It Takes a System: AI-Enabled Weapons and the Law of War

Mitt Regan *

I. INTRODUCTION

Incorporation of capabilities enabled by artificial intelligence ("AI") into weapons have raised concerns about the "loss of human control and judgement in the use of force" because of "the difficulties in anticipating and limiting their effects."[1] As two scholars put it, "[t]he development, testing, and usage of autonomous weapons systems (AWS) is arguably the defining security, legal, and ethical challenge of the twenty-first century."[2]

This Article first describes the primary international initiative to address these concerns, which emphasizes that use of such weapons must be guided by "context-appropriate human control and judgment."[3] I then use two scenarios to probe exactly what this standard means in practice. These scenarios suggest that fully understanding the standard requires that we appreciate that an AI-enabled weapon is a sociotechnical system comprised of humans, machines, and institutional processes that shape their interaction. By system, I mean simply "a number of elements (e.g. technologies, humans, rules) that somehow are interrelated and function together to fulfill a shared goal or objective."[4] With respect to an AI-enabled weapon, the targeting cycle is a sociotechnical system that is embedded in the larger sociotechnical system of the AI weapon lifecycle. This perspective illuminates that ensuring that human moral and legal judgment inform the use of a weapon requires that we focus not simply on whether a human or machine directly identifies and attacks a target, but on whether the system as a whole operates to accomplish this goal.

---

* McDevitt Professor of Jurisprudence; Director, Center on Ethics and the Legal Profession; Coordinator, Program on Law, Ethics, and International Security, Georgetown University Law Center; Senior Fellow, Stockdale Center on Ethical Leadership, United States Naval Academy.

[1] *ICRC Position on Autonomous Weapon Systems*, INT'L COMM. RED CROSS (May 12, 2021), https://www.icrc.org/en/document/icrc-position-autonomous-weapon-systems [https://perma.cc/8S9S-NYBE].

[2] INGVILD BODE & TOM WATTS, MEANING-LESS HUMAN CONTROL: LESSONS FROM AIR DEFENCE SYSTEMS ON MEANINGFUL HUMAN CONTROL FOR THE DEBATE ON AWS 6 (Maya Brehm et al eds., 2021).

[3] United Nations Office for Disarmament Affairs, GGE on LAWS: Rolling Text, Status Date: 8 November 2024 (Nov. 8, 2024), https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2024)/Revised_rolling_text_as_of_8_November_2024_final.pdf [https://perma.cc/HY3C-MS3Y] [hereinafter *Rolling Text*].

[4] Olya Kudina & Ibo van de Poel, *A Sociotechnical System Perspective on AI*, 34 MINDS & MACHS. 1, 2 (2024).

II. THE CONVENTION ON CERTAIN CONVENTIONAL WEAPONS

As concern about the incorporation of artificial intelligence into the use of weapons has increased over the last several years, the main effort to address this concern on the international level has been a United Nations working group focusing on the Convention on Certain Conventional Weapons ("CCW").[5] The purpose of that Convention is to "ban or restrict the use of specific types of weapons that are considered to cause unnecessary or unjustifiable suffering to combatants or to affect civilians indiscriminately."[6] Among the weapons that have been prohibited under the Convention are non-detectable fragments, mines, booby-traps, and incendiary weapons.

In 2013, following a report by U.N. Special Rapporteur Christof Heyns, the CCW's States Parties convened an Informal Meeting of Experts to address emerging technologies related to "lethal autonomous weapon systems" ("LAWS").[7] At the Fifth Review Conference in 2016, the High Contracting Parties established an open-ended Group of Governmental Experts ("GGE") on Emerging Technologies in the Area of LAWS. The task of the GGE was to explore and recommend approaches to these technologies that were consistent with the CCW's objectives.[8]

On December 2, 2024, the U.N. General Assembly adopted a resolution on LAWS that proposed a two-tiered approach: prohibiting certain LAWS while regulating others under international law.[9]

Because AI is a capability that can be incorporated into a wide variety of different objects, states and the GGE have struggled to reach agreement on the basic question of what exactly is a lethal autonomous weapon system. As the Institute of Electrical and Electronics Engineers ("IEEE") has noted, "[a]utonomous and AI capabilities pose different challenges and have different requirements from previous weapons. This does not mean AIS [autonomous and

---

[5] *Convention on Certain Conventional Weapons — Group of Governmental Experts on Lethal Autonomous Weapons Systems*, U.N. OFF. FOR DISARMAMENT AFFS., https://meetings.unoda.org/ccw/convention-on-certain-conventional-weapons-group-of-governmental-experts-on-lethal-autonomous-weapons-systems-2025 [https://perma.cc/6JKV-V5Z7] (last visited Apr. 22, 2026).

[6] *The Convention on Certain Conventional Weapons*, U.N. OFF. FOR DISARMAMENT AFFS., https://disarmament.unoda.org/en/our-work/conventional-arms/convention-certain-conventional-weapons [https://perma.cc/4X98-LM5J] (last visited Apr. 22, 2026). For the text of the Convention itself, *see* U.N. Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects (with Protocols I, II, and III), Oct. 10, 1980, 1342 U.N.T.S. 137.

[7] Mia Gandenberger, *CCW Adopts Mandate to Discuss Killer Robots*, REACHING CRITICAL WILL, https://reachingcriticalwill.org/news/latest-news/8583-ccw-adopts-mandate-to-discuss-killer-robots [https://perma.cc/X5M8-42Z5] (last visited Apr. 23, 2026); *see also 2013 Meeting of States Parties*, REACHING CRITICAL WILL, https://reachingcriticalwill.org/disarmament-fora/ccw/2013 [https://perma.cc/2H4D-UX4U] (last visited Jan. 17, 2026).

[8] Fifth Review Conference of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, *Report of the 2016 Informal Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS),)*, app. ¶ 3, U.N. Doc. CCW/CONF.V/2 (June 10, 2016).

[9] G.A. Res. 78/241 (Dec. 22, 2023); G.A. Res. 79/62, at 2 (Dec. 2, 2024).

intelligent systems] represents a discrete class of weapons; rather, these capabilities can be added to all kinds of weapons."[10]

In November 2024, the GGE prepared a report that said that "with the understanding that nothing has reached consensus until everything has reached consensus,"[11] the Group provisionally found consensus formulations of some concepts. Among these was that "[a] lethal autonomous weapon system can be characterized as an integrated combination of one or more weapons and technological components that enable the system to identify and/or select, and engage, a target without intervention by a human user in the execution of these tasks."[12] To "engage" a target is to strike or fire on it.[13]

The report also said that states must "ensure that human responsibility for decisions on the use of LAWS is retained . . . across the entire life cycle of the weapon system."[14] This means that LAWS must be "operated with context-appropriate human control and judgment," which "includes ensuring moral and legal consideration is given by a human, in particular, during identification and/or engagement stages."[15]

What does it mean for "context-appropriate human control and judgment" to govern the process of identifying and striking a target? As reflected by mandate of the UN GGE, the predominant focus historically has tended to be on when a weapon is used: the selection and engagement of targets. In this Article, I argue that clarifying what it means to exercise human control and judgment in selecting and striking a target requires widening the focus beyond the point at which these two activities occur. This requires appreciating that a military weapon is a sociotechnical system, and that using a weapon is the result of a complex process within this system that involves multiple interactions between humans and machines that are shaped by institutional practices.

III. TWO SCENARIOS

Let's consider two scenarios and assess whether human control and judgment guides the identification and engagement of a target in either one. The first scenario is one that Merel Ekelhof has presented.[16] Consider an air mission to destroy an enemy target. The strike will be conducted by an F-16 fighter jet using two 2000-pound bombs guided by Joint Direct Attack Munitions

---

[10] STEN ALLIK ET AL., IEEE SA INDUS. CONNECTIONS RSCH. GRP. ON ISSUES AUTONOMY AND AI IN DEF. SYSS., A FRAMEWORK FOR HUMAN DECISION-MAKING THROUGH THE LIFECYCLE OF AUTONOMOUS AND INTELLIGENT SYSTEMS IN DEFENSE APPLICATIONS 30 (2024).

[11] *Rolling Text*, *supra* note 3, at 1.

[12] *Id.*

[13] DEPT. OF DEF., JOINT PUB. 1–02, DICTIONARY OF MILITARY AND ASSOCIATED TERMS 81 (2017).

[14] *Rolling Text*, *supra* note 3, at 3.

[15] *Id.* at 2.

[16] Merel Ekelhof, *Autonomous Weapons: Operationalizing Meaningful Human Control*, INT'L COMM. RED CROSS: HUMANITARIAN L. & POL'Y (Aug. 15, 2018), https://blogs.icrc.org/law-and-policy/2018/08/15/autonomous-weapons-operationalizing-meaningful-human-control [https://perma.cc/YPW2-2QL6] (details that follow are drawn from this source).

("JDAMs"), which use the Global Positioning System (GPS) to guide the bombs to the target once they are released. The final stage known as mission execution involves:

(1) finding or locating the target;

(2) fixing the target, which involves positive identification of it;

(3) tracking the target if it is mobile;

(4) confirming that the target may be attacked consistently with law and rules of engagement and any other policies; and

(5) engaging, or firing on, the target.

The operators in the aircraft who will be carrying out the mission receive a pre-flight briefing about two hours before taking off. They are briefed on the mission and receive information that includes:

(1) a description of the target based on all available intelligence;

(2) fix the target's location – that is, its coordinates;

(3) weapon characteristics that will destroy the target and avoid or minimize civilian harm, in this case two 2000-lb. bombs guided by JDAMs and a delayed fuze that does not detonate the bombs until after they enter the facility;

(4) the desired bomb point of impact as guide for aiming them; and

(5) the weather forecast, which in this case describes a night with overcast conditions and some rainfall.

In terms of the tasks involved in executing the mission, the operators don't find the target—others have identified the target as lawful and have loaded its coordinates provided the coordinates to load into the aircraft computer. Operators don't visually fix the location of the target but rely on its coordinates to do that since the target is a structure in a specific place. Furthermore, weather conditions in this case may make visual inspection impossible. Operators needn't track the target, since it is not mobile. Confirmation that striking the target is consistent with law, rules of engagement, and other policies has been done by others earlier in the planning process, so the operators do not do that. Operators will fire—release the bombs—when the GPS in the bombs indicates the best time to do so. They don't aim them, because the GPS will guide the bombs to the target.

