent military necessity warranted the decision [for the scorched earth policy] made," thus rendering his actions to be one of bad judgement rather than a criminal act.[92] Indeed, Congress, when ratifying a number of LOAC treaties, implicitly recognized the Rendulic Rule, by attaching the following stipulation:

> [A]ny decision by any military commander, military personnel, or any other person responsible for planning, authorizing, or executing military action shall only be judged on the basis of that person's assessment of the information reasonably available to the person at the time the person planned, authorized, or executed the action under review, and shall not be judged on the basis of information that comes to light after the action under review was taken.[93]

As applied to AI-augmented targeting systems, particularly if the targeting system can process and disseminate a large amount of intelligence, surveillance, and reconnaissance ("ISR") information in creating its outputs, the Rendulic Rule would likely shield a commander from criminal liability.[94] The ability of such a targeting system to synthesize information would likely be more efficient and comprehensive in comparison to a traditional targeting cell relying on more manual processes to feed information to commanders. Thus, by relying on such a targeting system, a commander's defense would be that his decisions were based on reasonably available information provided by the system.

3. Analogies to Medical Practice

As the analysis of Article 119 above suggests, the use of AI in targeting may have implications for applying the standard of recklessness or intent on the part of a commander.[95] Within the medical community, a similar topic of discussion has been the integration of AI into medical diagnoses and the corresponding legal incentives for medical practitioners.[96] Specifically, as AI becomes more accurate, its recommendations may become the new standard of care for diagnoses.[97]

---

92. *Id.* at 1297.

93. REITZ ET AL., *supra* note 86, at 56 (quoting S. EXEC. DOC. NO. 110-22, at 13 (2008) (limiting, inter alia, the use of incendiary weapons and blinding laser weapons)).

94. This conclusion assumes that the commander otherwise acted in good faith based on reasonably available information at the time of the decision.

95. *See* A. Michael Froomkin et al., *When AIs Outperform Doctors: Confronting the Challenges of a Tort-Induced Over-reliance on Machine Learning*, 61 ARIZ. L. REV. 33, 62 (2019).

96. *See* Tinglong Dai & Shubhranshu Singh, Artificial Intelligence on Call: The Physician's Decision of Whether to Use AI in Clinical Practice 31 (Nov. 29, 2023) (unpublished manuscript), https://perma.cc/U3DE-6FLE.

97. *See* Froomkin et al., *supra* note 95, at 62.

Relying solely on human diagnoses could expose one to risk because of the failure to use an increasingly common technology, which could produce an inference that the practitioner did not use the appropriate standard of care, i.e., one with AI input.[98] Medical practitioners would thus be incentivized towards relying on AI-generated recommendations given the fear of malpractice, unless there is an articulable reason not to do so.[99] By analogy, a commander that ignores AI recommendations during the targeting cycle may be exposing himself to a similar type of risk under the UCMJ or international humanitarian law.

B. Training and Doctrine

If commanders are analogous to the physicians who are incentivized to use AI, the psychological dynamics of military training and doctrine may further drive the needle forward with respect to potential overreliance on machine outputs.

1. Trusting Your Equipment

At the outset, the idea of "trusting your equipment" is taught early on as a terminal objective in United States Army courses,[100] from new recruits going through basic training[101] to cadets going through obstacle courses,[102] and reiterated at higher ranks by senior officers.[103] Again, if an AI-augmented targeting system is authorized for use, the assumption is that the weapon has undergone the appropriate weapons review in accordance with Army Regulation 27-53,[104] has been vetted at a higher authority level and, thus, should be trusted and relied on in accordance with training. This

98. *See id.*

99. *See id.* at 62–63. The reality is that the incentives governing the use of AI in medical settings is highly nuanced and subject to much ongoing research. A recent study showed that the diagnosing physician's decision for AI use was pulled by a number of factors, including non-clinical objectives, such as the privacy costs incurred by the patient when using AI, which can be difficult to apply directly to a targeting context, but the point remains that a commander will face numerous competing considerations when making a targeting decision, e.g., weighing the military advantage of a successful strike against the number of potential civilian casualties. *See* Dai & Singh, *supra* note 96, at 32.

