CyCon 2024: Over the Horizon
16th International Conference on Cyber Conflict
C. Kwan, L. Lindström, D. Giovannelli, K. Podiņš, D. Štrucl (Eds.)
2024 © NATO CCDCOE Publications, Tallinn

Targeting in the Black Box

Scott Sullivan*
Professor
Army Cyber Institute
United States Military Academy
West Point, NY, United States
scott.sullivan@westpoint.edu

Iben Ricket
Research Scientist
Dartmouth College
Hanover, NH, United States
iben.sullivan@dartmouth.edu

Abstract: Artificial intelligence (AI) is poised to become pervasive in military operations worldwide. In the coming decades, AI-based systems will revolutionize logistics, dramatically change targeting, and ultimately power autonomous weapons systems. Unfortunately, many of the most potent AI-based systems are unintelligible to their developers and offer unexplained outputs to their users—a phenomenon called the "black-box problem." This paper first describes the basic AI architecture that is giving rise to the black-box problem. It then shifts to consider the unexplored question of whether black-box models comport with the international humanitarian law (IHL) principles of distinction, proportionality, and precaution, which are fundamentally rooted in nuanced context and subjective judgment. After describing the mismatch between black-box models and existing IHL principles, the paper compares existing NATO doctrine with emerging "soft law" embraced by NATO member States. Identifying a nascent movement away from explainable AI, the paper concludes by setting out the importance of interpretability and the aspects therein that should be considered by policymakers in constructing future legal norms and by military officials in assessing AI models to be used in future operations.

Keywords: *AI, targeting, autonomous weapons, Israel, neural networks, explainable AI*

* The views expressed in this article are personal and do not reflect the policy or position of any US government entity or organization.

1. INTRODUCTION

At the end of 2023, multiple outlets reported on Israel's widespread use of an artificial intelligence (AI) system to identify targets in its ongoing conflict with Hamas in Gaza.[1] Dubbed "Habsora," or "the Gospel" in English, Israeli Defense Forces (IDF) officials credited the AI system with the ability to increase the number of targetable sites in Gaza from 50 each year to over 100 each day.[2] Commentators quickly recognized AI-based targeting as "an intermediate step [to] autonomous systems that will eventually be deployed to the battlefield."[3]

While various forms of AI targeting and autonomous weapons systems have long been in service, their use has primarily been restricted to circumstances where the risk to civilians was minimal and the targeting question at issue was easy to identify definitively.[4] The introduction of more generally purposed AI-based targeting systems, like Habsora, in service of a conflict against a terrorist organization based in one of the most densely populated areas of the world marks a definitive shift.

AI systems that can target and use force without human intervention, often called lethal autonomous weapons systems (LAWS), have engendered the greatest attention and concern. Proponents suggest that well-designed LAWS may operate more humanely and lawfully by avoiding classically human errors and cognitive biases. Skeptics believe LAWS usage will generally increase armed conflict and risk unpredictable and perhaps unknowable dangers to combatants, civilians, and the larger global order. However, all parties acknowledge that the development and deployment of such systems are inevitable, leading to a recent call to "establish specific restrictions on autonomous weapons systems" by the leaders of the United Nations and the International Committee of the Red Cross (ICRC).[5]

Unfortunately, the attention garnered by the specter of robot soldiers tends to obscure the fact that AI systems powering autonomous weapons, and not the execution of force itself, pose the most imminent and widespread challenge for IHL regulation. Targeting systems like Hasbora exemplify this concern. Much of the system's operations remain

1 *See e.g.*, Harry Davies, Bethan McKernan & Dan Sabbagh, *"The Gospel": How Israel Uses AI to Select Bombing Targets in Gaza*, Guardian (UK) (Dec. 1, 2023), https://www.theguardian.com/world/2023/dec/01/the-gospel-how-israel-uses-ai-to-select-bombing-targets.

2 *Id*. The United States has also recently begun using an AI-based targeting system for operations against Yemen's Houthis and Iran proxy forces. Katrina Manson, *AI Warfare Is Already Here*, Bloomberg (Feb. 28, 2024), https://www.bloomberg.com/features/2024-ai-warfare-project-maven/.

3 Geoff Brumfiel, *Israel Is Using an AI System to Find Targets in Gaza. Experts Say It's Just the Start*, Morning Edition, NPR (Dec. 14, 2023), https://www.npr.org/2023/12/14/1218643254/israel-is-using-an-ai-system-to-find-targets-in-gaza-experts-say-its-just-the-st.