The responsibility of the operators therefore is to:

(1) load target coordinates into the aircraft computer;

(2) load the impact angle of the bombs into the aircraft computer;

(3) load the aircraft velocity into the aircraft computer;

(4) navigate to the "weapons envelope," the area within which the weapon can reach its target as identified by its coordinates. This navigation is guided by the aircraft computer;

(5) once in the weapons envelope, wait for the signal from the bombs' computers to release the bombs;

(6) let GPS guide the bombs to the target.

Throughout the mission, including "aiming" and releasing the weapon, the operators follow prescribed steps based on guidance provided by computers that, in varying degrees, rely on capabilities enhanced by AI. We have "humans in the loop" in the final steps involved in firing a weapon. Does human judgment guided by moral and legal considerations inform selection and engagement of the target in this case?

Let's consider a second scenario. A Phalanx Close-In Weapon System ("CIWS") on a naval vessel uses a radar-guided, 20-millimeter Vulcan cannon mounted on a swiveling base to defend a naval vessel against attacks by aircraft, missiles, and small boats. It was developed for situations where a vessel will soon be under attack from so many incoming weapons that it is impossible for humans to select and then fire on each of them.

When a ship gets information about an incoming attack, the captain can order that the self-defense system be changed from human to automatic mode. In this mode, the system uses radar and electro-optical/infrared sensors to detect incoming projectiles to determine their trajectory, speed, and point of origin, and uses acoustic sensors to help locate the source.[17] The Phalanx CIWS uses algorithms to rapidly process this data to identify incoming threats and to prioritize among them. It then fires a wall of projectiles to destroy the threats, all much faster than humans could ever do. In this mode, the system has fully autonomous capability: it can identify and attack targets without any human intervention. It is "the only deployed close-in weapon system capable of autonomously performing its own search, detect, evaluation, track, engage and kill assessment functions."[18]

In this scenario, we have no humans in the loop in the final steps involved in firing a weapon. Does human judgment guided by moral and legal considerations inform selection and engagement of the targets in this case?

I argue that selection and engagement of targets is informed by human moral and legal judgment in both scenarios. To appreciate why, we need to widen the aperture to understand how selecting and attacking a target are the last two stages of a complex system involving a multitude of operations whose combined function is to develop and use a weapon.

---

[17] Alain Servaes, *How U.S. Navy Nimitz-Class Carrier USS Abraham Lincoln Is Armed to Counter Drones and Swarm Boats*, GLOB. DEF. NEWS ARMY RECOGNITION GRP. (Jan. 13, 2026, at 4:08 CT), https://www.armyrecognition.com/focus-analysis-conflicts/navy/naval-technology/how-u-s-navy-nimitz-class-carrier-uss-abraham-lincoln-is-armed-to-counter-drones-and-swarm-boats [https://perma.cc/6MED-4P47]; *Mk-15 Phalanx Close-In Weapon System (CIWS)*, SEAFORCES–ONLINE, https://www.seaforces.org/wpnsys/SURFACE/Mk-15-close-in-weapon-system.htm [https://perma.cc/9QCX-AC49] (last visited Dec. 3, 2025, at 15:38 CST).

[18] *MK 15 – Phalanx Close-In Weapon System (CIWS)*, AMERICA'S NAVY (Sep. 20, 2021), https://www.navy.mil/resources/fact-files/display-factfiles/article/2167831/mk-15-phalanx-close-in-weapon-system-ciws [https://perma.cc/PRD8-NBJH].

IV. SOCIOTECHNICAL SYSTEMS

A core concept in applying a systems perspective to humans' use of technology is a "sociotechnical system." It reflects the idea that any human use of technology is comprised of social and technical subsystems that work together to accomplish the goals of the system. The unit of analysis is neither subsystem alone, but their interaction. "In other words, it is not possible to understand the 'social' without the 'technical,' nor the 'technical' without the 'social.' Explaining the outcomes of any technology requires focusing on the messier 'middle ground' between these two realms."[19]

This serves as a correction to the tendency to focus solely or mainly on technology in attempting to minimize potential harms from the use of AI. Chen and Metcalf note that, in addressing AI risks:

> [M]any regulators and technologists are overly focused on technical solutions: finetuning algorithms for fair outputs, documenting the engineering process, correcting the training dataset, and finding vulnerabilities. These actions are all necessary for safe AI, but such a limited lens implies a narrow scope of responsibility that begins and ends with the technical system.[20]

As these scholars observe, "[a]n AI system can be perfectly functional yet fail at the point of use due to the way people use it or the unintended ways it interacts with institutions and processes."[21] A sociotechnical perspective focuses "not [on] how sophisticated is the technology's design or how advanced are its capabilities, but rather what are the technical and societal dynamics that *together* make the system operate as it does, and produce the outcomes that it does."[22] Peter Kroes and his colleagues maintain that sociotechnical systems have three basic elements: technology, humans, and institutions.[23] Institutions include both formal rules and informal practices that shape how humans interact with each other and with machines, and how machines interact with each other.

Ibo van de Poel characterizes sociotechnical systems as composed of "three basic building blocks: 1. Technical artifacts[;] 2. Human agents[;] [and] 3. Institutions (rules to be followed by the agents)."24 AI systems are distinctive,

---

[19] BRIAN J. CHEN & JACOB METCALF, EXPLAINER: A SOCIOTECHNICAL APPROACH TO AI POLICY 1 (Data & Society ed., 2024), https://datasociety.net/wp-content/uploads/2024/05/DS_Sociotechnical-Approach_to_AI_Policy.pdf [https://perma.cc/VFV2-RDH5].

[20] *Id.*

[21] *Id.* at 7.

[22] *Id.* (emphasis in original).

[23] Peter Kroes et al., *Treating Socio-Technical Systems as Engineering Systems: Some Conceptual Problems*, 23 SYSS. RSCH. & BEHAV. SCI. 803, 803 (2006).

[24] Ibo van de Poel, *Embedding Values in Artificial Intelligence (AI) Systems*, 30 MINDS & MACHS. 385, 391 (2020).

he says, because AI is "autonomous, interactive, and adaptive."25 That is, AI can learn from and adapt itself to its environment.

AI systems therefore contain "an artificial variety of building blocks two and three," which van Poel calls "artificial agents" and "technical norms," respectively. Human agents and rules generally are understood as guided by purposive intention. Artificial agents and technical norms, however, are understood as guided by causal and physical rules rather than intention.26 Non-human agents that lack intention nonetheless can autonomously adapt themselves in response to their environment. Technical rules regulate the interaction between artificial agents and other technologies in the way that social institutions do in traditional sociotechnical systems.

Viewing technologies as sociotechnical systems is not a novel approach, even if new technological developments such as AI can lead to a focus on technology rather than the system of which it is a part. The concept originated with British sociologists who founded the Tavistock Institute for Human Relations at the conclusion of their work for the military during World War II.[27] Members of the Institute engaged in research and consulting. Their work for one of their first clients generated insights that provided the basis for the sociotechnical approach.[28] It is worth describing this project in some detail to illustrate the basic elements of this perspective.

The client was the owner of a large coal mine who was adopting new large-scale industrial mining machinery in its operations. This machinery could extract massive amounts of coal from the coal face, which promised a significant increase in worker efficiency.[29] This promise was not realized, however, and the Tavistock consultants sought to understand the reasons for this.

The consultants found that the new technology disrupted work arrangements under traditional mining operations that were crucial to miners' sense of identity and worth. In those operations, miners were divided into small permanent teams that were responsible for mining coal from small tunnels, and exercised autonomy in determining how to organize their work.[30] As the researchers reported:

> A primary work-organization of this type has the advantage of placing responsibility for the complete coal-getting task squarely on the shoulders of a single, small, face-to-face group which experiences the entire cycle of operations within the compass of its membership. For each participant the task has total

[25] *Id.* at 400.

[26] *Id.* at 391.

[27] *About Our Organisation*, TAVISTOCK INST. HUM. RELS., https://tavinstitute.org/who-we-are/our-story [https://perma.cc/QV6N-V3VJ] (last visited Dec. 3, 2025, at 14:35 CT).

[28] E. L. Trist & K. W. Bamforth, *Some Social and Psychological Consequences of the Longwall Method of Coal-Getting*, 4 HUM. RELS. 3, 4 (1951).

[29] *Id.* at 11–14.

[30] *Id.* at 6.

> significance and dynamic closure. . . . Leadership and "supervision" were internal to the group, which had a quality of *responsible autonomy*. The capacity of these groups for self-regulation was a function of the wholeness of their work task, this connection being represented in their contractual status. A whole has power as an independent detachment, but a part requires external control.[31]

By contrast, mechanized mining organized the work into a standard set of discrete activities that followed one another in rigid succession over three shifts of seven and a half hours each, "so that a total coal-getting cycle may be completed once in each twenty-four hours of the working week."[32] Workers on these shifts were segregated by job categories with each miner devoting himself exclusively to his specific job. In addition, the miners on each shift never met one another.[33]

The work was organized according to a single 24-hour cycle, in which the three shifts were reintegrated into a technological whole by the sequence of tasks.[34] This technological integration did not mean, however, that the different workers over the course of the cycle were reintegrated into a *social* whole. As the researchers noted, "[d]ifferentiation gives rise to the need for social as well as technological integration. No attempt seems to have been made in the [mechanized] method to achieve any living social integration of the primary and shift groups into which the cycle aggregate has been differentiated."[35]

The impairment of productivity connected with this sense of isolation was reinforced by an informal norm of low productivity. This was a response to management's failure to adapt the new technology to the radically different vicissitudes and uncertainties of working underground. Workers faced the demands of the work cycle even as they confronted "the need to contend with interferences, actual or threatened, emanating from the underground situation. The activity of the 'ground' has always to be dealt with, and ability to contend with this second or background task comprises the common fund of underground skill shared alike by all experienced face-workers."[36] Workers responded to the failure of the mechanized system to take this into account by adopting a norm of low productivity.[37]

The Tavistock consultants commented that miners who experienced the shift from the old to the new form of mining expressed "confused mourning for

---

[31] *Id.* at 6 (emphasis in original).

