MACHINE LEARNING WEAPONS AND INTERNATIONAL HUMANITARIAN LAW: RETHINKING MEANINGFUL HUMAN CONTROL

SHIN-SHIN HUA*

ABSTRACT

AI's revolutionizing of warfare has been compared to the advent of the nuclear bomb. Machine learning technology, in particular, is paving the way for future automation of life-or-death decisions in armed conflict.

But because these systems are constantly "learning," it is difficult to predict what they will do or understand why they do it. Many therefore argue that they should be prohibited under international humanitarian law (IHL) because they cannot be subject to meaningful human control.

But in a machine learning paradigm, human control may become unnecessary or even detrimental to IHL compliance. In order to leverage the potential of this technology to minimize casualties in conflict, an unthinking adherence to the principle of "the more control, the better" should be abandoned.

Instead, this Article seeks to define prophylactic measures that ensure machine learning weapons can comply with IHL rules. Further, it explains how the unique capabilities of machine learning weapons can facilitate a more robust application of the fundamental IHL principle of military necessity.

[*] Advanced Master of Laws (Leiden University, Netherlands), JD-equivalent (University of Cambridge); Research Affiliate, Centre for the Study of Existential Risk, University of Cambridge; Attorney, BT Group; former Attorney at Cleary Gottlieb Steen & Hamilton. With thanks to Horst Fischer, Professor of International Humanitarian Law at Leiden University and Adjunct Professor of International and Public Affairs at Columbia University, and Jacob Turner, barrister at Fountain Court Chambers in London, for their valuable guidance and feedback. All opinions are those of the author and do not reflect the views of any organization.

I. INTRODUCTION

Machine learning is the buzzword of our age. Instead of relying on pre- programming, these systems can "learn" how to do a task through training, use, and user feedback.[1] Having revolutionized fields from medicine to finance, machine learning is propelling a new artificial intelligence (AI) arms race among the world's major military powers to deploy these technologies in warfare.[2] Indeed, the rise of military AI has been compared to the advent of the nuclear bomb.[3]

[1] STEPHAN DE SPIEGELEIRE, MATTHIJS MAAS & TIM SWEIJS, ARTIFICIAL INTELLIGENCE AND THE FUTURE OF DEFENSE: STRATEGIC IMPLICATIONS FOR SMALL- AND MEDIUM-SIZED FORCE PROVIDERS 35–39 (2017).

[2] *America v China-The Battle for Digital Supremacy*, THE ECONOMIST (Mar. 15, 2018), https://www.economist.com/leaders/2018/03/15/the-battle-for-digital-supremacy; Karla Lant, *China, Russia and the US Are in an Artificial Intelligence Arms Race*, FUTURISM (Sept. 12, 2017), https://futurism.com/china-russia-and-the-us-are-in-an-artificial-intelligence-arms-race.

[3] Tom Simonite, *AI Could Revolutionize War as Much as Nukes*, WIRED (July 19, 2017), https://www.wired.com/story/ai-could-revolutionize-war-as-much-as-nukes/.

But machine learning technology challenges human control as a core tenet of International Humanitarian Law (IHL). IHL is the body of law that governs the conduct of belligerents in armed conflict, seeking to balance the necessity of weakening the adversary with the desire to minimize unnecessary suffering.[4] The requirement of a human operator to control the effects of weapons is an idea deeply embedded in IHL. The International Court of Justice, for example, stated in its *Nuclear Weapons Advisory Opinion* that nuclear weapons are by their nature "scarcely reconcilable" with IHL rules prohibiting unnecessary suffering and indiscriminate harm. [5] This is due to the inability to contain their destructive force "in either space or time."[6]

An Autonomous Weapons System with machine learning capabilities ("Learning AWS") may break this paradigm. While life-or-death decisions on the battlefield currently remain firmly within the control of human operators, the future automation of these decisions cannot be ruled out.[7] Machine learning systems are also developing a unique ability to adapt to uncertainties in their environment and to make complex decisions based on large volumes of data. This makes them potential candidates for replacing humans in selection of and engagement with military targets.[8]

However, a future Learning AWS's ability to "learn" from its environment would also make its behavior difficult to predict (*i.e.*, how a new input will be processed) and difficult to understand (*i.e.*, why a decision was made).[9] The question is whether a human can still be deemed to "control" a Learning AWS with unforeseeable behavior and opaque decision-making processes.[10]

Some scholars argue that if AWSs are unpredictable and inscrutable, humans cannot meaningfully control them.[11] It follows that these AWSs would be unlawful under the IHL doctrine of "meaningful human control."[12] Much of the present scholarship, however, focuses

[4] *See* discussion *infra* Section III.B.

[5] Legality of the Threat or Use of Nuclear Weapons, Advisory Opinion, 1996 I.C.J. Rep. 226, ¶¶ 92, 95 (July 8).

[6] *See, e.g.*, *id.* ¶¶ 35, 92, 95.

[7] *See* discussion *infra* Sections II.B–II.C.

[8] *See* discussion *infra* Section II.B.

[9] *See* discussion *infra* Section II.B.

[10] *See generally* Brent Daniel Mittelstadt et al., *The Ethics of Algorithms: Mapping the Debate*, 3 BIG DATA & SOC'Y 1, 11 (2016).

[11] *See, e.g.*, Andreas Matthias, *The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata*, 6 ETHICS & INFO. TECH. 175 (2004).

[12] *See infra* Section III.D (discussing the doctrine of "meaningful human control").

on the *absence* of human control over lethal autonomous weapons.[13] Less attention has been paid to the implications of machine learning as a *substitute* for human control in the targeting process.[14] This failure to consider the use of machine learning in future AWSs has led scholars to overlook the more basic question: do machine learning systems require human control in order to comply with IHL?[15]

This Article will first provide a technological overview of Learning AWSs. It will explain how such a system's ability to constantly "learn" and adapt from experience leads to highly unpredictable outcomes and inscrutable decision-making processes. This makes it difficult for any human to meaningfully control their use. At the same time, they hold great potential to enhance compliance with IHL due to their ability to process large volumes of data at speed and to use such data to make nuanced, strategic decisions.