4 Jack M. Beard, *Autonomous Weapons and Human Responsibilities*, 45 Geo. J. Int'l L. 617, 628–32 (2014).

5 *Joint Call by the United Nations Secretary-General and the President of the International Committee of the Red Cross for States to Establish New Prohibitions and Restrictions on Autonomous Weapon Systems*, International Committee of the Red Cross (Oct. 5, 2023), https://www.icrc.org/en/document/joint-call-un-and-icrc-establish-prohibitions-and-restrictions-autonomous-weapons-systems.

out of view. However, its apparent discovery of hundreds of "new" targets, coupled with a large and growing civilian death toll and the widespread destruction of civilian objects, has engendered tremendous criticism of Israeli targeting decisions and its dependence on AI.[6]

This paper seeks to build on the existing literature regarding autonomous weapons to discuss the legal and policy implications surrounding AI modeling systems, such as artificial neural networks, which are poised to fuel the targeting and autonomous capabilities of the future. These AI systems, often referred to as "black-box models," function unintelligibly to the States that develop them and generate outputs that offer little, if any, underlying reasoning to the personnel operating them.

Black-box models, such as artificial neural networks, sometimes called "deep learning," are widely considered the most powerful and accurate AI prediction systems on offer. However, their unintelligible operations and unexplained outcomes stand in contrast to prior iterations of machine learning methods and the software powering earlier "automated" or "autonomous" weapons, which function through increasingly complex, albeit explainable technology. Consisting of multiple layers of interconnected nodes, artificial neural networks require enormous volumes of data and produce outcomes or classifications reflecting "extremely complex non-linear statistical models and innumerable parameters" that lack any "reason or suitable explanation" discernible to the user, or even the architects of the system.[7]

The use of black-box models in targeting poses significant practical and legal challenges under established principles of IHL, a body of law whose principles are built upon a bedrock of context and subjectivity.

2. THE EXPLAINABILITY PROBLEM IN AI

Inherent to all AI systems is the ability to learn patterns in the data to generate a prediction.[8] The nature of the relevant patterns and predictions, of course, depends upon the field. AI models in medicine can identify patients at risk for hospital

6 Evan Hill, Imogen Piper, Meg Kelly & Jarrett Ley, *Israel Has Waged One of This Century's Most Destructive Wars in Gaza*, Washington Post (Dec. 23, 2023), https://www.washingtonpost.com/investigations/interactive/2023/israel-war-destruction-gaza-record-pace/; Jared Malsin & Saeed Shah, *The Ruined Landscape of Gaza After Nearly Three Months of Bombing*, Wall Street Journal (Dec. 30, 2023), https://www.wsj.com/world/middle-east/gaza-destruction-bombing-israel-aa528542.

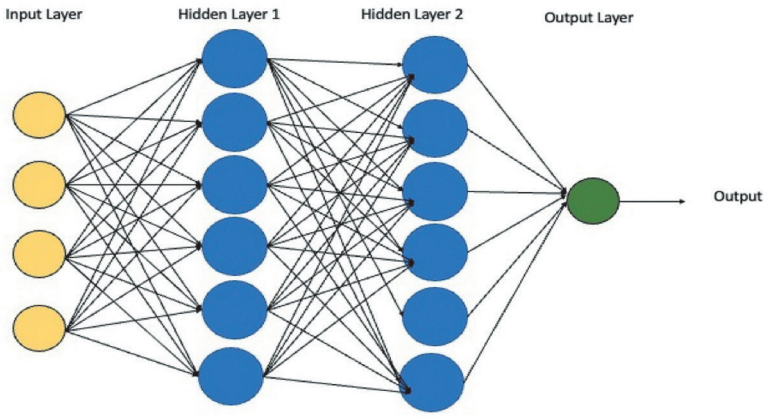
7 Vikas Hassija, Vinay Chamola, Atmesh Mahapatra, Abhinandan Singal, Divyansh Goel, Kaizhu Huang, Simone Scardapane, Indro Spinelli, Mufti Mahmud & Amir Hussain, *Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence*, 16 Cognitive Computation 45 (2023).

8 Iben Sullivan (Ricket), Michael Matheny & Jeremiah Brown, *Breaking Down the "Black-box" of Machine Learning for Predictive Analytics: Results from Models Predicting 30-Day Readmissions Following an AMI*, Nurs. Sci. Conf. Proc. (2023) (on file with the authors).

readmissions or predict the presence of early cancer.[9] Credit card companies and banks use AI algorithms in anomaly detection to identify fraudulent or abusive transactions. Autonomous vehicles rely on AI systems trained on data to identify appropriate and safe driving performance.