[32] *Id.* at 11.

[33] Trist & Bamforth, *supra* note 28, at 13–14.

[34] *Id.* at 14.

[35] *Id.* at 19.

[36] *Id.* at 23.

[37] *Id.*

the past that still goes on in them together with a dismay over the present coloured by despair and indignation."[38] Thus, as two observers note:

> [E]fficiency of the mine would require understanding the machinery and social structures of the miners as a single unit: a "socio-technical system." As social theorists in a very pragmatic environment, the leverage point they found was not in manipulating the social or the technical alone, but in the middle ground between those domains, at the interaction between humans and machines.[39]

A sociotechnical perspective has been used in the years since to analyze subjects such as human-computer interaction,[40] cybersecurity,[41] information and communication technologies,[42] systems engineering,[43] workplace technology,[44] enginering ethics,[45] and in recent years the use of AI in different settings,[46] including weapons with AI capabilities.[47] A 2023 EU-U.S. Roadmap for Terminology and Taxonomy for Artificial Intelligence includes "socio-technical system" as a term under "technical system attributes," and says, "[t]echnology is not only design and material appearance but also sociotechnical;

---

[38] Trist & Bamforth, *supra* note 28, at 10.

[39] CHEN & METCALF, *supra* note 19, at 3.

[40] *See, e.g.*, BRIAN WHITWORTH, *A Brief Introduction to Sociotechnical Systems*, *in* ENCYCLOPEDIA OF INFORMATION SCIENCE AND TECHNOLOGY 394, 394 (2d ed. 2009).

[41] *See, e.g.*, Masike Malatji, Sune Von Solms & Annlizé Marnewick, *Socio-Technical Systems Cybersecurity Framework,* 27 INFO. & COMPUT. SEC. 233, 233 (2019).

[42] *See, e.g.*, Steve Sawyer & Mohammad Hossein Jarrahi, *Sociotechnical Approaches to the Study of Information Systems, in* COMPUTING HANDBOOK: INFORMATION SYSTEMS AND INFORMATION TECHNOLOGY 5-1, 5-14 (Heikki Topi & Allen Tucker eds., 3d ed. 2014).

[43] *See, e.g.*, Gordon Baxter & Ian Sommerville, *Socio-Technical Systems: From Design Methods to Systems Engineering*, 23 INTERACTING WITH COMPUTS. 4, 4 (2011).

[44] *See, e.g.*, Peter M. Bednar & Christine Welch, *Socio-Technical Perspectives on Smart Working: Creating Meaningful and Sustainable Systems*, 22 INFO. SYSS. FRONTIERS 281, 292 (2020).

[45] *See, e.g.*, Jason Borenstein, Joseph R. Herkert & Keith W. Miller, *Self-Driving Cars and Engineering Ethics: The Need for a System Level Analysis*, 25 SCI. & ENG'G ETHICS 383, 394–96 (2019).

[46] *See, e.g.*, Melissa D. McCradden et al., *A Normative Framework for Artificial Intelligence as a Sociotechnical System in Healthcare*, 4 PATTERNS 1, 1 (2023) (applying a sociotechnical perspective to AI in healthcare); Mohammad Hossein Jarrahi, *Artificial Intelligence and the Future of Work: Human-AI Symbiosis in Organizational Decision Making*, 61 BUS. HORIZONS 577, 577–78 (2018) (applying a sociotechnical perspective to AI in corporate workplaces); Kyle J. Behymer & John M. Flach, *From Autonomous Systems to Sociotechnical Systems: Designing Effective Collaborations*, 2 SHE JI 105, 105 (2016) (applying a sociotechnical perspective to AI in academic and commercial collaborations); Andrew Jones, Alexander Artikis & Jeremy Pitt, *The Design of Intelligent Socio-Technical Systems*, 39 A.I. REV. 5, 5 (2013) (applying a sociotechnical perspective to the use of AI in systems design more broadly).

[47] *See, e.g.*, Ilse Verdiesen, Filippo Santoni de Sio & Virginia Dignum, *Accountability and Control Over Autonomous Weapon Systems: A Framework for Comprehensive Human Oversight*, 31 MINDS & MACHS. 137, 137 (2021).

that is, a complex process constituted by diverse social, political, economic, cultural[,] and technological factors."[48]

Even though the powerful technological features of AI can capture our attention, its uses are embedded in larger sociotechnical systems. It is therefore essential that we understand how AI functions in specific settings, what can promote or hinder its effectiveness, the risks that can arise from how it is used, and what opportunities exist for humans to ensure that outcomes are consistent with human values. This perspective expands analysis to encompass not simply end users of the system but actors such as developers and designers.

V. AI-ENABLED WEAPONS AS SOCIOTECHNICAL SYSTEMS

With respect to the incorporation of AI into weapons, Ilse Verdiesen, Filippo Santoni de Sio, and Virginia Dignum have described how conceptualizing such weapons as "complex socio-technical systems" helps clarify how humans can exercise oversight over such systems.[49] This is prompted by the practical concern that "notions of accountability, moral responsibility and controllability are often used as high-level and overarching constructs in the Autonomous Weapon Systems debate, but their definition and relations are not always fully clarified. Nor is it clear how to operationalize these notions into working and verifiable requirements for practical use."[50]

The authors characterize AI weapon systems as constituted by the interaction among what they call technical, sociotechnical, and governance components. The technical component consists of the technical engineering element, the sociotechnical is the interaction between humans and machines, and governance consists of the mechanisms designed to ensure accountability for outcomes. Each of these should provide a form of control that, in combination with the others, enables effective human oversight over the system as a whole. These controls should function before, during, and after deployment of a weapon.[51]

The technical layer of the system should provide control by being able to "receive the right input from the human operator."[52] The system's feedback mechanism should "verifiably check the difference between output and goals during development . . . in order to keep responding to the reasons (goals and norms) of the human operators."[53] Finally, "after deployment it should be

---

[48] EUR. COMM'N, EU-U.S. TERMINOLOGY AND TAXONOMY FOR ARTIFICIAL INTELLIGENCE 10 (1st ed. 2023), https://digital-strategy.ec.europa.eu/en/library/eu-us-terminology-and-taxonomy-artificial-intelligence [https://perma.cc/C2HM-JYVP].

[49] Verdiesen, Santiano de Sio & Dignum, *supra* note 47, at 139.

[50] *Id.* at 138.

[51] *Id.* at 153.

[52] *Id.* at 151.

[53] *Id.* at 151–52.

technically possible to verify and understand the output and the processes behind them."[54] In the sociotechnical layer:

> [T]he human operators should be able to set the right control measures before deployment and to correctly appreciate the capabilities and limitations of the systems . . . During the use they should have the capacity to have a meaningful interaction with the system and understand what it is doing in order to supervise the system to have ongoing control . . . After deployment they should be able to inspect and assess the behaviour of the system to be able to account for its actions.[55]

Finally, the governance component involves "the political and institutional conditions and the oversight mechanisms required for the system to remain under control."[56] This ensures that there is clear identification of responsibilities and accountability before deployment, and a process to demand accountability and explanation, and impose sanctions if need be.

The different forms of control "are interconnected and depend on each other for information."[57] The authors find that in military operations there is a control mechanism in each layer of the system before, during, and after deployment of a weapon.[58] Effective functioning and interaction among these components then has the potential to provide what they describe as "Comprehensive Human Oversight" over weapons that incorporate AI capabilities.[59]

Verdiesen and her colleagues note that failure to achieve this integration of control mechanisms can result in "accountability gaps" on each level of the system.[60] "[I]f the system is designed to be technically inaccessible," they argue, a technical accountability gap can occur because "human operators cannot give a meaningful account of an action mediated by this machine [since] information on decisions of the machine cannot be retrieved."[61] A socio-technical accountability gap can occur if:

> [H]uman operators do not have sufficient capacity (skill or knowledge) to interpret the behaviour of the machine even though the behaviour is accessible to, for example, an expert . . . Also, motivation to interpret the behaviour of a system could be lacking if there aren't sufficient mechanisms for accountability in place.[62]

---

[54] Verdiesen, Santoni de Sio & Dignum, *supra* note 47, at 152.

[55] *Id.*

[56] *Id.*

[57] *Id.*

[58] *Id.* at 153.

[59] Verdiesen, Santoni de Sio, & Dignum, *supra* note 47, at 159.

[60] *Id.* at 145.

[61] *Id.*

[62] *Id.*

Finally, a governance accountability gap can arise if:

> [A]n institutional setting is lacking to pressure human operators and other personnel (e.g. commanders, engineers) to account for their (mediated) actions even when the human operator may have the capacity to give a meaningful account. [This] also prevents providing protection of the individuals at the lower levels of institutional decisions and omissions.[63]

The work of Verdiesen and her colleagues illustrates how conceptualizing weapons with AI capabilities as sociotechnical systems clarifies whether the development and use of these weapons is informed by human moral and legal judgment and requires analyzing how the system as a whole functions, and not simply particular steps in that process. This framework can enable us to identify "which mechanism in which layer will lead to sufficient human oversight" over the system.[64]

The sociotechnical perspective illuminates why the term "lethal autonomous weapon system" is misleading, because it implies that an entire system operates without any human involvement. This would indeed be problematic because there would be no opportunity for human judgment to determine how a system will operate. A better term is "AI-enabled weapon system."[65] This reflects the fact that performing certain tasks in the functioning of the system may involve reliance on AI. That does not mean, however, that humans are not involved in the system. Rather, it means that AI makes contributions that, in combination with humans, enable the overall system to operate.

With this analytical framework in mind, let's return to the two scenarios that I posed at the outset, the bombing mission and the Phalanx self-defense system in autonomous mode.