100. @MCoEFortMoore, X (Oct. 6, 2023, 8:10 AM), https://perma.cc/TZ9M-AVUA.

101. *See* Dave Ress, *Mentorship, Not Yelling: The U.S. Military's Basic Training is Changing*, YORK DISPATCH (June 15, 2022), https://perma.cc/EBG5-BR57.

102. *See* Nia Fields, *Be Confident, Trust Your Equipment*, FUTURE ARMY OFFICERS (June 25, 2017), https://perma.cc/A8SG-K36A.

103. *See* Ruth Steinhardt, *'Trust Your Equipment,'* GW TODAY (Mar. 23, 2015), https://perma.cc/S5TH-HT86.

104. Army Regulation 27-53 governs weapons reviews "ensure they are consistent they are consistent with the international legal obligations of the United States, including law of war treaties and arms control agreements to which the United States is a party, customary international law, and other applicable U.S. domestic law and policy." *See* U.S. DEP'T OF THE ARMY, REGUL. 27-53 ¶ 1 (2019).

becomes relevant in a combat setting where, "[w]hen under stress, fast and effortless heuristics may dominate over slow and demanding deliberation in making decisions under uncertainty."[105] In other words, a commander may be prone to defaulting to learned training during combat, i.e., trusting in the equipment which, in this case, is the AI-augmented targeting system.

2. The OODA Loop

Acknowledging that the adage of "trusting your equipment" is often accompanied by "trust but verify,"[106] the potential for overreliance on machine outputs is compounded by the fact that doctrine may further push commanders to towards reliance. A core tenet of military operations is accelerating the OODA loop: observe, orient, decide, and act.[107] Developed by Colonel John R. Boyd, a former fighter pilot, the OODA loop is, in essence, a disciplined, iterative process of decision-making designed to maintain a competitive advantage over an opponent by continuously assessing, orienting, deciding, and acting upon information in a dynamic environment, thereby forcing the opponent to react—that is, bringing about changes to the situation faster than an opponent can comprehend, effectively "[g]enerat[ing] uncertainty, confusion, disorder, panic, chaos . . . to shatter cohesion, produce paralysis and bring about collapse."[108] More specifically, the observation phase brings in information from the external world, including unfolding circumstances and interactions with the environment.[109] Observation feeds into the orientation phase, which interprets the information gathered based on an individual's existing knowledge, experience, and mental models.[110] The decision phase then entails weighing the available options and their potential outcomes to arrive at a decision on how to respond.[111] Finally, the decision made in the previous phase is implemented through action.[112] This concept has been built into existing United States Army doctrine[113]

105. Rongjun Yu, *Stress Potentiates Decision Biases: A Stress Induced Deliberation-to-intuition (SIDI) Model*, 3 NEUROBIOLOGY STRESS 83, 83 (2016).

106. *E.g.*, Major David J. Devine, *The Trouble with Mission Command: Army Culture and Leader Assumptions*, 101 MIL. REV. 36, 40 (Sept.-Oct. 2021).

107. *See* Lieutenant Colonel Jeffrey N. Rule, A Symbiotic Relationship: The OODA Loop, Intuition, and Strategic Thought 5 (Mar. 2013) (unpublished manuscript) (on file with the United States Army War College).

108. *Id.* at 2, 5; Chet Richards, *Boyd's OODA Loop*, 5 NECESSE 142, 147 (2020); JOHN R. BOYD, PATTERNS OF CONFLICT 132 (2007); *see* Kimberly Wright, *OODA Loop Makes its Mark on Maxwell*, MAXWELL AIR FORCE BASE (Aug. 24, 2010), https://perma.cc/E642-2T5G.

109. *See* Rule, *supra* note 107, at 6.