While different AI modeling techniques present varying degrees of explainability problems, this paper focuses on neural networks and related deep-learning-based models, which are widely considered the most effective in classification and evaluation and also among the most opaque of AI systems.[10]

FIGURE 1: BASIC NEURAL NETWORK ARCHITECTURE. THE NEURAL NETWORK INCLUDES AN INPUT LAYER WHERE DATA IS FED INTO THE ALGORITHM, TWO HIDDEN LAYERS, AND AN OUTPUT LAYER WHERE DATA IS SCALED TO AN APPROPRIATE RANGE USING AN ACTIVATION FUNCTION



At their most basic structure, as shown in Figure 1, neural networks contain an input layer, one or more hidden layers, and an output layer.[11] Each layer includes nodes or neurons.[12] Data is passed through each layer via a neuron. Each neuron can be considered its own model, with input data, weights, a threshold (bias), and an output.[13] Each input is assigned a weight with larger values signifying greater importance. All inputs are multiplied by their respective weights, summed, and passed through an activation function.[14] If this final output exceeds a specific threshold, the

9 Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis & Dimitrios I. Fotiadis, *Machine Learning Applications in Cancer Prognosis and Prediction*, 13 Computational and Struct. Biotech. J. 8, 8 (2015).

10 Manuel Carabantes, *Black-Box Artificial Intelligence: An Epistemological and Critical Analysis*, 35 AI & Soc. 309, 313 (2020).

11 Rene Y. Choi, Aaron S. Coyner, Jayashree Kalpathy-Cramer, Michael F. Chiang & J. Peter Campbell, *Introduction to Machine Learning, Neural Networks, and Deep Learning*, 9 Transl. Vis. Sci. Technol. no. 2, art. 14 (2020).

12 *Id.*

13 Bogdan M. Wilamowski, *Neural Network Architectures and Learning Algorithms*, 3 IEEE Indust. Electr. Mag. no. 4, 2009, at 56–63.

14 *Id.*

neuron is activated, and the data is passed to the next network layer. As such, one neuron's output is the next layer's input.[15] Data is passed from one layer to the next until it reaches the final output layer, which defines the acceptable range of output (e.g., -1–1 or 0–1) by scaling or transformation.[16] While simple feedforward neural networks may contain a few hidden layers, where data moves through the network in one direction (forward), more complex and deeper networks can contain hundreds or thousands of hidden layers and feed data forward and backward through the layers.[17] Special neural network architectures exist for specific data inputs or use cases.[18] For example, convolutional neural networks are designed for images, and recurrent neural networks are used for text.[19]

The architecture and processes inherent to neural networks pose two fundamental problems of understandability. The first relates to complexity. Even a cursory explanation of neural network architecture illustrates the intractable challenge of following inputs through the network to a final predicted output.[20] As the complexity of the system increases, its opacity rises as well.[21] The most accurate and powerful neural network models—presumably the type of models States would desire to build and use in armed conflict—require endless volumes of data and create millions, if not billions, of parameters.[22] Thus, a single prediction generated from a deep learning algorithm can involve millions of mathematical and computational operations, making traceability from data to prediction (or back) functionally impossible.[23]

Furthermore, the nature of a neural network's architecture and operations, regardless of type, size, or complexity, is so cognitively dissimilar to humans that there exists a fundamental "mismatch between [the model's] nature and [human] understanding," and as a result, the models "are not intelligible no matter how much knowledge we possess on mathematics, computation, or any other related science."[24]

Collectively, these challenges for users and developers to understand how certain AI models operate and formulate specific predictions are commonly called the "black-box problem."[25] The use of black-box models has proliferated in recent years in areas of

15 Choi et al., *supra* note 11, at 14.
16 Wilamowski, *supra* note 13, at 56–63.
17 Choi et al., *supra* note 11, at 14.
18 *Id*.
19 *Id*.
20 Plamen Angelov & Eduardo Soares, *Towards Explainable Deep Neural Networks (xDNN)*, 130 Neural Netw. 185, 185–89 (2020).
21 Choi et al., *supra* note 11, at 20.
22 Angelov & Soares, *supra* note 20, at 188.
23 *Id*.; Rabia Saleem, Bo Yuan, Faith Kurugollu, Ashiq Anjum & Lu Liu, *Explaining Deep Neural Networks: A Survey on the Global Interpretation Methods*, 513 Neurocomputing 165, 165–68 (2022).
24 Carabantes, *supra* note 10, at 314.
25 Pantelis Linardatos, Vasilis Papastefanopoulos & Sotiris Kotsiantis, *Explainable AI: A Review of Machine Learning Interpretability Methods*, 23 Entropy, no. 1, 2020, at 3; Scott M. Lundberg et. al., *From Local Explanations to Global Understanding with Explainable AI for Trees*. 2 Nature Mach. Intell. 56 (2020).

decision-making with high-stakes outcomes, such as healthcare and criminal justice.[26] In such areas, the perceived higher performance of black-box models relative to their simpler counterparts is pitted against the inscrutability—and potential hidden biases and errors—intrinsic to black-box AI.[27]

3. EXISTING LAW AND THE ROLE OF EXPLANATIONS

Explanations are fundamental to functioning legal systems. In domestic legal systems, law enforcement officers must be able to articulate the evidence justifying their arrests. Judges author opinions to explain how they arrived at their decisions and to further define the contours of legal rules for future controversies. Administrative agencies are required to explain the reasoning behind their establishment of regulations.