VI. THE BOMBING MISSION AS A SOCIOTECHNICAL SYSTEM

Recall that in the bombing mission against an adversary's command and control facility, the operators appear to exercise minimal moral and legal judgment in dropping the missiles on the target. This raises the question: Is the selection and engagement of the target in this case informed by human moral and legal considerations as the CCW GGE says it must be?

To answer this question, we need to widen the aperture in order to appreciate that selecting and engaging a target is just the last stage in a multi-stage targeting cycle that comprises a sociotechnical system. In other words, targeting is "a process, not an action."[66] For the U.S. and NATO, there are five stages that lead to firing a weapon, with a sixth stage afterward to assess the

---

[63] *Id.*

[64] Verdiesen, Santoni de Sio & Dignum, *supra* note 47, at 159.

[65] *See* Jovana Davidovic & Mitt Regan, *Jus Ante Bellum and AI-Enabled Weapons*, *in* JUST WAR THEORY AND ARTIFICIAL INTELLIGENCE 1, 1 (Maria Power & Maggi Savin-Baden eds., 2026).

[66] Merel A.C. Ekelhof, *Lifting the Fog of Targeting: "Autonomous Weapons" and Human Control Through the Lens of Military Targeting*, 71 NAVAL WAR COLL. REV. 61, 64 (2018).

effects of the operation. The process is a cycle in the sense that lessons learned from the last stage should be incorporated into future targeting operations. Each stage requires the performance of many different tasks, which involve humans working in concert with machines that often utilize AI. As U.S. Joint Chiefs of Staff Publication 3-60 states, targeting is an "iterative process."[67] The stages "may occur concurrently, but [the targeting cycle] provides an essential framework to describe the steps that must be satisfied to conduct joint targeting successfully."[68]

A. *The Targeting Cycle*

Let's take a closer look at the five stages of the targeting process that culminate in firing a weapon and how humans and AI-enabled machines can collaborate in each. I will focus on deliberate targeting, when there is more time to deliberate compared to when there is an urgent need to act. In the latter case, known as dynamic targeting, the stages may be compressed.

Stage One involves establishing the commander's intent, objectives, and guidance. This is guided by an articulation of the desired "military end state" that is the goal of a campaign or operation.[69] What political and strategic goals does a commander want to accomplish by striking enemy military targets? "Attainment of clear, measurable, and achievable objectives is essential to the successful realization of the desired end state."[70] Once these objectives are established, the next step is to translate them into specific tasks that must be accomplished. This "ensures that each target can be traced back to clearly defined and attainable goals for military operations and, perhaps even more important, that everyone in the targeting process is aware of the objectives and guidance."[71]

Artificial intelligence can make substantial contributions to the tasks in Stage One. It can rapidly analyze vast amounts of intelligence data from various sources to identify characteristics of broad categories of potential targets and relationships among them. It can help assess the relative value of those assets in the adversary's asset network as a whole, based on models that reflect the relationships among them. What would be the effects on the adversary's network military assets of striking different targets? How would these impacts affect the adversary's ability to conduct effective operations? This can enable generations of different scenarios to identify how targeting can best achieve the desired objectives.

Stage Two is target development, which involves a more detailed analysis of the categories of potential targets that Stage One generates:

---

[67] CURTIS M. SCAPARROTTI, U.S. JOINT CHIEFS OF STAFF, JOINT PUB. 3–60, JOINT TARGETING, at II–3 (2013) [hereinafter JP 3–60].

[68] *Id.*

[69] *Id.* at II–4.

[70] *Id.* at II–4 to II–5.

[71] Ekelhof, *supra* note 66, at 67.

> This includes physical, logical, and complex social systems, and the interaction among them. While a single target may be significant because of its own characteristics, *the target's real importance lies in its relationship to other targets within an operational system*. A target system is most often considered as a collection of assets directed to perform a specific function or series of functions.[72]

The goal is to determine the "function, criticality, and vulnerabilities of each potential target," in order to determine "how much each contributes to the adversary's ability to wage war."[73]

Once an entity has been nominated as a potential target, there is an assessment of the intelligence supporting this designation, and then validation to ensure that it meets the objectives and criteria set forth in Stage One and that targeting would be consistent with IHL and Rules of Engagement.[74] This process is the basis for establishing a restricted target and no-strike list along with priorities among targets, which results in a proposed target list that is given to the Joint Force Commander.

JP 3-60 emphasizes that "[e]stablishing intelligence requirements is critical to the success of target development and to the entire targeting process."[75] AI can be essential in generating the intelligence that is crucial in Stage Two, given the massive amount of information generated from a multitude of sources such as drone full-motion video, geospatial data, imagery data, various forms of signals data, human intelligence, open source information and others. It may, for instance, detect patterns and anomalies that human analysts might miss, cross-reference potential targets against databases of protected sites, conduct sentiment analysis of intelligence and diplomatic reports to assess likely reaction to destroying particular targets, and use detailed models of relationships among entities to generate scenarios based on different combinations of strikes on different targets.

JP 3-60 notes that in Stage Two target system analysis "[a]utomation is often used to record the structure of target systems and model various functional impacts on them. Automated models are also used to study the cascading effects and coupling of target systems to show how they could affect one another."[76]

Stage Three of the targeting cycle is capabilities analysis: identifying what assets are available to conduct targeting of which entities. The goal is to efficiently achieve the desired effects while minimizing unintended harm.[77] "Weaponeers" conduct this analysis drawing on "probabilistic mathematical models . . . that take into account the target's critical vulnerabilities,

---

[72] JP 3–60, *supra* note 67, at II–5 (emphasis in original).

[73] *Id.* at II–6 to II–7.

[74] *Id.* at II–7.

[75] *Id.* at II–6.

[76] *Id.* at B–2.

[77] JP 3–60, *supra* note 67, at II–13.

performance data on the assets considered for application against the target, and means of delivery."[78]

In addition, "CDE [Collateral Damage Estimation] is a critical part of the effects estimate step in the joint targeting process phase 3. . . when munitions are used on facilities."[79] These estimates are based on models that incorporate a range of factors such as weapon blast and fragmentation radius, angle of the strike, time of fuze detonation, characteristics of the target and the surrounding terrain, environmental conditions, civilian presence and patterns of life, and other variables.[80] Stage Three culminates in the presentation of a list of targets on a Joint Integrated Prioritized Target ("JIPTL") list to the commander for authorization to begin planning specific strikes, based on an assessment of the probability of success and risk associated with striking each potential target with each type of weapon.

As should be apparent, AI can be used in this stage for tasks such as modeling different weapons systems and their effects when used on specific targets. This can generate a large number of scenarios for consideration and may enable automated pairing of assets to targets based on analysis of each. JP 3-60 says that during capabilities analysis "weaponeering information for a particular target is entered into an automated weaponeering system. Automated weaponeering programs utilize approved weapons data, delivery parameters, and accuracies to provide optimal weapon and platform (or capability) pairings to minimize forces required to meet the commander's objectives."[81]

Stage Four is known as Commander's Decision, Force Planning, and Assignment. This involves final approval of targets and assignment of them to different units to plan and execute missions, based on a determination of the best available weapons to use. "The process of resourcing JIPTL targets with available forces or systems and ISR [intelligence, reconnaissance, and surveillance] assets lies at the heart of force assignment. This process links theoretical planning to actual operations."[82] Tasking orders then provide "information necessary for detailed tactical-level planning of operations."[83]

The activities in Stage Four can draw on AI to analyze historical data and current intelligence to assess risks to friendly forces and civilians; identify adjustments that may lower civilian harm below the collateral damage estimate; generate automated strike packages to achieve objectives while minimizing collateral harm; and optimize the effects of a mission based on the available forces and capabilities.

---

[78] *Id.* at II–14.

[79] *Id.* at II–16.

[80] *See generally* STANLEY A. MCCHRYSTAL, J–2, CJCSI 3160.01, INSTRUCTION: NO-STRIKE AND THE COLLATERAL DAMAGE ESTIMATION METHODOLOGY (2009), https://www.aclu.org/sites/default/files/field_document/drone_dod_3160_01.pdf [https://perma.cc/8YV9-63QM].

[81] JP 3–60, *supra* note 67, at B–3.

[82] *Id.* at II–16.

[83] *Id.*

Stage Five is Mission Planning and Execution. This is the stage of the bombing mission that I discussed earlier that involves finding the target, fixing its location, tracking its movement if necessary, confirming it is a legitimate target, and firing on the target.

There are a variety of tasks in each step in this stage which AI can be useful. These include target selection and verification, use of route-planning algorithms to determine the safest and most effective routes for aircraft or other targeting platforms, assistance in deconfliction of airspace operations, helping optimize timing and sequencing in complex operations with multiple assets, and terminal guidance to home in on selected targets.

B. *Human Judgment in the Bombing Mission*

Despite the minimal judgment that operators exercise in the last stage of the bombing mission, it should be clear that this last stage is the culmination of previous stages comprised of multiple occasions on which humans must take moral and legal considerations into account in performing the tasks necessary for the mission to occur. In other words, we need to appreciate that releasing the missiles is but one part of a much larger sociotechnical system comprised by the targeting cycle and ask whether that system as a whole sufficiently incorporates moral and legal concerns.

The modern targeting sociotechnical system utilizes AI for many of its functions, and reliance on it is likely to increase.[84] There is a tendency, however, to neglect analysis of these decision support functions and to focus simply on whether a machine uses AI to select and engage a target.[85] Merel Ekelhof cautions, however, that "even if these technologies are playing a supporting role, and even if a human being ultimately makes the decisions, the technologies can influence critical targeting decisions—which could be both positive and negative."[86] On the one hand, they could accelerate the targeting process and help ensure that it is based on accurate information. On the other hand, we need to:

> [c]onsider the effect of autonomous technologies that decide, out of large amounts of data, what specific data to show to their operators and what data to ignore, thereby influencing or shaping situational awareness. Another consideration relevant to assessing human control in the targeting process would be the effect of data labeling on target selection. What labels are being applied (e.g., *weapon, attack, combatant, hostile intent*)? And

---

[84] *See* Ekelhof, *supra* note 66, at 74–81 (describing various potential use for AI in the targeting cycle); *see also* JP 3–60, *supra* note 67, at B–1 to B–4 (containing an Appendix devoted to Targeting Automation).