The Article next discusses the precautionary obligation to take constant care under IHL.[16] It considers the point at which the lack of human control over a machine learning weapon may breach this obligation, and concludes that it is far from clear that the law requires any minimum level of human control over a Learning AWS. Machine learning technologies may render human control unnecessary or even detrimental to a Learning AWS's ability to comply with IHL. A blanket requirement of *ex ante* human approval before each attack[17] or of the possibility of human override at any time[18] should therefore be reconsidered. Otherwise we risk losing a potential future tool for minimizing civilian casualties in armed conflict.

[13] *Id. But see infra* Section III.D.3.d (discussing Schuller's alternative theory).

[14] *See infra* Section III.D (discussing the doctrine of "meaningful human control"). *But see* Ashley Deeks, Noam Lubell & Daragh Murray, *Machine Learning, Artificial Intelligence, and the Use of Force by States*, 10 J. NAT'L SECURITY L. & POL'Y 1 (2019); Matthias, *supra* note 11; Alan L. Schuller, *At the Crossroads of Control: The Intersection of Artificial Intelligence in Autonomous Weapon Systems with International Humanitarian Law*, 8 HARV. NAT'L SECURITY J. 379 (2017).

[15] *But see infra* Section III.D.3.d (discussing Schuller's approach).

[16] Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (Protocol I), art. 57, June 8 1977, 1125 U.N.T.S. 3 (1977), https://ihl-databases.icrc.org/applic/ihl/ihl.nsf/7c4d08d9b287a42141256739003e636b/f6c8b9fee14a77fdc125641e0052b079 [hereinafter Additional Protocol I].

[17] HEATHER M. ROFF & RICHARD MOYES, "MEANINGFUL HUMAN CONTROL, ARTIFICIAL INTELLIGENCE AND AUTONOMOUS WEAPONS": BRIEFING PAPER PREPARED FOR THE INFORMAL MEETING OF EXPERTS ON LETHAL AUTONOMOUS WEAPONS SYSTEMS, 4–5 (April 2016), http://www.article36.org/wp-content/uploads/2016/04/MHC-AI-and-AWS-FINAL.pdf.

[18] Peter Margulies, *Making Autonomous Weapons Accountable: Command Responsibility for Computer-Lethal Force in Armed Conflicts*, *in* RESEARCH HANDBOOK ON REMOTE WARFARE 405–42, 433–34 (Jens David Ohlin ed., 2017).

In this light, the Article endorses the legal standard proposed by Alan Schuller, which asks whether it is reasonably predictable that a Learning AWS will comply with IHL.[19] If this test is satisfied, there is no further requirement of human control during the system's deployment. Schuller proposes, further, a number of prophylactic measures to help ensure development of lawful Learning AWSs in the first place.

Whilst Schuller's theory better accommodates the unique characteristics of machine learning technology, it is incomplete in two key respects, which this Article seeks to address. First, the Article offers suggestions as to how Schuller's prophylactic measures can better accommodate unpredictability that arises not only between the Learning AWS and its operating environment, but also between the Learning AWS and its human controller. Second, the Article argues that Schuller's standard of reasonably predictable IHL compliance does not go far enough. It fails to recognize that many machine learning systems can be programmed to optimize the probability of achieving a certain goal. In the future, it may be possible to program a Learning AWS to optimize the objective of minimizing civilian harm. The fundamental IHL principle of military necessity therefore dictates that a Learning AWS should be programmed to comply with IHL not only to a reasonable level of predictability, but also to an optimal level of predictability.

In conclusion, this Article urges a fresh approach that moves away from the idea that if a Learning AWS cannot be meaningfully controlled by a human, it cannot comply with IHL. A more nuanced approach is required in order to realize the opportunities for machine learning technology to more robustly apply the rules of IHL.

II. Overview of the Technology

A. *Defining Autonomous Weapons Systems*

There is no commonly accepted definition of an AWS.[20] Narrower definitions describe AWSs that have the ability to autonomously use lethal force or to carry out "critical functions."[21] In comparison, a 2016 United Nations meeting of governmental experts defined an AWS more broadly as "weapons systems that are capable of carrying out tasks governed by IHL in partial or full replacement of a human in the use of

[19] *See infra* Section III.D.3.d (discussion of Schuller).

[20] *See, e.g.*, Int'l Comm. of the Red Cross, Autonomous Weapon Systems: Technical, Military, Legal and Humanitarian Aspects 7 (2014), https://www.icrc.org/en/document/report-icrc-meeting-autonomous-weapon-systems-26-28-march-2014 [hereinafter ICRC Report].

[21] *Id.*

force, notably in the targeting cycle.'[22] This definition has two key elements: (1) the balance of human-machine control ("partial or full replacement of a human"); and (2) the function being carried out by the AWS ("tasks governed by IHL . . . notably in the targeting cycle"). These two aspects are defined in further detail below.

1. Human-Machine Interactions

Autonomous systems can be categorized in the following ways according to the distribution of control between human and machine:

- "Human-in-the-loop" weapons: systems select targets and use force only via human command.
- "Human-on-the-loop" weapons: systems select targets and use force under human supervision. The human operator can override the system.
- "Human-out-of-the loop" weapons: systems select targets and use force with no human input or control.[23]

For the purposes of this analysis, references to AWSs include "human-on-the-loop" and "human-out-of-the-loop" weapons. The latter category is particularly relevant to discussions around the minimum level of human control required under IHL.

2. The Task Performed

Discussions of autonomy should also consider the specific tasks to be carried out by the AWS. A useful analytical framework is the "OODA Loop," which sees decision-making as a continuous process with four stages: Observe, Orient, Decide, and Act.[24]

In warfare, machines have long been used to carry out the "Observe" stage of the targeting process.[25] The use of machine sensors to observe

[22] Rep. of Switzerland to Convention of Certain Conventional Weapons Meeting of Experts, Towards a "Compliance-Based" Approach to LAWS 1 (Mar. 30, 2016) (working paper), https://www.unog.ch/80256EDD006B8954/(httpAssets)/D2D66A9C427958D6C1257F8700415473/$file/2016_LAWS+MX_CountryPaper+Switzerland.pdf.

[23] *See* HUMAN RIGHTS WATCH , MIND THE GAP: THE LACK OF ACCOUNTABILITY FOR KILLER ROBOTS (2015), https://www.hrw.org/report/2015/04/09/mind-gap/lack-accountability-killer-robots.

[24] William C. Marra & Sonia K. McNeil, *Understanding "The Loop": Regulating the Next Generation of War Machines*, 36 HARV. J.L. & PUB. POL'Y 1139, 1145 (2013).