Explainability may be even more significant in the international legal system. International legal obligations are typically drafted at a much higher level of abstraction than their domestic counterparts due to their subjects' tremendous economic, cultural, and legal differences.[28] Despite this, the absence of a "centralized enforcement mechanism" to coerce compliance means that the force of international legal rules flows, at least in part, on States' ability to understand the boundaries of the rule and the reasoning underlying such boundaries.[29]

Within the specific context of IHL, the legality of State action frequently turns on the veracity of the proffered explanation. Such explanations can, for instance, serve to distinguish between an accepted act of war and a war crime. However, this paper focuses on how the larger structural principles of IHL embed the concept of explainability of targeting decisions and how reliance on black-box AI systems threatens the fabric of that system.

There is general agreement that AI weapons systems, autonomous or not, must conform with existing IHL. NATO doctrine provides that AI weapons systems must be "developed and used in accordance with … international humanitarian law and human rights law."[30] The more recent US-led *Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy* similarly stipulates that the "use of AI in armed conflict must be in accord with States' obligations under international

26 *See* Cynthia Rudin, *Stop Explaining Black-Box Models for High Stakes Decisions and Use Interpretable Models Instead*, 1 Nature Mach. Intell. 206, 207–10 (2019).

27 *Id*. The exchange between a model's inherent understandability and its performance is often referred to as the "performance–interpretability tradeoff."

28 *See* Jack Goldsmith & Daryl Levinson, *Law for States: International Law, Constitutional Law, Public Law*, 122 Harv. L. Rev. 1791, 1824 (2009).

29 *Id*.

30 Zoe Stanley-Lockman & Edward Hunter Christie, *An Artificial Intelligence Strategy for NATO*, NATO Review (Oct. 25, 2021), https://www.nato.int/docu/review/articles/2021/10/25/an-artificial-intelligence-strategy-for-nato/index.html.

humanitarian law including its fundamental principles."[31] One of the significant challenges with assessing the legal status of AI on the battlefield, especially in complex combat circumstances, arises from the fact that the applications and architecture of such systems are largely theoretical or legally protected secrets. As legal scholars argued nearly a decade ago, "there is no reason, in principle, why a highly automated or autonomous system could not satisfy the requirements of targeting law."[32] The same could be said of an AI-based targeting system.

With the dawn of AI-based targeting upon us, our attention must turn to the designers of these systems and the commanders employing them, who will be required to exercise judgment in a manner consistent with IHL. Of course, responsible judgment can only occur when it is well-informed. It is this informed judgment requirement, which is embedded in existing humanitarian law and the emerging norms surrounding AI weapons systems, that black-box AI models challenge.

A. The Principles of Distinction and Proportionality

Distinction and proportionality reflect two of the most basic legal principles underlying IHL.[33] Distinction is often considered the principle upon which AI-based weapons systems are most likely to succeed, especially if success is judged by their perceived ability to outperform their human counterparts. Distinction requires a combatant to use "reasonable judgment" to differentiate combatants and military objects from civilians and civilian objects.[34] A variety of peculiarly human traits with tragic consequences often compromises human targeting decisions.[35] Untouched by these factors, machine-learning systems armed with vast quantities of high-quality data gathered over time would presumably be well-positioned to identify the patterns of combatants and military objects and distinguish them from civilians and civilian objects. Such a system would be especially effective where, such as in the war in Ukraine, most combatants would be targetable based on their status as members of a State's armed forces or another organized armed group.

Legal nuances quickly proliferate when considering the targeting of civilians. Civilians are generally protected from attack but forfeit that protection when "directly

31 *Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy*, U.S. Dep't of State (Nov. 9, 2023), https://www.state.gov/wp-content/uploads/2023/10/Latest-Version-Political-Declaration-on-Responsible-Military-Use-of-AI-and-Autonomy.pdf.

32 Kenneth Anderson, Daniel Reisner & Matthew Waxman, *Adapting the Law of Armed Conflict to Autonomous Weapons Systems*, 90 US Int'l L. Stud. 386, 406 (2014).