[85] *See* Ekelhof, *supra* note 66, at 80 ("These types of technologies are often neglected in the discourse on autonomous weapons owing to the fact that they are not weaponized."); *see also* Klaudia Klonowska, *Article 36: Review of AI Decision-Support Systems and Other Emerging Technologies of Warfare*, *in* 23 YEARBOOK OF INTERNATIONAL HUMANITARIAN LAW 123, 127 (Terry D. Gill et al. eds., 2020) (describing the crucial role that AI decision–support systems can play in the targeting process).

[86] Ekelhof, *supra* note 66, at 82–83.

> how is this information presented to the human; is there a risk of either automation bias toward or mistrust of the system?[87]

This requires examining the human-machine interactions that constitute the system to determine if humans are exercising distinctive human judgment in these interactions or are simply deferring to machine outputs. If the latter occurs at any given point, having an operator involved in the last stage will not be sufficient to ensure that use of the weapon is informed by moral and legal considerations. By contrast, if the process that leads up to the strike involves repeated exercises of human judgment in assessing and relying on machine contributions, we can say that the system in which the bombing mission is embedded incorporates these concerns. As Ekelhof says, "[I]f we fail to consider these types of technologies for intelligence tasks and the manner in which they are implemented within the military architecture, we risk losing a valuable opportunity to examine potential ways to manage them."[88]

The International Committee of the Red Cross ("ICRC") has recognized the importance of these technologies for IHL compliance, and has sponsored commentary and commissioned reports that address issues raised by use of AI-enabled decision support systems ("DSS") in military operations.[89] These have focused on "how the human judgement required by IHL may be affected or impeded by the military use of increasingly complex machine learning-based DSS."[90] This is a commendable example of widening the focus in order to conceptualize the targeting process as a sociotechnical system comprised of this interaction between humans and this technology.

VII. THE PHALANX AS A SOCIOTECHNICAL SYSTEM

The second scenario involves the use of a Phalanx CWIS in autonomous mode. Determining whether this use of the weapon is guided by moral and legal considerations requires that we appreciate that, while the targeting cycle is a

---

[87] *Id.* at 80 (emphasis in original).

[88] *Id.*

[89] *See, e.g.*, Jimena Sofía Viveros Álvarez, *The Risks and Inefficacies of AI Systems in Military Targeting Support*, INT'L COMM. RED CROSS: HUMANITARIAN L. & POL'Y (Sep. 4, 2024), https://blogs.icrc.org/law-and-policy/2024/09/04/the-risks-and-inefficacies-of-ai-systems-in-military-targeting-support [https://perma.cc/57QT-V2FK]; Wen Zhou & Anna Rosalie Greipl, *Artificial Intelligence in Military Decision–Making: Supporting Humans, Not Replacing Them*, INT'L COMM. RED CROSS: HUMANITARIAN L. & POL'Y (Aug. 29, 2024), https://blogs.icrc.org/law-and-policy/2024/08/29/artificial-intelligence-in-military-decision-making-supporting-humans-not-replacing-them [https://perma.cc/JZV3-EHPZ]; ARTHUR HOLLAND MICHEL, *ICRC Observations on External Report*, *in* DECISIONS, DECISIONS, DECISIONS: COMPUTATION AND ARTIFICIAL INTELLIGENCE IN MILITARY DECISION-MAKING 3, 3–8 (2024), https://library.icrc.org/library/docs/DOC/icrc-4757-002.pdf [https://perma.cc/5DLC-NLD5]; ANNA GREIPL, NEIL DAVISON & GEORGIA HINDS, INT'L COMM. RED CROSS, GENEVA ACAD., EXPERT CONSULTATION REPORT: ARTIFICIAL INTELLIGENCE AND RELATED TECHNOLOGIES IN MILITARY DECISION-MAKING ON THE USE OF FORCE IN ARMED CONFLICTS: CURRENT DEVELOPMENTS AND POTENTIAL IMPLICATIONS 4 (2024), https://geneva-academy.ch/wp-content/uploads/2025/09/Artificial-Intelligence-And-Related-Technologies-In-Military-Decision-Making.pdf [https://perma.cc/88Q9-NKLV].

[90] HOLLAND MICHEL, *supra* note 89, at 4.

sociotechnical system, it is nested within a larger sociotechnical system consisting of an AI weapon life cycle.

A good description of this lifecycle is contained in the IEEE *Framework for Human Decision-Making through the Lifecycle of Autonomous and Intelligent Systems in Defense Applications*.[91] The Framework underscores that ensuring that human judgment guides the use of such systems requires directing beyond the point at which a weapon is deployed. "Much of the focus in international debates on AIS [Autonomous and Intelligent Systems] in defense," the Framework says, and "centers on the potential for human control to be exercised at the point of deploying the technology. However, many people are involved across the entire lifecycle of such systems, including before their development has even started."[92] The Chair of the GGE has recently underscored this point by noting the possibility that actions taken "prior-to-the-loop" that bound the parameters of a LAWS can be a way of securing human judgment and control over the use of a LAWS.[93] The IEEE Framework emphasizes that "there is no such thing as a completely autonomous system. Humans are always involved somewhere along the way. The Framework, therefore, aims to inform discussions about manifold technical, ethical, and legal issues arising out of pursuing a human-centric approach across the entire lifecycle of AIS, especially with regard to human responsibility and accountability."[94]

The Framework identifies nine stages in the lifecycle of AIS in the defense setting "in which human decision-makers play especially key roles, and between which there is likely to be a transfer of responsibility and accountability to new human decision-makers."[95] It also identifies five ongoing activities that require human decision-making throughout the lifecycle.[96] Conceptualizing these stages and the human decision-making within and across them as a sociotechnical system underscores that:

> Everyone associated with deploying an AIS for military purposes needs to have a common understanding of its capabilities and limitations. Individual, incremental steps in design and development may seem unproblematic, but then trigger cascading effects when combined with other technological components. A system cannot be understood in isolation from other technical components, the logistical environment of use,

---

[91] ALLIK ET AL., *supra* note 10; *see also* GINA M. RAIMONDO, NAT'L INST. STANDARDS & TECH., NIST AI 100-1, ARITIFICAL INTELLIGENCE RISK MANAGEMENT FRAMEWORK 10–11 (2023), https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf [https://perma.cc/RST2-F2BT] (providing a more general description of the AI lifecycle with slightly different stages).

[92] ALLIK ET AL., *supra* note 10, at 29.

[93] Grp. of Gov't Experts on Emerging Techs., Background Paper on Context-Appropriate Human Judgement and Control and the Critical Functions of LAWS, ¶ 22, U.N. Doc. CCW/GGE.1/2025/CRP.1 (May 20, 2025).

[94] ALLIK ET AL., *supra* note 10, at 6.

[95] *Id.* at 7.

[96] *Id.*

> the human environment, or the personnel it will interact with for a given task.[97]

I describe in some detail below the activities in these lifecycle stages, as well as the ongoing activities across the stages, to underscore the multiple points at which human judgment needs to inform the interdependent tasks that must be accomplished to enable an AI-enabled weapon to function in compliance with IHL.

A. *The AI Lifecycle*

1. Stage One: Deliberation Prior to Development

The first stage in the lifecycle is deliberation before development of an AIS. This involves consideration of the rationale for possible development and use of an AIS, assessment of the requirements it would need to satisfy, and anticipation of legal, ethical, and technical issues that will need to be resolved if development proceeds. Actors need to consider questions such as: "Is it clear what problem is being solved by using autonomous and/or AI capabilities? Are these capabilities the best solution for that problem? Is it clear what the desired function is for the system? Is it clear that autonomy or AI are the best solutions to perform that function?"[98]

Actors at this stage need to define clear requirements for intended use of a potential military AIS that can guide analysis of trade-offs involved in using an AIS compared to a traditional weapon, what a clear set of requirements for intended use would look like for an AIS used for a particular purpose, how to ensure human oversight throughout the lifecycle, and documentation of development and design decisions if development were to proceed.

Furthermore, "[b]efore any potential system development, it must be identified who should be involved throughout the lifecycle. This includes figuring out how to consider and address ethical, moral, and legal considerations throughout the lifecycle."[99] This requires determining appropriate levels of oversight at various stages and how to integrate legal and policy reviews that occur in them.

The Framework notes that, while IHL applies to AIS, it is not always clear exactly how. Thus, there needs to be agreement on the answers to questions such as what it means to "automate" the targeting process and what is necessary to regulate it; whether to allow and regulate the use of systems that cannot always communicate with humans, or whether to require continuous unbroken communication; and how to ensure sufficient time to verify that systems function as intended when decisions need to be made quickly.[100]

---

[97] *Id.* at 8.

[98] *Id.* at 30.

[99] ALLIK ET AL., *supra* note 10, at 31.

[100] *See id.* at 29.

2. Stage Two: Research and Development

The Framework notes that "[p]roblems with respect to human control or oversight at the deployment stages may arise because they were not addressed at earlier stages, including research and development . . . It must . . . be considered what responsibility means at the research and development stage, and how the role of the developer needs to differ for systems with autonomy and/or AI compared to traditional weapons."[101] Thus, as the Framework puts it:

> Any potential system development should start with a set of requirements which, when met, enable the system to perform its intended functions. Once a system design is proposed, specifications are developed that describe what components must be used and subsystems must do in order for the system to satisfy the requirements. Components and subsystems can then be built or bought to meet the specifications, and assembled into a system which can be expected to satisfy the system requirements.[102]

To this end, people involved in developing an AIS need to understand what it is supposed to accomplish and the context in which it will be used, and to appreciate that system development and design includes assessment and documentation of a system's features that will affect its operation.[103]

One challenge is that individual AI components that perform various functions in an AIS often will have been developed for civilian purposes, without attention to the types of risks that can arise in military use. There needs to be assurance at the development stage that they will be tested for their suitability for military purposes, and that the components will be integrated so that the system operates effectively and safely. Thus, "[t]he implications of the dynamic and integrated nature of potential AIS for use and testing need to be considered at the research and development stages."[104]

The research and development stage should also identify who is responsible for ensuring that the AIS meets legal and policy requirements. Developers themselves may not be familiar with these requirements but should have at least a basic understanding and should know whom to consult for more details. The Framework observes that "addressing the gap between the human language in which laws are expressed, and the specification languages used to characterize the behaviors of technologies, could be an opportunity for collaboration between engineers, legal experts, and policymakers."[105] This requires that "mechanisms and processes must be defined to assess acceptable risk quickly given the circumstances of operational use."[106]

---

[101] *Id.* at 32.