[25] VINCENT BOULANIN & MAAIKE VERBRUGGEN, MAPPING THE DEVELOPMENT OF AUTONOMY IN WEAPON SYSTEMS 27–29 (2017), https://www.sipri.org/sites/default/files/2017-11/siprireport_mapping_the_development_of_autonomy_in_weapon_systems_1117_1.pdf.

military operations and detect potential military targets is generally uncontroversial from an IHL perspective.[26] Similarly, once a human operator has decided on the military target and how lethal force is delivered (e.g. choice of weapon, timing, ability to loiter) the delivery of lethal force itself has already been widely automated through use of remote warfare.[27] However, machine learning technologies open up the possibility that AWSs might also be used to carry out also the "Orient" and "Decide" stages of the targeting process.

In the "Orient" phase, the AWS autonomously reviews "[c]urrent intelligence estimates, sensor collection and battlefield reports. . .[and] the tactical and strategic implications [are] weighed, as are countless other military and non-military considerations."[28] Usually, the experience of a human commander will play a key role in identifying and weighing up the potential courses of action at this stage.[29] However, in this Article's hypothesis, the Learning AWS would use its machine learning functionality to identify potential courses of action at the "Orient" stage. Finally, it would use machine learning to determine the best course of action at the subsequent "Decide" stage of the OODA loop. The "Decide" stage constitutes the final deliberative step in the decision-making process and ultimately results in the delivery of lethal force in the "Act" stage.[30]

At the "Orient" and "Decide" stages of the targeting cycle, the delegation of discretionary, value-laden judgments to the machine dilutes the causal link between a human's decision to kill and the delivery of lethal force.[31] It is therefore this delegation of *discretionary* decision-making in the targeting cycle that gives rise to new questions under IHL and is the focus of this analysis.

B. *Machine Learning*

A machine's control system governs its decision-making process. Control systems can be categorized based on their capacity to govern their own behavior and deal with environmental uncertainties.[32]

Automatic systems, for example, rely on a series of pre-programmed "if-then rules" which prescribe how the system should react to a given

[26] Schuller, *supra* note 14, at 394.

[27] BOULANIN & VERBRUGGEN, *supra* note 25, at 47–49.

[28] Schuller, *supra* note 14, at 394.

[29] *Id.*

[30] *Id.* at 396–97.

[31] *Id.* at 394–97.

[32] BOULANIN & VERBRUGGEN, *supra* note 25, at 6.

input.[33] Automatic systems have no ability to handle uncertainties in their operating environment.[34]

Learning systems are a more sophisticated form of control system that can improve their performance over time through experience. A key advantage of learning systems over automatic systems is that they do not require a human to specifically define the problem or solution.[35] Instead, learning systems can "learn" by extracting statistical relationships or patterns from data. The knowledge gained is then used to automatically improve the performance of the system through changing its structure, program, or data.[36]

If a future AWS is to replace discretionary human decision-making in the "Orient" and "Decide" steps of a targeting process, as discussed above, it must be able to carry out nuanced decision-making that takes into account the uncertainties of the battlefield. This requires that it have the ability—characteristic of learning systems—to improve its performance over time through interactions with its surroundings.[37]

The following analysis considers two subtypes of machine learning in order to isolate attributes of these technologies that may be relevant under IHL: deep learning and reinforcement learning. For reasons discussed below, deep and reinforcement learning seem the most likely replacement for humans in carrying out discretionary decision-making in the targeting cycle.

1. Deep Learning

Deep learning is a type of representation learning method, which denotes systems that can "learn how to learn."[38] These systems can work off raw data, extracting representations (features) that are useful to their specific machine learning tasks.[39] They do this through deep neural networks, which are networks of hardware and software that are inspired by the human brain.[40]

The key advantage of deep learning compared to older types of machine learning is that it does not require manual feature engineering, which involves the refinement of each raw dataset before it can be

[33] *Id.* at 9–11.

[34] *Id.Id.* at 6.

[35] *Id.* at 16–17.

[36] *Id.*

[37] *Id.* at 113–14.

[38] *Id.* at 17.

[39] *Id.*

[40] BOULANIN & VERBRUGGEN, *supra* note 25, at 17.

processed by a machine learning system.[41] A further benefit is that deep learning systems can make distinctions that a human trainer would be unable to represent through algorithms.[42] This makes deep learning attractive for military deployment, as it can accurately and efficiently interpret intelligence data. The diminished involvement of a human programmer means, however, that it can be difficult to understand how a deep learning system makes its decisions.[43]

2. Reinforcement Learning

Reinforcement learning technology merges the training and application phases of a machine system, which are distinct in traditional neural networks. A reinforcement learning system trains within its operating environment by pursuing various alternative action routes in a trial-and-error fashion, using the results to continuously hone its own parameters.[44] A machine that can learn "on the job" is far better at adapting to uncertain surroundings.[45]

A recent example of reinforcement learning is AlphaGo Zero, a system designed by the AI company DeepMind. AlphaGo Zero was trained to play Go, a game considered far more difficult than chess for machines to master due to the enormous number of possible moves.[46] While its predecessor AlphaGo first trained on thousands of human amateur and professional games, AlphaGo Zero was able to skip this step and learn simply by playing games against itself. In doing so, it swiftly and dramatically exceeded human playing capabilities.[47]

AlphaGo Zero demonstrated the great potential of reinforcement learning for use in future AWSs. First, reinforcement learning has the potential to greatly surpass human abilities in carrying out the kind of complex problem-solving required to wage war.[48] Second, reinforcement learning systems can generate novel solutions unconstrained by

[41] *Id.*

[42] Matthias, *supra* note 11, at 179.

[43] Mittelstadt et al., *supra* note 10, at 6.

[44] Matthias, *supra* note 11, at 179.

[45] *Id.*

[46] David Silver et al., *AlphaZero: Shedding new light on chess, shogi, and Go*, DEEPMIND BLOG (Dec. 6, 2018), https://deepmind.com/blog/article/alphazero-shedding-new-light-grand-games-chess-shogi-and-go, (last visited Aug. 19, 2019).

[47] David Silver & Demis Hassabis, *AlphaGo Zero: Starting from Scratch*, DEEPMIND BLOG (Oct. 18, 2017), https://deepmind.com/blog/alphago-zero-learning-scratch/.