110. *See id.*

111. *See id.*

112. *See id.*

113. *See, e.g.*, U.S. DEP'T OF THE ARMY, FIELD MANUAL NO. 3-0, OPERATIONS ¶ 1-12 (2022) [hereinafter FM 3-0] ("Army forces must accurately see themselves, see the enemy

and into future command and control concepts that emphasize the need to "exploit the operational initiative and establish overall decision dominance."[114]

AI-augmented targeting systems, as tools that can help expedite decision-making,[115] would, in turn, help commanders adhere to doctrine by capitalizing on the element of speed, thus serving as a further incentive for reliance. In particular, during the orient phase, AI algorithms can analyze complex datasets and rapidly identify patterns, trends, and anomalies to help decision-makers understand a situation more comprehensively and accurately.[116] This enhanced situational awareness would allow for quicker assessments of threats, opportunities, and potential courses of action. Additionally, during the decision phase, the AI system can automate routine decisions or provide recommendations.[117]

C. Behavioral Influences

Two behavioral influences are likely to lead to reliance on machine outputs, particularly in Large-Scale Combat Operations (LSCO), defined as extensive military campaigns involving multiple branches of the armed forces and significant numbers of troops, aimed at achieving strategic objectives over a broad area:[118] (1) cognitive burden and (2) the perceived reliability of the system.

1. Cognitive Burden

In an LSCO-type scenario, commanders involved in targeting must contend with managing a wide range of tasks at scale, including the high-payoff target list, target selection standards, and strategic considerations such as the positioning of artillery for shaping and counterfire operations.[119] A combination of these tasks likely comes with an immense cognitive burden, in terms of the amount of information the commander

or adversary, and understand their operational environment before they can identify or exploit relative advantages.").

114. ARMY FUTURES COMMAND CONCEPT FOR COMMAND AND CONTROL 2028: PURSUING DECISION DOMINANCE iii (2021), https://perma.cc/P9XH-UJ4X.

115. *See* Ali Rogan & Harry Zahn, *How Militaries are Using Artificial Intelligence On and Off the Battlefield*, PBS NEWS (July 9, 2023), https://perma.cc/9JNS-CGMG (noting that "one of the things that AI is doing is helping process information faster").

116. *See, e.g.*, Schultz & Clarke, *supra* note 43.

117. *See, e.g.*, Harry Davies et al., *'The Gospel': How Israel Uses AI to Select Bombing Targets in Gaza*, GUARDIAN (Dec. 1, 2023), https://perma.cc/U7J4-5GLS.

118. *See* FM 3-0, *supra* note 113, ¶ 1-10.

119. *See* Colonel Michael J. Simmering, *Working to Master Large-Scale Combat Operations: Recommendations for Commanders to Consider During Home-Station Training*, MIL. REV., May-June 2020, at 20, 21.

can process.[120] Studies have noted that the scenario in which an individual's attention is divided across multiple, concurrent tasks gives rise to overreliance on machine outputs where some tasks can be automated.[121] In one study, . . . participants simultaneously performed tracking and fuel management tasks manually and had to monitor an automated engine status task. Participants were required to detect occasional automation failures by identifying engine malfunctions not detected by the automation. In the constant reliability condition, automation reliability was invariant over time, whereas in the variable reliability condition, automation reliability varied from low to high every 10 min. Participants detected more than 70% of malfunctions on the engine status task when they performed the task manually while simultaneously carrying out tracking and fuel management. However, when the engine status task was under automation control, detection of malfunctions was markedly reduced in the constant reliability condition.[122]

Additionally, individuals "may be less likely to track automation performance and instead rely on previous judgements of reliability during periods of higher workload, essentially pausing learning by adaptively trading-off information access costs against information utility[,] a known strategy to manage time pressure."[123] Again, holding the assumption that the AI-augmented targeting system can achieve some degree of parity with human-decision making, a marker of reliability, the reflexive tendency will be to rely on the machine.

On a related note, time pressure can further compound cognitive burden. During combat, commanders are often under immense time pressure to make quick decisions to respond to threats[124] and to maintain the OODA loop advantage.[125] If the commander relies on the AI's outputs without too much independent verification, he gains the performance advantage of time.[126] This, however, has a corresponding disadvantage "as

120. *See* Nilli Lavie, *Attention, Distraction, and Cognitive Control Under Load*, 19 CURRENT DIRECTIONS PSYCH. SCI. 143, 145–46 (2010).

121. *See* Luke Strickland et al., *How Do Humans Learn About the Reliability of Automation?*, 9 COGNITIVE RSCH.: PRINCIPLES & IMPLICATIONS 1, 16 (2024).

122. Raja Parasuraman & Victor Riley, *Humans and Automation: Use, Misuse, Disuse, Abuse*, 39 HUM. FACTORS 230, 240–41 (1997).