33 *Id*. at 401.

34 Geoffrey S. Corn, *Targeting, Command Judgment, and a Proposed Quantum of Information Component: A Fourth Amendment Lesson in Contextual Reasonableness*, 77 Brook. L. Rev. 437, 454 (2012) ("each ad hoc targeting decision must be the result of a reasonable judgment").

35 *See generally*, Kevin Jon Heller, *The Concept of "the Human" in the Critique of Autonomous Weapons*, 15 Harv. Nat'l Sec. J. 1 (2023).

participating" in hostilities.[36] While many instances of direct participation are uncontroversial—for example, firing upon combatants—the scope of what behaviors constitute direct participation is undefined and largely undefinable. For example, *The Commander's Handbook on the Law of Naval Operations*, issued by the US, indicates that "there is no definition of direct part in hostilities in international law" and, as such, combatants "must make an honest determination" based on "all relevant available facts in the circumstances prevailing at the time."[37] Even when a civilian might be found to be directly participating, a temporal question arises. Because civilians are only targetable "so long as" they directly participate in hostilities, their targetability ceases when their participation ceases. This temporal aspect has led to competing standards, both of which require a nuanced, case-by-case assessment that is also highly transient, given the innumerable ways by which participation might cease.[38]

Despite these intricacies, the determination emanating from a black-box AI system would ultimately be binary, with each person or object delineated as either a target or non-target. Beneath the binary nature of the classification would be a probabilistic determination of the model's prediction. In the absence of any explanation from the system regarding how the model came to its conclusion, there is little "judgment" to be exercised by the commander who receives the model's determination. He is either to trust the system or not. The "honest judgment" of the commander is replaced by the decision to trust the AI system or not.[39]

Proportionality analysis poses far more substantial challenges. The principle of proportionality demands that "the incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof" of an attack not "be excessive in relation to the concrete and direct military advantage" to be accomplished.[40] In essence, proportionality requires the balancing of two qualitatively different interests. Of these, insofar as the reliable identification of civilian persons and objects can be

36 Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts, June 8, 1977 [hereinafter AP I], art. 51; *see also* Nils Melzer, Interpretive Guidance on the Notion of Direct Participation in Hostilities under International Humanitarian Law (International Committee of the Red Cross 2009), https://www.icrc.org/en/doc/assets/files/other/icrc-002-0990.pdf.

37 Dep't of the Navy, The Commander's Handbook on the Law of Naval Operations, NWP 1-14M/MCTP 11-10B/COMDTPUB P5800.7A, § 8.2.2 (2022). An earlier iteration of the handbook similarly stated that "direct participation in hostilities must be judged on a case-by-case basis" as to each "particular civilian" that might be subject to attack. Dep't of the Navy, The Commander's Handbook on the Law of Naval Operations, NWP 1-14M, § 8.2.2 (2007).

38 The "continuous combat function" espoused by the ICRC requires a fine-grained analysis of acts that give rise to a quantitative and qualitative assessment of when sporadic involvement in hostilities passes the relevant threshold. *See* Melzer, Interpretive Guidance, *supra* note 36, at 33–6. In contrast, the membership test poses a similarly intensive examination of whether an individual's social network and behaviors are sufficiently tied to an armed group so as to render him targetable as a member of that group.

39 While the prudential considerations attached to exercising judgment on whether to attack a targetable person or object are not strictly required by distinction, the practical limitations on a commander's options are significant.

40 Customary International Humanitarian Law (Jean-Marie Henckaerts & Louise Doswald-Beck eds., International Committee of the Red Cross 2005), Rule 14.

achieved, identifying the anticipated cost of an attack might be the easiest aspect for an AI-based targeting system to accomplish. To the extent that a system can correctly identify civilians and civilian objects, their presence (or absence) combined with the means of the attack potentially used in the operation would pose a relatively straightforward assessment of civilian cost. In contrast, the variations on the "concrete and direct military advantage" of a specific attack are much more subjective and, as such, difficult to capture by quantitative means. Put simply, "measuring concrete and direct military advantage will always be part of subjectivity" as it weighs elements "that are not quantifiable."[41]

Understanding an operation's direct and concrete military advantage requires an appreciation of how a specific attack fits within a larger operation and the tactical and strategic advantages the attacking party seeks to capture. Consequently, the quantitative and qualitative aspects of military advantage are not only highly nuanced—similar to distinction—but also highly dynamic and fluid. For example, the military advantage of destroying an arms depot is much higher when new intelligence has identified it as the only repository of specific munitions upon which your enemy relies. States and scholars alike disagree on the scope of considerations relevant to assessing the "concrete and direct military advantage" of a target and whether the assessment should be limited to individual attacks or extended to a broader scale.[42]