[48] *See* Elsa B. Kania, *Quest for an AI Revolution in Warfare*, THE STRATEGY BRIDGE (June 8, 2017), https://thestrategybridge.org/the-bridge/2017/6/8/-chinas-quest-for-an-ai-revolution-in-warfare.

human capacities and preconceptions.[49] At the same time, this makes the behavior of these machines both unpredictable (difficult to anticipate beforehand) and inscrutable (difficult to explain afterwards).[50]

3. Legally Relevant Attributes of Machine Learning Systems

From the above overview of machine learning technologies, it is possible to distill a number of legally relevant attributes. A Learning AWS's decision-making process may be inscrutable to a human controller, making human supervision difficult, especially in time-critical combat scenarios. This inherent uncontrollability arguably renders Learning AWSs unlawful under IHL, for example under the precautionary obligation to take constant care to spare the civilian population, civilians, and civilian objects.[51]

Furthermore, as Learning AWSs can adapt and "learn" from their experiences, even programmers and developers may find it difficult to predict how they will eventually behave.[52] In the case of a malfunction leading to a breach of IHL, this could make it difficult to find the mens rea required to establish individual liability under international criminal law (ICL).[53]

Despite these compelling concerns, machine learning systems can process vast volumes of intelligence data at speeds far surpassing human capabilities.[54] In addition, this Article argues that techniques such as reinforcement learning may make it possible for AWSs to carry out complex, strategic decision-making on a future battlefield.[55] Both of these attributes could facilitate targeting decisions that *improve* compliance with IHL, for example by minimizing civilian harm.

The challenge for IHL is to harness the potential of Learning AWSs to minimize civilian harm in armed conflict, while prohibiting Learning AWSs that are dangerously unpredictable or inscrutable.

C. *Current Military Uses of Machine Learning Systems*

Machine learning is currently used in a variety of military applications. One example is the use of deep learning in developing precision

[49] *See* Silver & Hassabis, *supra* note 47.

[50] Mittelstadt et al., *supra* note 10, at 3–4.

[51] Additional Protocol I, *supra* note 16, art. 57.

[52] Will Knight, *The Dark Secret Heart at the Heart of AI*, MIT TECH. REVIEW (2017), https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/ (last visited Jan. 18, 2019).

[53] *See* discussion *infra* Section III.A–III.B.

[54] *See generally* CHRISTOPHER M. BISHOP, PATTERN RECOGNITION AND MACHINE LEARNING (2006).

[55] Kania, *supra* note 48.

automatic target recognition (ATR) systems.[56] Machine learning plays a key role in ATR systems by minimizing false-alarm rates in complex environments, for example by ensuring that the ATR system is not distracted by decoys or mistakes.[57]

Presently, deep learning in ATR targeting systems only acts as a decision aid to human operators.[58] Significant technical obstacles must still be overcome before achieving fully autonomous targeting using a Learning AWS. The main difficulty is that designing a machine learning system which can handle all possible eventualities, even in relatively stable environments, could require "impossibly" large data sets.[59]

Nevertheless, given the rapid pace of development of machine learning technologies and the clear intention of some of the world's major military powers to implement these technologies for military uses,[60] this Article will not confine itself to current applications of machine learning technology in warfare. Instead, it will consider potential *future* uses of machine learning technology in armed conflict, focusing on the automation of discretionary decision-making in the targeting cycle.

III. Are Learning Autonomous Weapons Systems "Unlawfully Autonomous"?

A. *Unpredictability and IHL*

"Automatic" machine systems that follow simple "if *X*, then *Y*" rules are predictable: their programming governs how these systems will respond to environmental input.[61] Machine learning systems, on the other hand, are, "by definition, unpredictable," because they are constantly "learning" and adapting to their surroundings.[62]

The unpredictability of a Learning AWS gives rise to a problematic future scenario under IHL. Take a Learning AWS that meets all of the requirements of IHL when functioning *properly*. Even there, a

[56] *See, e.g.*, Pat Host, *Deep Learning Analytics Develops DARPA Deep Machine Learning Prototype*, DEFENSE DAILY (Nov. 5, 2016), https://www.defensedaily.com/deep-learning-analytics-develops-darpa-deep-machine-learning-prototype/advanced-transformational-technology/.

[57] SPIEGELEIRE, MAAS, & SWEIJS, *supra* note 1, at 88–89.

[58] BOULANIN & VERBRUGGEN, *supra* note 25, at 25–26.

[59] Schuller, *supra* note 14, at 410; *see also* BOULANIN & VERBRUGGEN, *supra* note 25, at 65–82.

[60] *See supra* Section I.

[61] PAUL SCHARRE, AUTONOMOUS WEAPONS AND OPERATIONAL RISK 12 (Feb. 2016), https://s3.amazonaws.com/files.cnas.org/documents/CNAS_Autonomous-weapons-operational-risk.pdf?mtime=20160906080515.

[62] INT'L COMM. OF THE RED CROSS, AUTONOMOUS WEAPON SYSTEMS: IMPLICATIONS OF INCREASING AUTONOMY IN THE CRITICAL FUNCTIONS OF WEAPONS 13 (2016), https://www.icrc.org/en/publication/4283-autonomous-weapons-systems.

malfunction may still lead to civilian harm. In such a scenario, no human could be held responsible because it was not foreseeable that the Learning AWS would "fail" in this particular way.[63] The unpredictability of Learning AWSs thus presents a potential obstacle to establishing liability under ICL, the system of law that holds individuals responsible for serious violations of IHL.[64] Under ICL, combatants are generally responsible for the reasonably foreseeable consequences of their actions.[65]

Given the potential difficulty of establishing responsibility *ex post*, this Article turns its attention instead to possible measures to prevent violations of IHL from occurring *ex ante*. In particular, it will explore what minimum level of human supervision over a Learning AWS is required by IHL precautionary obligations.

B. *Precautionary Obligations Under IHL*

At its core, IHL balances several contradictory fundamental principles. The fundamental principles of IHL guide the conduct of belligerents at all times.[66] They are general in nature, and inform and underpin the specific treaty rules that apply them.[67] While seeking to mitigate the effects of conflict, according to the principle of humanity, IHL also recognizes that belligerents must be permitted to weaken their enemy, according to the principle of military necessity.[68] It is therefore lawful under the fundamental principle of distinction to launch attacks that pursue a valid military purpose, as long as any collateral damage to those who are not, or no longer, participating in

[63] MICHAEL HOROWITZ, PAUL SCHARRE & CENTER FOR A NEW AMERICAN SECURITY, MEANINGFUL HUMAN CONTROL IN WEAPON SYSTEMS: A PRIMER 7–8 (2015), http://www.cnas.org/sites/default/files/publications-pdf/Ethical_Autonomy_Working_Paper_031315.pdf.