123. *Id.* at 16 (citations omitted).

124. *See* Kevin Mullaney & Mitt Regan, *One Minute in Haditha: Ethics and Non-Conscious Decision-Making*, 18 J. MIL. ETHICS 75, 76 (2019).

125. *See* Rule, *supra* note 107, at 5.

126. *See* J. Elin Bahner et al., *Misuse of Automated Decision Aids: Complacency, Automation Bias and the Impact of Training Experience*, 66 INT'L J. HUM.-COMPUT. STUD. 688, 697 (2008).

high levels of complacency were shown to result in an elevated risk of commission errors."[127]

2. Perceived Reliability

The DoD and its components devote considerable attention to ensuring that AI is safe and reliable enough to engender user trust.[128] For example, while the United States declined to ratify Additional Protocol I of the 1949 Geneva Conventions, which requires all contracting parties to, "[i]n the study, development, acquisition or adoption of a new weapon, means or method of warfare, . . . determine whether its employment would, in some or all circumstances, be prohibited . . . ,"[129] the United States nevertheless mandated that "[s]ystems . . . go through rigorous hardware and software V&V [verification and validation] and realistic system developmental and operational T&E [testing and evaluation], including analysis of unanticipated emergent behavior."[130]

Moreover, looking specifically to the acquisitions process, testing and evaluation is a rigorous means through which "engineers and decision-makers . . . [can] characterize operational effectiveness, operation suitability, interoperability, survivability (including cybersecurity), and lethality.[131] The terms effectiveness and suitability presumably entail the following factors: (1) whether a system can dependably do what it is intended to do, (2) whether a system can dependably not do undesirable things, and (3) whether a system will be employed correctly when paired with humans.[132] In turn, trustworthiness is established to the extent that those factors are satisfied in the affirmative.[133]

However, to the extent that this is successful, the risk of overreliance on machine outputs is higher. According to a command-and-control experiment conducted to explore the correlation between automation trust and individual task load,[134] the derived evidence "suggest[ed] an

127. *Id.* An additional related consideration is that AI can help counteract and reduce the influence of situational emotional responses that may distort judgment and perception. *See* Etzioni & Etzioni, *supra* note 1, at 74.

128. *See, e.g.*, KELLEY M. SAYLER, CONG. RSCH. SERV., IF11150, DEFENSE PRIMER: U.S. POLICY ON LETHAL AUTONOMOUS WEAPON SYSTEMS 1–2 (2024)

129. Protocols Additional to the Geneva Conventions of 12 August 1949 art. 36, June 8, 1977, 1125 U.N.T.S. 3.

130. U.S. DEP'T OF DEF., DIRECTIVE 3000.09, AUTONOMY IN WEAPON SYSTEMS § 3 (2023).

131. U.S. DEP'T OF DEF., INSTRUCTION 5000.89, TEST AND EVALUATION § 3.1(a) (2020).

132. *See* DAVID M. TATE, TRUST, TRUSTWORTHINESS, AND ASSURANCE OF AI AND AUTONOMY 3 (2021).

133. *See id.*

134. *See* David P. Biros et al., *The Influence of Task Load and Automation Trust on Deception Detection*, 13 GRP. DECISION & NEGOT. 173, 187 (2004).

individual's use of a system's automation capability is directly and positively related to the level of perceived reliability of that system's automation, which leads to trust in machine outputs."[135] A separate study showed that "[a]ll participants [within the study] behaved complacently towards the diagnoses generated by the automated aid at least to some extent," and such complacency did not abate even where a group of participants experienced automation failures.[136] While there is evidence that perceived reliability can shift based on the actual reliability of the system,[137] the argument holds that the perceived reliability of a system by a commander would nevertheless be high if the AI-augmented targeting system is at some level of parity with human decision-making, and if the acquisitions process described above is organized to ensure reliability as much as possible.[138] Ultimately, once there are repeated experiences of reliability, it may give rise to complacency that makes one inattentive to other information that may contradict machine outputs.