While States have proposed different formulations of engaging in proportionality analysis, there is consensus that human judgment governs the analysis. The German government, for instance, has stated that the relevant calculations must be made on "a case-by-case basis and that no abstract calculations [are] possible."[43] Canada requires its commanders to possess a subjective "honest and reasonable expectation that the attack will make a relevant contribution to the success of the overall operation."[44] Further, the gravity of the military advantage includes the "security of the attacking forces."[45]

AI targeting platforms like Habsora undermine rather than effectuate the subjective standards enshrined in distinction and proportionality. Israel, like other States, has explicitly rejected the existence of a uniform calculation for balancing concrete military advantage in proportionality analysis.[46] However, descriptions of how Habsora communicates proportionality judgments to the units carrying out attacks

41 AP I, *supra* note 36, art. 51(5).

42 *See* Nelleke H. Hoffs, *Deducing the Measuring Standard of "Concrete and Direct Military Advantage Anticipated," Referred to in Article 51(5)(b) of Additional Protocol I to the 1949 Geneva Conventions* (Sep. 21, 2013) at 8, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2329212.

43 *See ICRC—Customary IHL—Practice Relating to Rule 14 Proportionality in Attack*, Government of Germany https://ihl-databases.icrc.org/en/customary-ihl/v2/rule14?country=de.

44 Office of the Judge Advocate General, Canada, The Law of Armed Conflict at the Operational and Tactical Level, p. 4-3, §§ 20 and 21 (1999).

45 *Id*. Presumably the more insecure the attacking forces, the lower the threshold of the advantage requirement.

46 IDF Sch. of Military Law, Isr., Rules of Warfare on the Battlefield 27 (2nd ed. 2006).

do not include the subjective balancing requirement the law requires. Instead, targets displayed to commanders are accompanied by either a "green, yellow, [or] red [light], like a traffic signal."[47] The details of how Habsora determines when the red light turns to yellow or green are unknown to the public and apparently unknown to the IDF units relying on these outputs. However, the interface itself suggests a simplicity that belies the complicated analysis accompanying proportionality and encourages human deference to the algorithmic determination that the target is a "go" for a strike.

The "case-by-case" emphasis in assessing civilian "direct participation" and proportionality in attacks is deliberate and reflects the considered judgment of IHL that such legal principles are intended to comport with standards rather than rigid rules. Standards "rely on case-by-case decisionmaking and render a decision that is ostensibly limited to the facts of the case."[48] Standards broaden discretion and, consistent with the distinction and proportionality described above, do so based on the subjective assessment of a commander in the field's evaluation of the distinctive set of facts. As a result, standards "allow for the decrease of errors of underinclusiveness and overinclusiveness" and "allow the decisionmaker to take into account all relevant factors or the totality of the circumstances."[49]

In short, standards like "direct participation" and proportionality analysis exist because of the judgment that the amount of factual variance is near infinite, information is imperfect, and the error costs (both to commanders and civilians) demand each situation to be judged individually. Such circumstances are a poor fit for AI systems dependent upon enormous volumes of high-quality data and unable to absorb qualitative, subjective content such as command intentions that "cannot be automated with narrow [i.e., non-AGI] AI technology."[50] Worse, their use, especially when unaccompanied by an understanding of the variables upon which the system came to its output, is more likely to impair rather than augment the human judgment upon which these complex legal standards rely.

B. The Principle of Precaution and "Constant Care"

At its core, the precautionary principle is an obligation effectuated in planning. The obligation, reflected in customary law and Article 57 of Additional Protocol I, creates a general obligation applicable to all military operations and imposes specific requirements regarding attacks. Specifically, Article 57(1) imposes an affirmative obligation on States to take "constant care" in the "conduct of military operations" to spare civilians and civilian objects.[51]

47 Davies, McKernan & Sabbagh, *supra* note 1.
48 Edward Lee, *Rules and Standards for Cyberspace*, 77 Notre Dame L. Rev. 1275, 1295 (2002).
49 Kathleen M. Sullivan, *The Supreme Court, 1991 Term—Foreword: The Justices of Rules and Standards*, 106 Harv. L. Rev. 22, 58–9 (1992).
50 Avi Goldfarb & Jon R. Lindsay, *Prediction and Judgement: Why Artificial Intelligence Increases the Importance of Humans in War*, 46 Int'l Sec. no. 3, Winter 2021/22, at 9.
51 AP I, *supra* note 36, art. 57(1).