[102] *Id.* at 44.

[103] *Id.* at 37.

[104] ALLIK ET AL., *supra* note 10.

[105] *Id.* at 38.

[106] *Id.* at 42.

Based on this understanding, developers should establish a process to identify what problems that arise later in the lifecycle should stop development or use of the problem; establish processes and procedures to enable a user to take over the system and the conditions under which this should occur; and document circumstances "under which humans would not be able to intervene or override system behaviors."[107]

A particularly important task at the development stage of AIS is to "define failsafe behaviors that take effect when systems correctly identify certain states—for example, loss of communication or moving out of a certain region."[108] This requires that developers and designers identify undesirable states, these states are detected, and "behaviors created during the design phase are then executed and cause the system to exit that state.".[109] Furthermore, "[t]here may need to be a process by which human operators can designate states as undesirable and create corrective behaviors after fielding, as the use of the system changes."[110]

A focus on sociotechnical systems "undermines the notion that developers are only responsible for their products up to the point of deployment and no further. . . . Treating technology and society as one coherent unit means that developers, deployers, and regulators are accountable not just for the mechanical workings of technical machines, but for how those machines integrate with, reshape, and sometimes harm social systems."[111]

This makes clear that, as Olya Kudina and Ibo van de Poel observe, "how we should understand values like fairness and explainability in the design of AI systems will partly depend on the specific context and the concrete sociotechnical systems [that technology] will be embedded in. Designers should not only rely on general design principles or general fairness metrics, but should also tailor these concerns to the specific context and the sociotechnical systems in which the relevant AI technology is to function."[112] This underscores that "technologies are not neutral tools to be used by people," but have design features that "help to give shape to specific moral intuitions, avenues for moral decisions and even the infrastructure for moral decision-making."[113]

3. Stage Three: Procurement and Acquisition

In order to approve procurement of a system, government personnel need to verify that a system is developed in accordance with specified requirements. This means that these personnel need to understand the technical details that

---

[107] *Id.* at 38–39.

[108] *Id.* at 42.

[109] ALLIK ET AL., *supra* note 10.

[110] *Id.*

[111] CHEN & METCALF, *supra* note 19, at 6.

[112] Kudina & van de Poel, *supra* note 4, at 4.

[113] *Id.* at 5.

characterize normal system performance and the system's failure modes.[114] They must ensure that system training and certification requirements include the potential ramifications of both normal operating modes and system failure modes for compliance with IHL.[115]

4. *Stage Four: Testing, Evaluation, Verification, and Validation TEVV)*

Testing software typically involves identifying the inputs software will receive and confirming that it will produce appropriate outputs.[116] Evaluation involves assessing test results in terms of how well the system performs specified requirements, and, apart from requirements, may also assess "human usability, environmental impact, interoperability, or cybersecurity.[117]

Verification ensures that the system meets formal specifications, including legal and operational ones. This includes assessments of components and subsystems, which can be done either by testing or by purchasing components with appropriate certification. Thus, "verification needs to confirm that the system, once built, meets the requirements."[118] Finally, validation is the process of checking that the system will be suitable for its intended purpose. This may involve deploying the model in realistic operational settings or using simulations to evaluate performance under real-world scenarios and conditions. It also involves determining the ability of eventual system users such as commanders and operators to understand the operation of the system sufficiently to trust and rely on it in decision-making in an appropriate extent.

Assurance of a system's ability to comply with IHL requiring translating legal terms into a form that can be tested. "Meeting requirements such as 'the system will never target a noncombatant,'" for instance, "depend[s] on being able to create specifications for the system, subsystems, and components that can be verified."[119] Testing and evaluation should be part of a process that enables parties to have confidence that a system will be safe, secure, and reliable, and the conditions under which this confidence is and is not warranted. This will enable commanders to determine when it is appropriate to deploy a system and the risks involved when doing so. User trust in the system also will depend on the ability of the machine to present an interface that is transparent and easily interpretable in providing situational awareness.[120]

Autonomous capabilities complicate the test, evaluation, verification, and validation processes compared to use of these processes for conventional software. As a 2020 report by former high-level national security officials notes, machine and deep learning:

---

[114] ALLIK ET AL., *supra* note 10, at 43.

[115] *Id.*

[116] *Id.* at 45.

[117] *Id.*

[118] *Id.*

[119] ALLIK ET AL., *supra* note 10, at 45.

[120] *Id.* at 46.

> [H]as trouble functioning correctly if the inputs or environmental conditions change. However, testing these systems in all possible scenarios and with all ranges of inputs is simply not feasible. It is nearly impossible to predict all the ways a system could break, or an adversary could manipulate or spoof it.[121]

Furthermore, "[a]s these systems become more sophisticated, as is the case with deep learning, their output becomes even less transparent, making it harder to determine the conditions under which they might fail and what steps could correct system behavior."[122] The report offers suggestions for how to make progress in tackling these issues.[123]

Effectively addressing these and other distinctive challenges will be crucial in ensuring that use of AI-enabled weapons comply with IHL. This underscores that IHL compliance depends on the effective functioning of an entire sociotechnical system that includes humans and machines collaborating on a multitude of tasks well before a target is identified and attacked.

5. Stage Five: Education, Training, and Human-System Integration

Education and training is necessary to ensure that developers, commanders, and users understand how a system works, its capabilities and limitations, the appropriate conditions under which it can be used, and the risks for IHL compliance of unexpected system behavior. This understanding "is necessary for political decision-makers, commanders, and users to be reasonably held responsible for the results of decisions to deploy or not deploy a system."[124]

The Framework notes:

> Humans can more easily predict the types of mistakes that other humans will make than the types of mistakes that machines will make. Is it possible to train users to anticipate the types of mistakes that will likely be made by a system they are working with? If not, a determination needs to be made if the AIS considered for use in a military setting is too unpredictable, potentially putting it in violation of IHL.[125]

---

121 MICHÈLE A. FLOURNOY, AVRIL HAINES & GABRIELLE CHEFITZ, BUILDING TRUST THROUGH TESTING: ADAPTING DOD'S TEST & EVALUATION, VALIDATION & VERIFICATION (TEVV) ENTERPRISE FOR MACHINE LEARNING SYSTEMS, INCLUDING DEEP LEARNING SYSTEMS 8 (2020), https://cset.georgetown.edu/wp-content/uploads/Building-Trust-Through-Testing.pdf [https://perma.cc/CUL2-GFDF].

122 *Id.*

123 *Id.* at 19–28.

124 ALLIK ET AL., *supra* note 10, at 48.

125 *Id.* at 49–50.

6. Stage Six: Political and Strategic Considerations

Political leaders and senior military officials need to understand the costs, benefits, and risks of a system with AI capabilities compared to one without them.[126] The also need to be aware of need to understand the chain of command, and command and control procedures, including whether it is possible to recall a weapon once it is activated.

Article 36 of II to the Geneva Conventions requires that consideration of a new weapon, means or method of warfare" requires a state to determine whether using it would, "in some or all circumstances," violate international law, most notably IHL.[127] The challenges posed by the characteristics of weapons that incorporate AI have led to the recommendation that states conduct this legal review earlier in the lifecycle than for a conventional weapon, probably at the development stage, to provide sufficient assurance.[128] This is because lawyers will need to "take into account the potential legal ramifications of degraded system performance or system failures, including how deviations from conditions experienced during testing could affect system performance in unexpected ways."[129]

7. Stage Seven: Operational-Level Command and Control

Assuming prior approval of the use of a system, commanders at the operational level need to understand "normal system performance, system limitations, restraints, constraints, risks, risk mitigation measures, and potential failure modes," including the risk of IHL violations in the case of unexpected system behavior.[130] Commanders need to know under what conditions a human may intervene in the system's operation, and whether it is possible to recall a weapon.

In choosing to deploy an AIS, operational-level commanders should require a comparison of the advantages and risks of this type of weapon to traditional ones. Commanders "should be able to clearly articulate their decision to request AIS based on understanding the technology, advantages, and disadvantages of autonomy over other available systems, risks—including to mission and to force—legal considerations, and geopolitical constraints."[131] The Framework says that operational-level commanders also should require a formal legal review of an AIS prior to any use to consider the risk of IHL implications of any potential system failure modes.[132]

---

[126] *Id.* at 52.

[127] Protocol Additional to the Geneva Conventions of 12 August 1949 and Relating to the Protection of Victims of International Armed Conflicts, art. 36, June 8, 1977, 1125 U.N.T.S. 3, 21.

[128] Tobias Vestner & Altea Rossi, *Legal Reviews of War Algorithms*, 97 INT'L L. STUD. 509, 513 (2021).

[129] ALLIK ET AL., *supra* note 10, at 53.

[130] *Id.* at 55.

[131] *Id.*

[132] *Id.* at 56.

8. Stage Eight: Tactical Employment

Tactical-level personnel should receive in-depth training on a system prior to deploying it to their area of operations. The training should include information on the system's capabilities and limitations, its normal operating parameters, how a system will respond if there is a loss of communications, potential abnormal behavior, and legal and policy guidance on its use. Tactical commanders and personnel should be briefed on any legal review of a system, including the legal ramifications of failure modes, and should be able to articulate why they wish to use an AIS instead of other types of weapons.