IV. PONYING UP TO THE CHALLENGES AHEAD

The unique challenges above highlight the need to move away from the more generalized law of the horse and towards concrete solutions tailored to the impacts of AI on the targeting cycle. Although overreliance is not inherently negative, there should be an emphasis on the initial stages of the targeting cycle to guarantee that the information reaching the commander is accurate. Importantly, the commander's reliance on machine outputs must be a rebuttable presumption to minimize the effects of any potential automation bias.

A. Improving Validation in the Early Phases of the Targeting Cycle

While civilian applications of AI use have been subject to much pushback, which resulted in reconsideration of use in certain cases,[139] military applications may be an irrevocable inevitability on the battlefield, particularly as state actors worldwide see AI as a strategic priority that can

135. *Id.*

136. Bahner et al., *supra* note 126, at 696.

137. *See* Strickland et al., supra note 121, at 17 (noting that "in many circumstances, humans may rely on a mental model of how reliably automation performs with respect to task features or other contexts").

138. On a related note, individuals "could satisfice with respect to learning of automation reliability, either sampling automation reliability less and/or extracting less quality evidence from the task environment in situations where they perceive the automation's reliability to be of low importance to operational success." *Id.* at 16–17 (citations omitted). In other words, commanders may not sufficiently scrutinize AI outputs where he perceives that they are not critical to accomplishing the mission.

139. *See, e.g.*, Ellen P. Goodman, *The Challenge of Equitable Algorithmic Change*, 8 REGUL. REV. DEPTH 1, 1 (2019).

be used in a variety of ways, including target identification and early warning systems.[140] In other words, context matters in determining the appropriate course of action for AI use. For example, the gravity of overreliance on judicial use of algorithms to assess recidivism[141] is different from using the outputs of an AI-augmented targeting system within a combat zone, which necessitates a balance of considerations like efficiency against potential devastation to civilian objects and populations. Given the increasing likelihood of encountering AI technologies on the battlefield, deployed by other state actors, proactive risk management is essential to mitigate potential harms.

The concern above directs attention to earlier stages in the targeting process in which: (1) individuals should be expected to have enough technical knowledge to assess the accuracy and reliability of machine outputs and (2) have more time than a commander may have to take steps to ensure such accuracy and reliability. While the first prong can likely be addressed with additional training on AI technology, this may be difficult to put into practice with all military commanders.

Therefore, the latter prong bears more discussion. As alluded to in Part II of this Article, the targeting cycle is a deliberative process that "provides a coherent range of options and effects that aims to optimize military action by avoiding duplication of effort, effects negating each other[,] and ensures that the right targets are prosecuted in the right order, at the right time[,] by the right capabilities."[142] Again, the cycle is iterative and bidirectional, spanning six phases from the commander's objectives to combat assessment.[143] Importantly, Phases 1 through 3—end state and commander's objectives, target development and prioritization, and capabilities analysis—build towards Phase 4, in which the commander decides on what targets to engage and the means of engaging such targets.[144] Note that this decision is based on the planners' briefing to the commander regarding the recommendations and the rationale behind and target selection.[145]

To reiterate, AI can be integrated into the targeting cycle in a number of ways, including the task of "convert[ing] raw data into actionable

140. *See* Anna Nadibaidze & Nicolo Miotto, *The Impact of AI on Strategic Stability is What States Make of It: Comparing US and Russian Discourses*, 6 J. PEACE & NUCLEAR DISARMAMENT 47, 48 (2023).

141. *See* Jeff Larson et al., *How We Analyzed the COMPAS Recidivism Algorithm*, PROPUBLICA (May 23, 2016), https://perma.cc/S779-U3AJ.

142. N. ATL. TREATY ORG., ALLIED JOINT DOCTRINE FOR JOINT TARGETING, AJP-3.9, para. 1.2.1 (2021).

143. *See* JOINT TARGETING, *supra* note 47, at II-3; Merel A. C. Ekelhof, *Lifting the Fog of Targeting*, 71 NAVAL WAR COLL. REV. 61, 66 (2018).

144. *See* JOINT TARGETING, *supra* note 47, at II-19 to -20.