[64] ICL is just one way in which IHL is enforced. Generally, ICL prosecutions are reserved for the "most serious crimes of concern to the international community." Rome Statute of the International Criminal Court preamble, art. 25(2), July 17, 1998, 2187 U.N.T.S. 90 (entered into force July 1, 2002). Therefore, not all violations of IHL automatically constitute international crimes.

[65] *See, e.g.*, *id.* art. 30 (explaining that (1) in order to establish a crime under the Rome Statute, the requisite mens rea must be present and (2) the general mens rea standard, short of intent, is knowledge i.*e.*, "*awareness that a circumstance exists or a consequence will occur in the ordinary course of events*").

[66] *See* Legality of the Threat or Use of Nuclear Weapons, Advisory Opinion, 1996 I.C.J. 226, 261, 493 (July 8) (dissenting opinion of Weeramantry J.).

[67] *Id.*

[68] *See, e.g.*, Hague Convention (IV) Respecting the Laws and Customs of War on Land and Its Annex: Regulations Concerning the Laws and Customs of War on Land, preamble, 26 Stat. 2277 ("the desire to diminish the evils of war, as far as military requirements permit").

hostilities is not excessive in relation to the expected gain in military advantage (the principle of proportionality).[69]

Precautionary obligations provide the practical means for belligerents to apply the fundamental IHL principles of humanity, distinction, military necessity, and proportionality. For example, the principle of distinction prohibits direct attacks against civilians.[70] The precautionary obligation to take constant care to spare the civilian population, civilians, and civilian objects therefore prescribes a practical application of the principle of distinction.[71]

The following analysis focuses on the duty to take "constant care"[72] as the core obligation upon which the other, more specific precautionary obligations are based and which they "materialize."[73]

C. *Autonomous Weapons Systems and the Duty of Constant Care*

The law provides little guidance on exactly what the constant care standard requires.[74] The term "constant care" is not defined under IHL and the ICRC's Commentary on the Additional Protocols of 8 June 1977 to the Geneva Conventions of 12 August 1949 simply refers to it as a "general principle."[75] It is said to apply to all domains of warfare and all levels of operation.[76] While this defines the scope of the obligation, it does little to explain its content.

[69] *See, e.g.*, the principle of distinction as set out in Article 48 and 52 of Additional Protocol I, *supra* note 16 (defining who is a combatant and a military object that may be permissibly attacked under IHL)*; see also* the principle of proportionality as embodied *inter alia* in Article 51(5)(b), 57(2)(a)(iii) and 57(2)(b) of Additional Protocol I, *supra* note 16.

[70] *Id.* art. 48, art. 52.

[71] Additional Protocol I, *supra* note 16, art. 57(1); JEAN-MARIE HENCKAERTS & LOUISE DOSWALD-BECK, INT'L COMM. OF THE RED CROSS, CUSTOMARY INTERNATIONAL HUMANITARIAN LAW: VOLUME I: RULES 51 (2005); CLAUDE PILLOUD ET AL., INT'L COMM. OF THE RED CROSS, COMMENTARY ON THE ADDITIONAL PROTOCOLS OF 8 JUNE 1977 TO THE GENEVA CONVENTIONS OF 12 AUGUST 1949 ¶ 2191 (1987)

[72] Additional Protocol I, *supra* note 16, art. 57(1).

[73] THEO BOUTRUCHE, EXPERT OPINION ON THE MEANING AND SCOPE OF FEASIBLE PRECAUTIONS UNDER INTERNATIONAL HUMANITARIAN LAW AND RELATED ASSESSMENT OF THE CONDUCT OF THE PARTIES TO THE GAZA CONFLICT IN THE CONTEXT OF THE OPERATION "PROTECTIVE EDGE" 8 (2015), https://www.diakonia.se/en/IHL/News-List/eo-on-protective-edge/.

[74] TERRY GILL ET AL., ILA STUDY GROUP 'THE CONDUCT OF HOSTILITIES AND INTERNATIONAL HUMANITARIAN LAW: CHALLENGES OF 21ST CENTURY WARFARE' - INTERIM REPORT 15 (2014), https://pure.uva.nl/ws/files/2346971/157905_443635.pdf.

[75] COMMENTARY ON THE ADDITIONAL PROTOCOLS OF 8 JUNE 1977 TO THE GENEVA CONVENTIONS OF 12 AUGUST 1949, *supra* note 71, ¶ 2191.

[76] PROGRAM ON HUMANITARIAN POLICY AND CONFLICT RESOLUTION, COMMENTARY TO THE HPCR MANUAL ON INTERNATIONAL LAW APPLICABLE TO AIR AND MISSILE WARFARE 124–125 (2010).

The report of a fact-finding committee established by the International Criminal Tribunal for the Former Yugoslavia (ICTY) to review the NATO bombing campaign against the Federal Republic of Yugoslavia[77] attempted to provide some clarification. The 2000 report confirmed that the obligation was one of feasibility only ("[t]he obligation to do everything feasible is high but not absolute") and that commanders enjoy "some range of discretion to determine which available resources shall be used and how they shall be used."[78]

Yet this quote strikes at the core of the definitional problem. The precautionary obligation to take constant care is couched in terms of what is *subjectively* practicable or feasible based on what a reasonable commander would do under the circumstances.[79] What is "feasible" requires a careful balancing of humanitarian and military considerations.[80] The highly subjective, judgment-laden nature of this concept makes it difficult to define with any precision.

The definitional fuzziness is compounded by new technologies in relation to which there is little jurisprudence or state practice to guide the practical application of legal standards. In response to similar difficulties around cyber warfare, an academic document called the Tallinn Manual on the International Law Applicable to Cyber Operations sought to clarify how international law applies to cyber warfare.[81] The Tallinn Manual, first published in 2012[82] by an initiative of the NATO Cooperative Cyber Defense Centre of Excellence, aims to objectively restate existing law according to groups of international legal experts.[83] It sheds some light on how the duty to take constant care might apply to Learning AWSs.