9. Stage Nine: Review, Reuse and/or Retire

In this final stage of the lifecycle, there is a review to determine if the system should be reused in its current form, updated or modified for future use, or retired from use. Paraphrasing the report, among the questions to address at this stage are:

> Was the intended purpose for which the system was used articulated and documented prior to deployment? Did the system's functions satisfy this intended purpose? Did it function as expected?
>
> If the system functioned fulfilled the purpose and performed as expected, do we know why? Can we determine if commanders and users did not fully understand the system but it nonetheless functioned as intended?
>
> If the system did not function as intended, do we know why? To what extent was it a result of technical error, or an error by the commander or operator, or that the system was not fit for the mission? Can we identify at what point in the lifecycle the problem(s) occurred? Is it clear who was responsible? Is there documentation to establish this?
>
> Is another round of TEVV necessary before the system can be reused? Who will need to do this: military, contractors, or an independent group?
>
> If the system functioned as intended, does it matter if the commanders and users understand why?[133]

B. *Ongoing Activities Across the Lifecycle*

The Framework describes five activities that involve "the most important aspects of human responsibility and accountability" with regard to an AIS in a military setting. Defining how these activities will be carried out and who is responsible for them should be done at the outset of a potential project.

---

[133] *Id.* at 60- 6.

1. Evaluation of Legal, Ethical, and Policy Concerns

Because AI capabilities and the nature of battlefields are dynamic and constantly evolving, "a number of legal, ethical, and policy concerns need to be addressed at every stage of a system's lifecycle. Legal experts should be involved from the conception of a system rather than waiting until it is in development to consider its legality."[134]

2. Responsibility, Accountability, and Knowledge Transfers

There must be documentation of knowledge about the system and who is responsible for what tasks at each stage of the lifecycle. This helps ensure that those who must rely on the work of others have a clear understanding where things stand and whom to contact for what information if the need arises. "At any stage of the lifecycle, it should be possible to look back to see who was responsible for each stage or project within a stage. It needs to be clear how, when, and to whom their responsibility was passed during transitions between stages."[135] This requires clear communication among people with disparate areas of expertise, such as developers, designers, test and evaluation experts, software engineers, procurement officials, commanders, and operators.

In particular, it needs to be clear when a human is in control of the system and when control is transferred to a machine. As the framework observes:

> This needs to take into account that control may shift back and forth between the human and the system during certain missions. Such handoffs are a technical problem that needs to be solved in advance. However, this is also a question of user interface, training, and contextual understanding that will be dependent on --and need to be addressed – for each system and each mission.[136]

3. Training, Education, and Human-System Integration

Everyone involved in a system needs to have an understanding of the capabilities, limitations, and risks of a system enabling them to play their roles. Training thus must be tailored to these roles. In addition, there must be attention to interfaces that enable operators to understand the state of the system as part of their situational awareness. More generally, system development and design must take into account research findings on how humans tend to interact with machines. "Humans have a tendency toward automation bias," for instance, "and other cognitive limitations, such as finding it difficult to maintain focus for extended time frames, monitor repetitive tasks, or stay engaged with secondary tasks as workload increases. These need to be

---

[134] ALLIK ET AL., *supra* note 10, at 24.

[135] *Id.* at 25.

[136] *Id.* at 31.

addressed in a system's design and considered and monitored through all stages."[137]

4. TEVV, Monitoring, Hardware System Or Software Updates And Interoperability, Maintenance

TEVV occurs during research and development, procurement and acquisition, and education, training, and human-system integration. TEVV will also be necessary in case of system updates based on new data or online updates. In addition, if those responsible for post-deployment monitoring flag a problem, the system probably will require maintenance or updates, which in turn requires a new TEVV process. Even routine maintenance may require updates that involve this. In both cases, there likely will need to be intensive monitoring for a period of time to ensure that the system is acting appropriately and unexpected behaviors do not occur as a result of the updates.[138]

5. Risk Assessments

Risk assessments must occur regularly throughout every stage of the lifecycle. This is because risks from a system can arise for a wide range of reasons:

> [S]uch as it did not function as intended; it was used incorrectly; it was poorly trained, or trained on incomplete data that do not span operational uses; it was used in the wrong situation; the user did not understand its limitations; the developers misunderstood how the system would be used; or it was hacked or otherwise manipulated. The use of a system might also pose political risks, risks of proliferation, or risks of escalating a situation.[139]

The National Institute of Standards and Technology *Risk Management Framework* notes that different risks may arise at different points in the AI lifecycle. For example, it says:

> [A]n AI developer who makes AI software available, such as pre-trained models, can have a different risk perspective than an AI actor who is responsible for deploying that pre-trained model in a specific use case. Such deployers may not recognize that their particular uses could entail risks which differ from those perceived by the initial developer.[140]

The analytical framework of a lifecycle thus "provides a granular means for identifying the technical considerations that are necessary to meet legal requirements, to inform and guide ethical discussions, and to determine human

---

[137] *Id.* at 26.

[138] *Id.* at 27.

[139] ALLIK ET AL., *supra* note 10, at 28.

[140] RAIMONDO, *supra* note 91.

responsibility and accountability across the entire lifecycle of AIS."[141] It therefore can enable development and design "around [human] intent as early as possible and maintain that throughout the lifecycle."[142]

C. *Human Judgment in the Autonomous Phalanx Weapon System*

The IEEE Framework enables us to appreciate the numerous instances in which humans exercise judgment informed by moral and legal considerations in the lifecycle of the Phalanx weapon system in autonomous mode. Ideally, the process would have proceeded in the following way. Humans determined prior to beginning development of the system that a problem could arise when naval vessels were attacked by a barrage of weapons that were too numerous for a human to identify and attack quickly enough to defend the vessel. The explicitly articulated purpose of the weapon was to address this threat, which limited its intended use to situations in which this extreme set of circumstances occurred. The fact that the system would be used at sea also minimized the likelihood of unintended civilian casualties if the system did not operate as intended.

This meant that the Phalanx would not be a fully autonomous system that operated in all cases without direct human involvement. Nor would it be an offensive weapon or one that automatically shifted to autonomous mode based on certain sensor inputs. Rather, autonomous operation was an option that a captain could choose based on his or her judgment that it was necessary to defend the ship. Furthermore, authorized targets would be only certain enemy military objects, not humans. All this provided the basis for a common understanding of the system that was to inform all tasks in every subsequent stage of the lifecycle.

Research and development relied on this understanding to translate it into technical specifications the system needed to meet, and how it needed to be designed, in order to accomplish its purpose. Determination of the specific enemy weapons that the system could target in autonomous mode, and the munition that it would use to do so, guided this process. Work in this stage also determined normal system performance and identified ways in which the system could fail. Developers may have consulted lawyers at this stage for guidance on how development and design choices could ensure compliance with IHL. In addition, the system needed to be designed so that the interface with a human user made it easy for the user to evaluate how the system was behaving.

The TEVV process then used the technical specifications of the system as the basis for determining that individual components, as well as their integration into the system as a whole, would meet those requirements. It also sought to assess the probability of and reasons for plausible failure modes. This analysis required reliance on human judgment to determine acceptable levels of performance, reliability, and safety. The work in this stage was guided by the goal of providing sufficient assurance to commanders and operators that they could trust the system in autonomous mode under the specified conditions to

[141] ALLIK ET AL., *supra* note 10, at 10.

[142] *Id.* at 15.

behave as intended in accomplishing the purpose of the system while complying with IHL.

Related to this last point, training and education ideally ensured that all relevant parties, especially operational commanders and tactical commanders and operators, understood the appropriate conditions for transfer to autonomous mode, how the system would behave once this was done, the risks of the weapon operating outside its intended parameters, how to determine when this occurred, and how to respond if it did. It also should have ensured that users understood how the system interface indicated the system's behavior and provided information on the situation more generally. Ideally, work at the design stage should have engaged with these stakeholders to make sure that the interface was easily interpretable for this purpose.

Similarly, there should have been education of political and senior military leaders regarding the features of the system, its operations, and the risks that it could pose so that they understood what considerations should inform the allocation of authority over its use. Education and training of all these parties would be designed to ensure that humans were able to exercise judgment informed by moral and legal considerations.

The final stage of the lifecycle offered the opportunity to review the performance of the Phalanx after deployment, make revisions to it based on these assessments, and retire systems as appropriate. As one source says, "[d]ue to the continuing evolution of both threats and computer technology, the Phalanx system has, like most military systems, been developed through a number of different configurations."[143]

Finally, there ideally have been activities "in which human decision-making elements occur repeatedly throughout the lifecycle."[144] As the IEEE Framework describes, these require human judgments about legal and ethical concerns; responsibility and accountability; training and education; human-system integration; TEVV; and risk assessments, all of which are intended to ensure that moral and legal consideration inform every activity in each stage. "The focus throughout," the Framework says, "is on the humans involved in decisions. The framework is designed to help identify the relevant decision-makers, their role at each stage, the knowledge and documentation they need from previous decision-makers, and the types of questions and issues to be addressed at each stage."[145]

Situating the Phalanx in autonomous mode within the sociotechnical system of its lifecycle and the ongoing activities in it thus enables us to say that human judgment informs this use of the system even though a human is not directly involved in identifying and engaging targets. Humans have determined the circumstances in which the system may operate autonomously, the objects that it can target when it does, and what it may use to do so. Humans also have

---

[143] *Mk-15 Phalanx Close-In Weapon System (CIWS)*, *supra* note 17.

[144] ALLIK ET AL., *supra* note 10, at 7.

[145] *Id.* at 8.

collaborated with machines to ensure that they operate within these parameters, to inform humans when they do not, and to establish processes for humans to follow if this occurs.

While not framed in terms of the AI lifecycle, ICRC recommendations to states regarding the focus of regulation are consistent with this focus. The ICRC calls for:

> (1) limits on the types of target, such as constraining them to objects that are military objectives by nature;
>
> (2) limits on the duration, geographical scope and scale of use, including to enable human judgement and control in relation to a specific attack;
>
> (3) limits on situations of use, such as constraining them to situations where civilians or civilian objects are not present; and
>
> (4) requirements for human–machine interaction, notably to ensure effective human supervision, and timely intervention and deactivation.[146]

A particular benefit of the lifecycle perspective is that it enables a more sophisticated understanding of system failure than simply attributing it to what is commonly described as "human error" or "technology malfunction."[147] The well-known incident of the *USS Vincennes'* mistaken destruction of an Iranian commercial airliner in 1988 illustrates this point.