145. *See id.* at II-19.

intelligence" outputs that can feed into recommendations to the commander in Phase 4.[146] By assuming the task of processing full-motion video captured by unmanned aerial vehicles, for example, AI can label data (e.g., hostile intent or weapon), determine interrelationships between data points, and suggest targets for engagement.[147] Potential challenges arise when considering issues such as how the information is presented to the human commander or operator and what specific information is being shown, out of the vast trove of intelligence synthesized by the AI, as these factors can significantly influence the human's perception of a situation.[148]

To the extent that the targeting cycle is followed at each step leading up to engagement, the cycle as a whole represents the exercise of meaningful human judgment. Planners must understand the commander's intent and end state objectives in Phase 1 and translate them into operational tasks.[149] In Phase 2, planners must develop potential targets through analysis, vetting, validation, nomination, and prioritization.[150] Then, in Phase 3, planners must determine the appropriate asset with which to engage the targets developed in Phase 2.[151] This feeds into Phase 4, where the commander makes the decision about target engagement, which in turn, proceeds to Phases 5 and 6, force execution and combat assessment, respectively.[152] Thus, to remedy the concerns for overreliance above, earlier phases of the targeting process should ensure that the commander's reliance on machine outputs is, in fact, reasonable. In other words, even if the commander at the tip of the spear is not able or likely to question machine output absent unusual circumstances, this is reasonable if there is justifiable reliance in the machine outputs by people involved in the previous phases.

As an example, focusing again on Phase 2 within an LSCO environment, urban areas make detection, tracking, and distinguishing between civilians and threats extremely difficult.[153] As threats blend in with the broader population, target development under Phase 2 using AI can be problematic, due to potential errors in categorizing threats, which would lead to an increase in false-positive or false-negative targets if the operator is uncritical of AI outputs or recommendations. A potential solution is to connect visual data with the system's inferences and outputs

146. Ekelhof, *supra* note 142, at 77.
147. *See id.* at 77–80.
148. *See id.*
149. *See id.* at 66–67.
150. *See id.* at 67.
151. *See id.*
152. *See* JOINT TARGETING, *supra* note 47, at II-4.
153. *Cf.* Russell, *supra* note 71, at 12 (noting the role of environmental complexity in targeting decisions).

so that the operator can correct reasoning errors.[154] The system interface would allow for a replay of the potential threat activity across space and time, with a high level of traceability.[155] Specifically, the focus would be on honing the human-machine interface to ensure operator engagement. The operator should be able to see a machine output of a potentially problematic classification and decide whether to agree with the machine by being able to replay and examine the threat activity that led the machine to arrive at the output.[156] Ultimately, addressing potential overreliance on machine outputs in earlier stages of the targeting cycle can provide a better measure of justified reliance on AI-augmented targeting systems, even when commanders cannot thoroughly interrogate such outputs by the time they receive them.

B. A Rebuttable Presumption

While commander reliance on the system may be reasonable, it should be a rebuttable, rather than a conclusive, presumption. Accordingly, it remains important, such as through measures in the Section above, to ensure that there is no automation bias that could increase the risk of false positives or negatives.

For example, the Recognition-Primed Decision (RPD) model "emphasizes that humans are embodied within situations and environments."[157] Specifically, "we filter the countless cues available in the environment by advancing or suppressing them in our minds based on their association with and relevance to the current goal."[158] In a combat environment, there is some evidence to support the idea that soldiers are especially likely to interpret environmental cues as threats. For example, "combat training increased the likelihood of seeing an individual with one hand behind the back as a threat," where a possible explanation is that "close combat training alter[s] the attentional set of an individual to look for someone who might draw a weapon."[159] Indeed, "[f]or armed conflict, stimulus-driven behaviors would likely create a stronger bias to fire upon any stimulus presented because it could be a threat—ostensibly entering a

154. *Cf. id.* at 8–10 (discussing potential mechanisms to promote human oversight of AI targeting).