Unfortunately, the Manual is vague. Its Commentary to Rule 114 (Constant Care) clarifies that the precautionary obligation requires

[77] Int'l Crim. Trib. for the Former Yugoslavia, Final Report to the Prosecutor by the Committee Established to Review the NATO Bombing Campaign Against the Federal Republic of Yugoslavia, 39 I.L.M. 1257 (June 13, 2000).

[78] *Id.* ¶ 29.

[79] Michael N. Schmitt, *Autonomous Weapon Systems and International Humanitarian Law: A Reply to the Critics*, HARV. NAT'L SECURITY J. FEATURES 1, 20 (2013); Markus Wagner, *The Dehumanization of International Humanitarian Law: Legal, Ethical, and Political Implications of Autonomous Weapon Systems*, 47 VAND. J. TRANSNAT'L L. 1371, 1397 (2014).

[80] *See, e.g.*, CCW Protocol (III) on Prohibitions or Restrictions on the Use of Incendiary Weapons, art. 1(5), Oct. 10, 1980, 1342 U.N.T.S. 71 (entered into force Dec. 2, 1983).

[81] TALLINN MANUAL 2.0 ON THE INTERNATIONAL LAW APPLICABLE TO CYBER OPERATIONS (Michael N. Schmitt ed., 2d ed. 2017) [hereinafter TALLINN MANUAL 2.0].

[82] TALLINN MANUAL ON THE INTERNATIONAL LAW APPLICABLE TO CYBER WARFARE (Michael Schmitt ed., 2013).

[83] TALLINN MANUAL 2.0, *supra* note 81, at 1–12.

commanders and all others involved in cyber operations to be continuously sensitive to the effects of their activities.[84] Further, this duty extends throughout the operations, including in planning and operational processes.[85]

However, these relatively general principles fail to sufficiently address the specificities of cyber technology. For example, the Tallinn Manual states that "[g]iven the complexity of cyber operations . . . mission planners should, where feasible, have technical experts available to assist them in determining whether appropriate precautionary measures have been taken."[86]

While this principle can be extrapolated to Learning AWSs, the guidance is far from illuminating. Obviously, with highly complex technologies such as AWSs, military commanders must seek the advice of technical experts. This basic guidance fails to clarify whether the duty to take constant care requires some minimum level of human control over a Learning AWS.

In light of these uncertainties, the following analysis explores whether the academic doctrine of "meaningful human control" (MHC) can usefully flesh out the duty to take constant care when applied to AWSs with learning capabilities.

D. *Meaningful Human Control and the Duty of Constant Care*

1. The "Meaningful Human Control" (MHC) Doctrine

The term "meaningful human control" was coined by Article 36, an NGO, in its 2013 report on the United Kingdom's approach to AWSs.[87] It is an academic concept that is not part of existing IHL, and for which there is no agreed-upon definition.

At its broadest, the MHC doctrine contains a number of elements to ensure that an AWS is lawful. These include: (1) predictable, reliable and transparent technology; (2) accurate information on the outcome sought and on the context of use; (3) timely human action and potential for timely human intervention; and (4) the ability to attribute legal responsibility for outcomes.[88]

[84] *Id.* at 477.

[85] *Id.*

[86] *Id.*

[87] ARTICLE 36, KILLER ROBOTS: UK GOVERNMENT POLICY ON FULLY AUTONOMOUS WEAPONS, POLICY PAPER (Apr. 2013), http://www.article36.org/wp-content/uploads/2013/04/Policy_Paper1.pdf.

[88] *See, e.g.*, ROFF & MOYES, *supra* note 17.

Two schools of thought exist on the doctrine's relationship with IHL. The maximalist position is that MHC forms a separate and additional principle alongside the other fundamental principles of IHL. Under the minimalist approach, MHC is not a standalone requirement under IHL. Instead, it is a guiding principle for the design and usage of weapons systems in order to facilitate their compliance with IHL.[89] This author prefers the latter approach because it acknowledges the reality that states today may be reluctant to agree to new, binding legal rules.[90]

The following discussion focuses on the third requirement of human intervention. It will address David Akerson's argument that the duty to take constant care is infringed where human input is replaced with robotic autonomy because the latter "represents a break between the [military] force owing the duty of care and its ability to exercise that duty."[91]

2. The Requirement of *Ex Ante* Human Approval

Military theorists distinguish between three different levels of action. The strategic level is where a nation or group of nations define their military objectives.[92] The operational level of war implements the plans set at the strategic level, *e.g.*, by designating the time, space, and purpose under which troops are deployed.[93] Finally, the tactical level deals with how individual battles and engagements are fought.[94]

In their briefing paper for the U.N. Convention on Certain Conventional Weapons, Heather Roff and Richard Moyes propose a standard requiring human control over attacks at least down to the tactical level of warfare in addition to the operational and strategic levels.[95] In other words, an AWS would be precluded from autonomously moving from one attack to another without *ex ante* "human legal judgments" applied to each attack.[96]

[89] HOROWITZ, SCHARRE, & CENTER FOR A NEW AMERICAN SECURITY, *supra* note 63, at 7.

[90] Nehal Bhuta, Susanne Beck & Robin Geiss, *Present Futures: Concluding Reflections and Open Questions on Autonomous Weapons Systems*, *in* AUTONOMOUS WEAPONS SYSTEMS LAW, ETHICS, POLICY 347, 375 (Nehal Bhuta et al. eds., 2016).

[91] David Akerson, *The Illegality of Offensive Lethal Autonomy*, *in* INTERNATIONAL HUMANITARIAN LAW AND THE CHANGING TECHNOLOGY OF WAR 65, 87 (Dan Saxon ed., 2013).

[92] *Volume 1: Basic Doctrine, Levels of War*, CURTIS E. LEMAY CENTER, https://www.doctrine.af.mil/Portals/61/documents/Volume_1/V1-D34-Levels-of-War.pdf (last visited Feb. 15, 2019).

[93] *Id.*

[94] *Id.*

[95] ROFF & MOYES, *supra* note 17 at 4–5.

[96] *Id.* at 5.