Article 57(2) imposes more specific precautions "with respect to attacks."[52] Under this provision, commanders must "do everything feasible to verify that the objectives to be attacked are neither civilians nor civilian objects" and to refrain from imposing civilian damage disproportionate to the anticipated military advantage.[53] Further, States are obligated to suspend any attack "if it becomes apparent that the objective is not a military one," provide "effective advance warning" of attacks when possible, and choose objectives causing lesser civilian damage when such choice is available "for obtaining similar military advantage."[54]

The inexplicability of black-box AI targeting systems poses fundamental problems under both the general duty of constant care of Article 57(1) and the precautions in attack obligations set out in Article 57(2).

First, the more specific precautionary obligations set out in Article 57(2) are dependent upon commanders understanding the basis of the targeting recommendation made by any AI-based system. For instance, the ability to verify a target as required by Article 57 can only occur if those tasked with verification are aware of the information giving rise to the targetability determination and how it was weighed. Similarly, awareness of the factors an AI model is basing its determination on might influence a commander's decision to engage in an alternative means of attack.

The broader "constant care" obligation imposed under Article 57(1) has implications for a State's obligations in attack as well as in its preceding design of targeting platforms. The duty of constant care requires States to possess "situational awareness at all times" in identifying and averting avoidable civilian harms associated with targeting decisions.[55] This includes understanding the variables giving rise to initial determinations of targetability (or non-targetability), the military advantages perceived, the civilian harm envisioned, *and* the recognition that such conditions are dynamic and that newly available information might alter the relevant legal calculus. In short, such situational awareness requires an understanding of the AI-based targeting model at a global level (how it operates generally) and at the local level (why it generated specific outputs)—neither of which is available via black-box models.

Moreover, the broad scope of the duty of constant care requires States to ensure explainability as a matter of design, not just at the time of the use of armed force. As the language suggests, the "constant care" obligation, unlike the distinction and proportionality requirements, extends beyond armed attacks and is "relevant in both

52 *Id*. art. 57(2).
53 *Id*.
54 AP I, *supra* note 36, art. 57(2), 57(3).
55 Eric Talbot Jensen, *Cyber Attacks: Proportionality and Precautions in Attack*, 89 Int'l L. Stud. 198, 202 (2013).

peacetime and *wartime*."[56] According to the ICRC, military operations covered under the duty include "any movements, manoeuvres and other activities whatsoever carried out by the armed forces with a view to combat."[57] As explained by Russell Buchan, an "operation possesses a military character where it is designed to advance combat."[58] As a result, there is reason to believe that the development and deployment of AI targeting systems would constitute "military operations" subject to the State's "constant care" obligations.[59] Just as the duty of constant care can be violated by a commander failing to weigh the relevant factors in executing an attack, States can violate the duty of constant care by designing and deploying a black-box targeting system that systematically precludes commanders from performing effective due diligence on suggested attacks.

4. BEYOND EXISTING LAW

There is consensus that existing legal principles are insufficient to govern the use of AI systems in armed conflict. Unfortunately, there is little agreement about the contours of new legal norms governing such systems. The ambiguity of what norms should govern AI systems includes a lack of clarity on the importance of avoiding black-box AI and the type of explainability that should be required. The Artificial Intelligence Strategy adopted by NATO identifies "explainability and traceability" as one of its principles and states that "AI applications will be appropriately understandable and transparent."[60] More recently, the *Political Declaration on Responsible Military Use of Artificial Intelligence* espouses a narrower understandability requirement and foregoes any reference to "explainability." Regarding development, the declaration states that military AI systems should be "developed with methodologies, data sources, design procedures and documentation that are transparent to and auditable by their relevant defense personnel."[61] While reprising the "transparency" language set out in NATO doctrine, the language only addresses the transparency of the constitutive components of the model rather than the model itself. In other words, personnel are to have awareness of (and presumably understand) the fundamental elements upon which a military AI system is based but such understanding will not, in and of itself,

56 Asaf Lubin, *Lieber Studies Big Data Volume—Algorithms of Care: Military AI, Digital Rights, and the Duty of Constant Care*, Articles of War (Feb. 13, 2024), https://lieber.westpoint.edu/algorithms-care-military-ai-digital-rights-duty-constant-care/; *see also* Jensen, *Cyber Attacks, supra* note 55, at 202; Russell Buchan, Data Protection in War (2023) (on file with author).

57 Int'l Comm. of the Red Cross (ICRC), Commentary on the Additional Protocols of 8 June 1977 to the Geneva Conventions of 12 August 1949, at ¶ 2191 (Yves Sandoz et al. eds., 1987).