155. *Cf. id.*

156. *Cf. id.*

157. Mullaney & Regan, *supra* note 124, at 79.

158. *Id.*

159. Adam T. Biggs et al., *When the Response Does Not Match the Treat: The Relationship Between Threat Assessment and Behavioural Response in Ambiguous Lethal Force Decision-Making*, 74 Q. J. EXPERIMENTAL PSYCH. 801, 821 (2021).

simple see-something/shoot-something mindset once the weapon is drawn."[160]

The Haditha incident can serve to demonstrate how the mechanics of RPD may work.[161] In 2005, a convoy of United States Marine vehicles was returning to base when an improvised explosive device (IED) detonated beneath one vehicle.[162] Sergeant Frank Wuterich pulled his vehicle over to the side of the road and proceeded to engage five men, who were later determined to be civilians.[163] Going through the RPD analysis, a number of cues pushed towards this reaction, which occurred within a mere minute of the explosion,[164] including the fact that insurgents hid among the civilian population, information that an IED attack was likely on the day of the incident and that small arms fire began when the IED detonated.[165] The convergence of these cues led Sergeant Wuterich to believe he was under attack and thus begin to "interpret[] . . . cues through the dominant lens of squad protection."[166]

Though the examples above focus on the perspective of the operator or the individuals pulling the trigger, the same cognitive framework can be applied to the commander who authorizes and approves AI-augmented targeting. While it would also depend on the specific AI output, it is foreseeable that a commander may engage in a form of confirmation bias. He may be likely to perceive threats in his environment, making him more inclined to accept AI outputs that identify ostensible valid targets because they offer means to address such threats. This underscores the importance of ensuring that even a presumption of reliance on machine outputs be rebuttable if a commander is aware of other information that may call the accuracy of those outputs into question.

V. CONCLUSION

The relationship and corresponding reliance that humans place on machines have evolved over the past century, from more rudimentary torpedoes guided by sound[167] to sophisticated avionics on fighter jets that can assist in firing missiles during aerial engagements.[168] With the impending integration of AI into targeting systems, an important question is whether commanders who authorize AI-augmented targeting can have

160. Adam T. Biggs, *Perception During Use of Force and the Likelihood of Firing Upon an Unarmed Person*, 11 SCI. REP. 1, 7 (2021).
161. *See* Mullaney & Regan, *supra* note 124, at 75.
162. *See id.*
163. *See id.* at 75–76.
164. *See id.* at 76.
165. *See id.* at 90–91.
166. *Id.* at 92.
167. *See* Wildenberg, *supra* note 21.
168. *See* Fino, *supra* note 5, at 37.

justified reliance on such systems, especially given the potential flaws currently inherent in AI, such as hallucinated outputs.[169] A number of factors push toward reliance on AI outputs, including legal, doctrinal, and behavioral factors, though such reliance is not fundamentally negative as long as it remains a rebuttable presumption. Accordingly, to ensure justified reliance on machine outputs, the United States military should focus on increasing explicability and traceability in the earlier Phases of the targeting cycle so that the information gleaned from machine outputs is more thoroughly vetted by the time it reaches the commander to make the final use of force decision. Ultimately, Judge Easterbrook's analogy regarding the law of the horse warrants reconsideration in the context of AI integration in military operations. AI represents a novel technology with distinct considerations pertinent to the battlefield, necessitating tailored solutions that address its unique challenges and potential.

This Article focuses on just one aspect of AI use on the battlefield and a plethora of additional considerations remain, as AI becomes increasingly prominent in combat applications. In addition to the aforementioned considerations, another aspect worth pondering is how the anticipated tendencies at the individual decision-maker level might intersect with the institutional incentives of the military. In many instances, automated decision-making currently appears slightly inferior to human decision-making, yet offers significant advantages in terms of cost-effectiveness and speed. If an AI tool proves to be 90% as effective as human decision-making but considerably faster, cheaper, and requires fewer personnel, as was the key assumption for this Article, there could be substantial pressure to adopt it. Coupled with the described reliance on AI, this dynamic may lead to situations where the overall efficacy of military actions diminishes with the introduction of AI.[170] Exploring potential institutional strategies to address this challenge, alongside individual-level considerations, warrants further investigation.

169. *See, e.g.*, Davis, *supra* note 54, at 121.

170. Indeed, the current size of the force, at least with respect to the United States Army has been a subject of discussion. In 2023, Secretary of the Army Christine E. Wormuth noted that "You could not fight a major war in Europe or in Asia very effectively with an Army that's smaller than 450,000," though the future of autonomous technologies can very well reduce the needed footprint. Joe Lacdan, *Army Leaders Stress Transformation as Service Adjusts to Evolving Battlefield*, U.S. ARMY (Sept. 24, 2023), https://perma.cc/KK8G-VX6Y.