Roff and Moyes's concern is that applying MHC only at the higher levels of warfare (*i.e.*, at the strategic and/or operational levels) could progressively dilute the quality of the legal and operational judgments reached.[97] This objection rests on the idea that greater physical distance between decision-makers and the battlefield could lead to poorer contextual awareness, for example of the geographic space and time in which the AWS would be used.[98] This lack of proximity between decision-makers and the battlefield reality could reach a point where "[the] ability to predict outcomes becomes either non-existent or minimal."[99]

3. Learning Autonomous Weapons Systems May Comply Better With IHL Without *Ex Ante* Human Approval

For the reasons discussed below, delegating legal and operational decisions to a Learning AWS could improve—rather than dilute—the quality of these decisions.

a. Big Data

Big data is a key component of military decision-making today.[100] Intelligence data can inform each of the steps in the targeting cycle, which typically consist of: (1) setting objectives; (2) developing and prioritizing targets; (3) analyzing capabilities; (4) assigning forces; (5) mission planning and execution; and (6) assessment.[101] Each of these steps contains its own feedback loop and attendant time lags, which is exacerbated by the need to process intelligence data.[102] Indeed, the amount of data available to inform targeting decisions can overwhelm human analysts.[103]

But where functions in the targeting cycle are delegated to machine learning systems, legal and operational judgments could improve as they are continuously and seamlessly updated according to realities on the ground. This is because learning systems, by their nature, use the knowledge gained through experience to automatically improve the performance of the system through changing its structure, program or

[97] *Id.*

[98] *Id.*

[99] *Id.*

[100] Kimberly Trapp, *Great Resources Mean Great Responsibility: A Framework of Analysis for Assessing Compliance with API Obligations in the Information Age*, *in* INT'L HUMANITARIAN LAW AND THE CHANGING TECH. OF WAR 159–60 (Dan Saxon ed., 2013).

[101] SPIEGELEIRE, MAAS & SWEIJS, *supra* note 1, at 89.

[102] *Id.*

[103] *Id.*

data.[104] In the future, it is therefore conceivable that a Learning AWS could continuously use its information and experience (at the tactical level) to inform and reconfigure plans on how tactical forces should be employed (at the operational level).[105] In this way, the use of a Learning AWS would effectively merge the three levels of war, eliminating the delay and miscommunications caused by the feedback loop occurring at each step.

Further, as demonstrated by reinforcement learning systems such as AlphaGo Zero, the most advanced machine learning technologies today may in the future be able to carry out relatively complex strategic decisions, similar to those involved in warfare, better than humans can.[106] This potential has caught the eye of the Chinese People's Liberation Army, for whom AlphaGo Zero has "decisively demonstrated the immense potential of artificial intelligence to take on an integral role in decision-making in future warfare."[107] In this way, future Learning AWSs may not only produce military outcomes that are militarily advantageous for the belligerents, but also facilitate more IHL-compliant targeting decisions. They may be able to do this by processing vast volumes of intelligence data faster and more accurately (*e.g.*, using facial recognition technology to distinguish between combatants and civilians). Learning AWSs can then use that data to inform the nuanced legal judgments required for IHL compliance, *e.g.*, more accurately assessing expected collateral damage for the purposes of a proportionality analysis.[108]

Yet it is unclear how meaningfully a human operator can supervise machine learning systems, especially in time-critical scenarios. As Andreas Matthias explains, meaningful human intervention ''is impossible when the machine has an informational advantage over the operator . . . [or] when the machine cannot be controlled by a human in real-time due to its processing speed and the multitude of operational variables. . ."[109] In this situation, the requirement of *ex ante* human approval fundamentally misunderstands how more sophisticated learning systems operate and is unlikely to enhance a Learning AWS's compliance with IHL.

[104] BOULANIN & VERBRUGGEN, *supra* note 25, at 16.

[105] *See* discussion *supra* Section II.A.2.

[106] Kania, *supra* note 48.

[107] *Id.*

[108] JACOB TURNER, ROBOT RULES: REGULATING ARTIFICIAL INTELLIGENCE 356 (2018).

[109] Matthias, *supra* note 11, at 182–83.

b. Inscrutability

Even if a Learning AWS works off of relatively limited or sparse data,[110] the processes of the most advanced machine learning technologies, such as deep learning and reinforcement learning, might be inscrutable to a human.[111]

A recent example of such a "black box" AI is NVIDIA's self-learning and self-driving car.[112] The car did not require a single instruction provided by an engineer or programmer.[113] It relied instead on a deep learning algorithm that had taught itself to drive by observing human driving behavior.[114] The problem with this self-taught driving ability is that it is not entirely clear how the car makes its decisions.[115] The system is so complicated that even the engineers who designed it might be unable to identify the reason for any single action, and there is currently no clear way to give the system the ability to explain why it did what it did in every case.[116]

With automatic weapons systems that follow more basic, "if-then" rules, irregularities in the decision-making process are easier to spot. These could give prior warning that an erroneous targeting decision was about to be made, at which point the human supervisor could override the system. The opacity of machine learning techniques, on the other hand, could make it impossible for a human supervisor to identify process irregularities and pre-empt a malfunctioning targeting decision.

The U.S. Department of Defense (DoD) has identified this "dark secret heart of AI" as a key stumbling block in the military use of learning machines.[117] The DoD has even initiated an Explainable Artificial Intelligence Program that is developing ways for machine learning systems to provide a rationale for their outputs.[118] However, these rationales have severe drawbacks. First, they will generally be simplified, meaning that vital information might be lost in transmission.[119] And

[110] *See, e.g.*, John Keller, *DARPA TRACE program Using Advanced Algorithms*, MILITARY AEROSPACE (July 24, 2015) (discussing DARPA's ATR system), https://www.militaryaerospace.com/articles/2015/07/hpec-radar-target-recognition.html.

[111] Mittelstadt et al., *supra* note 10, at 4, 6.

[112] Knight, *supra* note 52.

[113] *Id.*

[114] *Id.*

[115] *Id.*

[116] *Id.*

[117] *Id.*

[118] *Id.*

[119] Mittelstadt et al., *supra* note 10, at 4.

they might take time both to put together and to understand.[120] On the battlefield, these seconds could matter.