D. *The USS Vincennes Incident*

On July 3, 1988, the naval vessel *USS Vincennes* was on patrol in the Strait of Hormuz in the Persian Gulf along with other U.S. naval vessels assigned to protect commercial vessels from attacks by Iranian speedboats. The *Vincennes* had an advanced Aegis control system that could project an image of an air battle over several hundred square miles. It also could identify and track aircraft and could distinguish friendly and potentially hostile ones. Finally, when activated by the captain, it had the ability to autonomously attack a variety of targets such as aircraft and high-speed cruise missiles.

The *Vincennes* was on patrol because Iraq and Iran were at war and had begun targeting U.S. and neutral ships in the Persian Gulf. Both countries had antiship missiles that could fly several miles just above the surface. In addition, attacks by swarms of small armed boats were common. A year earlier, the USS Stark on patrol in the Gulf had been destroyed, and 37 sailors killed, by an attack by an Iranian F-1 fighter plane that fired two Exocet missiles into the ship.

At 9:45 a.m., the *Vincennes* received a report that another U.S. ship, the *USS Montgomery,* was under attack by Iranian gunboats. It moved to protect the *Montgomery* and joined a firefight with 13 boats at 10:13 a.m. During the fight,

---

[146] *ICRC Position on Autonomous Weapon Systems*, *supra* note 1.

[147] *See generally* RAIMONDO, *supra* note 91, at ii, 1; SIDNEY DEKKER, THE FIELD GUIDE TO HUMAN ERROR INVESTIGATIONS (2002).

the *Vincennes's* main machine gun malfunctioned. The captain ordered the crew to activate the other machine gun, which required an abrupt full-rudder turn to spin the ship at 30 knots per hour. This caused considerable disruption, with objects falling and sliding as the ship made a sharp turn.

As this was occurring, the *Vincennes*' crew detected an unidentified aircraft leaving Iran's dual military-civilian Bandar Abbas airfield in Iran on a path directly toward the ship. Around the same time, an Iran Air Airbus A300 civil airliner, Iran Air Flight 655, took off from Bandar Abbas heading for Dubai. The flight first appeared on the *Vincennes* radar at 10:17 a.m. just after the *Vincennes* had opened fire on the Iranian boats. The Aegis system onboard the *Vincennes* correctly indicated that Flight 655 was ascending and that its transponder signal indicated that it was a civilian aircraft.

Over the next three minutes, as the ship continued to turn and exchange fire with the gunboats, the responsible personnel on the ship: (1) mistakenly identified Iran Flight 655 as a possible Iranian F-14 fighter plane; (2) reported hearing transponder signals that indicated that the aircraft was military; and (3) reported the aircraft as descending towards the ship when it was in fact still climbing according to its usual flight plan.

At 10:19 a.m., the *Vincennes* began to issue warnings to Flight 655 on the Military Air Distress frequency, and at 10:20 a.m. began warnings on the International (civil) Air Distress frequency as well. Having informed Joint Task Force Command that a potentially hostile aircraft was 28 nautical miles away, was not responding to warnings, and was rapidly nearing potential missile attack range, the *Vincennes* sought and received permission to attack the plane.

At about 10:24 am, Iran Air Flight 655's flight was eight miles from the *Vincennes*. The ship then fired two missiles at it. A few seconds later, with the flight still on its climb, and well within the prescribed civil air corridor, the missiles hit the plane at an altitude of 13,500 feet. With 290 people aboard, it burst into flames and fell into the Persian Gulf. No one survived. The whole flight had taken less than seven minutes.

The official Investigation Report attributed the accident to human error arising from a stressful situation in which the *Vincennes* was simultaneously involved in a firefight with small, armed Iranian speedboats, the captain was receiving information that indicated a high risk that the aircraft was hostile, and he reasonably could have suspected that the aircraft was connected to the firefight with gunboats.[148] "Given the time available," the Report said:

> [T]he Commanding Officer could hardly meet his obligation to protect his ship and crew and also clear up all of the possible ambiguities . . . Although it might not seem fair, commanding officers do not have the luxury of reconciling all such questions

[148] U.S. DEP'T OF DEF., AD–A203 577, FORMAL INVESTIGATION INTO THE CIRCUMSTANCES SURROUNDING THE DOWNING OF IRAN AIR FLIGHT 655 ON 3 JULY 1988, at 6 (1988).

> before committing themselves. They have to go with the weight of evidence.[149]

The Report noted that the complex Aegis technology on board the *Vincennes* correctly indicated throughout this period that the Iranian plane was ascending after taking off from a nearby airport, and that its transponder signaled that it was a civilian aircraft.[150] In this sense, the incident was not due to a problem with the technology. The Report did note that there were some recommendations for improvement in the Aegis system when operating in a close-in environment rather than the open-sea environment for which it had been designed.[151]

One comment in the Report, however, went beyond separate assessment of humans and technology and focused on the interaction between the two. The Report suggested that "some additional human engineering be done on the display systems of AEGIS. The objective would be to better equip it for assisting with rapid decisions in a situation such as [the USS] *Vincennes* confronted."[152] Specifically, "means for displaying altitude information on a contact such as 'ascending' or 'descending' on the LSD [Large Screen Displays] should . . . be examined."[153] As Jeffrey Ding has noted concerning the incident, while the technology indicated the altitude of the plane, it was not designed to provide information about the rate of change in altitude. As a result, operators had to compare altitude data at different times and calculate the rate of change in their heads or by hand while engaged in combat.[154] Ding explicitly directs attention to the Aegis lifecycle by noting that a systems engineer had suggested an interface that indicated the rate of altitude change, but this had not been adopted in the design stage.[155] Decisions made at the software development stage therefore may have made the accident more likely.

More generally, Ding argues that the portion of the AI lifecycle that involves the military software development and acquisition process provides too little opportunity for early operator feedback to software developers. He observes that "software can 'fail' in the sense that it does not meet the needs of users, even if it passes all the technical requirements established in the system acquisition phase."[156] Studies by human-computer interaction specialists "have identified confusing interfaces as contributing to inadvertent military launches, fratricide, and other safety hazards."[157] The risk of these is intensified in situations in

---

[149] *Id.*

[150] *Id.* at 7.

[151] *Id.* at 8.

[152] *Id.*

[153] U.S. DEP'T OF DEF., *supra* note 148, at 8.

[154] Jeffrey Ding, *Machine Failing: How Systems Acquisition and Software Development Flaws Contribute to Military Accidents*, 8 TEX. NAT'L. SEC. REV. 9, 19 (2025).

[155] *Id.*

[156] *Id.* at 14.

[157] *Id.*

which operators must rapidly interpret data and make decisions under stressful conditions.

Thinking of an AI-enabled weapon as a sociotechnical system that involves a lifecycle thus illuminates the complex ways in which a multitude of interdependent activities can determine compliance with IHL and broader moral concerns.

VIII. CONCLUSION

As this Article has described, the CCW GGE has declared that AI-enabled weapons must be "operated with context-appropriate human control and judgment," which "includes ensuring moral and legal consideration is given by a human, in particular, during identification and/or engagement stages."[158] Many discussions of AI-enabled weapons that emphasize the need for human judgment that takes moral and legal considerations into account focus on the final steps of identifying and attacking targets. I have described how this neglects the fact that these steps are nested within a larger targeting sociotechnical system in which humans exercise moral and legal judgment on multiple occasions. This targeting system is further embedded within a larger sociotechnical system consisting of the lifecycle of an AI-enabled weapon in which there are even more points at which humans can exercise such judgment.

Whether the use of an AI-enabled weapon is informed by distinctive human moral and legal concerns therefore depends upon the operation of the system as a whole rather than on whether a human or machine performs a particular function. In the case of the *USS Vincennes*, for example, had the Aegis technology been in autonomous mode it would not have fired at the Iranian commercial airliner. It was the interaction between the humans and the technology that resulted in the terrible mistake. Had an earlier stage in the Aegis lifecycle addressed the potential problem with this interaction, the lives of 290 people might have been spared.

As one scholar writing on the *Vincennes* incident has said:

> If the large-scale social-organizational-technical system fails to produce the right outcome, then it has failed *as a system*. Whether the proximate cause was human action or malfunction of the equipment is a matter of detail . . . It does not absolve us of the need to enquire . . . whether the inability of the operators to avoid a mistake or recover a deteriorating situation occurred because of the structure and context of the large-scale technical system in which their decision was embedded.[159]

Determining whether use of an AI-enabled weapon is informed by human moral and legal judgment therefore needs to go beyond asking whether a human

---

[158] *Rolling Text*, *supra* note 3, at 2.

[159] Gene I. Rochlin, *Iran Air Flight 655 & the USS Vincennes: Complex, Large–Scale Military Systems and the Failure of Control* 23 (Inst. Governmental Stud. & Energy and Res. Grp., U.C. Berkeley, Working Paper No. 90-27, 1990) (emphasis in original).

or a machine performs the last steps of identifying and engaging a target. Helen Toner emphasizes this point:

> There's often an assumption that keeping a human 'in the loop' who has to approve targeting decisions solves the problem of LAWS. What matters is not just whether a human is involved, but how they interact with the AI system in practice—which will depend on questions of design, operator training, and testing in realistic conditions. If those questions are not thought through carefully, the results can be disastrous regardless of whether the human or machine has final say.[160]

A sociotechnical system perspective helps address these questions by directing attention to the numerous interactions between humans and machines, humans and humans, and machines and machines that are necessary to develop and use an AI-enabled weapon system. It is the exercise of distinctive human judgment throughout these interactions that will be most important in ensuring that weapons are guided by human concerns rather than machine logic.

[160] Alex Friedland, *IDF Uses AI to Accelerate Targeting According to Report, the White House Releases Guidance on Federal AI Use, and a Roundup of the Latest Computing and Funding News*, GEO. CTR. FOR SEC. & EMERGING TECH. (Apr. 11, 2024), https://cset.georgetown.edu/newsletter/april-11-2024 [https://perma.cc/CHK4-PDXQ].