58 Buchan, *supra* note 56, at *10.

59 There is substantial ambiguity (and thus debate) regarding where the duty of constant care begins and ends. *See e.g.*, Michael Schmitt, *Big Data: International Law Issues During Armed Conflict, in* Big Data and Armed Conflict (Laura A. Dickinson & Edward W. Berg eds., 2023). However, because inexplicability is persistently intrinsic to black-box AI models, their usage would necessarily involve contexts where the duty would undoubtedly apply.

60 Stanley-Lockman & Christie, *supra* note 30, at C.

61 *Political Declaration, supra* note 31.

guarantee any understanding of how the model is generating specific outputs or even much insight into its general operation.

The declaration also states that users should "be trained so they sufficiently understand the capabilities and limitations" of each system.[62] As with the language of "transparency," this requirement, while appropriate, does little to address the core black-box problem—the ignorance of how specific outputs are made. For example, knowing that an AI targeting platform sometimes might mistake a video camera for a weapon tells you little about whether that particular error might have occurred in any specific targeting decision.

Beyond legal obligations, an intractable concern attaches to black-box AI systems that counsel in favor of explainable alternatives: their resistance to adjustment and the hidden errors and biases their opacity may perpetuate. A limited understanding of how an AI system makes its prediction poses an inherent difficulty in identifying when it has made errors.[63] Such hidden errors would be especially challenging to detect in targeting circumstances where civilians and combatants are difficult to distinguish. When an AI system incorrectly identifies a civilian as a target, absent incontrovertible evidence to the contrary, the system user will likely characterize the subsequent strike as a success, creating a self-fulfilling prophecy problem, in which the reaction to the prediction effectively renders the prediction accurate. Just as problematically, black-box models are notoriously difficult to debug when inevitable prediction errors are discovered.[64] Imagine an AI system that targets an ambulance by mistake. Without knowing how the model came to its targeting conclusion, it is impossible to know how to ensure that such an error does not recur. The problem could relate to inappropriate weighting, missing parameters, or nuances within the original training data. The difficulty in detecting and remediating prediction errors is precisely why black-box models are considered especially vulnerable to security intrusions.[65]

An "appropriately" explainable and understood AI system depends on the domain and application in which it will be used. In the context of IHL, the explainability ought to be understood within the context of the aspects of judgment described above as embedded in the core principles of distinction, proportionality, and precaution discussed above and the balance between humanitarian interests and military necessity embedded in the regime. To that end, States need to emphasize interpretability as a necessary component of any AI-based targeting system—whether operating autonomously or not.

62 *Id*, at para. 6.
63 *See Carabantes, supra* note 10.
64 Thomas P. Quinn, Stephan Jacobs, Manisha Senadeera, Vuong Le & Simon Coghlan, *The Three Ghosts of Medical AI: Can the Black-Box Present Deliver?* 124 A.I. in Medic. 1, 3 (2022).
65 *Id*.

Model interpretability gauges the degree to which the user can understand how an AI system came to its prediction. Globally, interpretability ensures those using the AI system can verify that it performs as expected.[66] At this level, it is essential for users to understand the magnitude and direction of the input variables' impact on the final predicted outcome, enabling them to grasp, on average, the relationship between input data and predicted outcomes.[67] As to specific predictions, the user needs to understand how the model's individual features influence the final output and its predicted value.[68] This enables users to identify in detail how the value of each input combines to generate the predicted output for an individual (or single unit).

Model interpretability is also essential for understanding the mechanisms of the algorithm.[69] This becomes especially important when monitoring systems for bias. Breaking the algorithm into discrete steps illustrates the decision process.[70] This allows users to review each step the model took between input and final prediction. In reviewing the key steps, users can identify specific variables the model used to make decisions and any associated critical value it used to make the decision.

5. CONCLUSION

The growing capabilities of AI have provoked tremendous excitement and consternation. Nowhere are the stakes of this technology higher than in international security and armed conflict. However, amid the inevitable AI race gripping the world, it is crucial to remember that the legal rules underlying our interactions, both internationally and personally, were written from an unmistakably human perspective. The subjective standards and context-driven demands of IHL place human judgment at the center of authority. Relying on human subjectivity, these rules require AI systems that enable individuals to exercise their judgment in a manner consistent with existing legal constraints.

66 Franck Jaotombo, Luca Adorni, Badih Ghattas & Laurent Boyer, *Finding the Best Trade-off Between Performance and Interpretability in Predicting Hospital Length of Stay Using Structured and Unstructured Data*, 18 PLoS One, no. 11, 2023; Linardatos, Papastefanopoulos & Kotsiantis, *supra* note 25; Lundberg et al., *supra* note 25, at 56–67.

67 *Id*.

68 *Id*.

69 *Id*.; *see also* Linardatos, Papastefanopoulos & Kotsiantis, *supra* note 25.

70 *Id*.