Any future Learning AWS will likely make targeting decisions based on big data in time-critical situations, following inscrutable decision-making processes. In these circumstances, human supervision or input becomes practically meaningless. But despite being inscrutable and ungovernable by human operators during their operation, these Learning AWSs might still be capable of better IHL compliance than a human-controlled AWS. This might be due, for example, to their ability to process and analyse much larger quantities of data more accurately and swiftly than any human. They should not be prohibited simply because they cannot be meaningfully supervised by a human operator. Instead, we should consider why human supervision is necessary in the first place.

c. Margulies's Dynamic Diligence Theory

Roff and Moyes's focus on MHC at the tactical level fails to grasp the immense potential of Learning AWSs. Nevertheless, similar formulations of the MHC concept that require some minimum level of ex ante human approval prior to the delivery of force can be readily found in the academic literature and are widely endorsed by members of the international community.[121]

One alternative formulation of the MHC test that better accommodates the sophistication of Learning AWSs is Peter Margulies's "dynamic diligence" standard.[122] This standard requires that the distribution

[120] *Id.* at 4–6.

[121] *See, e.g.*, Steve Goose, *Statement To the Convention on Conventional Weapons Informal Meeting of Experts on Lethal Autonomous Weapons Systems*, HUM. RIGHTS WATCH (May 13, 2014), http://www.hrw.org/news/2014/05/13/statement-convention-conventional-weapons-informal-meeting-experts-lethal-autonomous; *Key Areas for Debate on Autonomous Weapons Systems: Memorandum for Delegates at the Convention on Certain Conventional Weapons (CCW) Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS)*, ARTICLE 36 (May 2014), http://www.article36.org/wp-content/uploads/2014/05/A36-CCW-May-2014.pdf (generally, and also referring to the broad agreement between states over the need to retain human control over the critical functions of weapon systems); Peter Asaro, *On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-making*, 94 INT'L REV. RED CROSS 687, 687-709 (2012), www.icrc.org/eng/assets/files/review/2012/irrc-886-asaro.pdf; European Parliament Resolution of 12 Sept. 2018 on Autonomous Weapon Systems, EUR. PARL. DOC. 2018/2752(RSP) (2018), http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+P8-TA-2018-0341 + 0+DOC+XML+V0//EN&language=EN.

[122] Peter Margulies, *Making Autonomous Weapons Accountable*, *in* RESEARCH HANDBOOK ON REMOTE WARFARE 405, 415 *et seq.* (Jens David Ohlin ed., 2017). Note that Margulies originally formulated the "dynamic diligence" standard as a form of superior responsibility that can apply to the deployment of an AWS as the "subordinate" in the superior-subordinate relationship. This

of control between human and machine be both "tactical" and "dynamic."[123] This entails both: (1) the ability of AWSs to actively request ex ante human review in riskier settings, such as urban areas with civilians; and (2) the possibility for humans to override an AWS's machine learning protocol at any time.[124]

Margulies goes further than Roff and Moyes by acknowledging that, in certain circumstances, human intervention might not be indispensable.[125] Where fast reaction time is crucial and an AWS could react more quickly than human operators, ex ante human authorization for each attack might not be "feasible" where it would interfere with achievement of the expected military objective.[126] It could not, therefore, be consistent with the duty to take constant care, which requires only what is "feasible."

Second, Margulies takes issue with the assumption that human supervision by itself necessarily leads to greater observance of the constant care obligation. He argues that whether human input "is a precaution against civilian casualties or an added risk factor is an empirical question."[127]

While it is true that the duty to take constant care requires only what is "feasible," and it may not always be feasible in time-critical situations to require ex ante human approval, this Article prefers Margulies's second justification. It crucially recognizes that, even where human approval is technically feasible, it may not actually enhance IHL compliance. In other words, human input may actually hinder the goal of "spar[ing] the civilian population, civilians, and civilian objects"[128] and therefore may not be required by the duty to take constant care. Margulies proposes that "while an AWS must have th*e capability* for human intervention, IHL would not *require* human intervention, if an AWS could do the job as well or better."[129] This is because the machine would not be clouded by "distortions in judgment caused by human anger, fear, and cognitive flaws [which] may exacerbate errors in the targeting process."[130]

Article will use it to assess the minimum level of human intervention that is required under IHL rules on precautions.

[123] *Id.* at 433.

[124] *Id.* at 433–34.

[125] *Id.*

[126] *Id.*

[127] *Id.* at 434.

[128] Additional Protocol I, *supra* note 16, art. 57.

[129] Margulies, *supra* note 18, at 434 (emphasis in original).

[130] *Id.*

Although this is a step forward from Roff and Moyes, Margulies still does not go far enough. For more basic weapons that employ reactive systems following simple "if-then" rules, it may be logical to require the possibility of human intervention. Their processes are transparent and their behaviors are predictable.[131] However, with future Learning AWSs, optimal IHL outcomes may require that there be *no* possibility for human override, contrary to Margulies's standard.

Take the example of reinforcement learning technology again.[132] AlphaGo Zero was more powerful than previous versions of AlphaGo because by training against itself it was unconstrained by "the preconceived notions, rules of thumb, and conventional wisdom upon which most human decision-makers rely."[133] By training purely against a machine, AlphaGo Zero was able to discover unconventional strategies and imaginative new moves.[134] Similarly, any future Learning AWS that uses reinforcement learning could make targeting decisions "that humans may not have considered, or that they considered and rejected in favor of more intuitively appealing options."[135]

Consider a hypothesis where a future Learning AWS plans to strike target *X*, leading to ten civilian casualties. If the Learning AWS includes the possibility of human override as advocated by Margulies, a human supervisor would most likely override the Learning AWS and opt instead for target *Y*, which might be more intuitively "lawful" under IHL but in fact leads to less than ten civilian casualties. Target Y might be a more appealing choice to a human supervisor because, for example, he or she assumes that the factual scenario fits a recurring factual pattern, but in fact some of the variables are different. In this scenario, it seems clear that the duty to take constant care would actually require the Learning AWS to be insulated from human control because this is the best way to ensure that "the civilian population, civilians and civilian objects" are spared.[136]

The duty to take constant care must always be interpreted in light of the purpose set out in Article 57 of Additional Protocol 1 of the Geneva Conventions of 12 August, 1949, which is to "spare the civilian population, civilians and civilian objects."[137] The paradigm of "the more

[131] BOULANIN & VERBRUGGEN, *supra* note 25, at 9.

[132] *See* discussion *supra* Section II.B.2.

[133] Matthew U. Scherer, *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies*, 29 HARV. J.L. & TECH. 353, 365 (2016) (discussing AI systems more generally).

[134] Silver & Hassabis, *supra* note 47.

[135] Scherer, *supra* note 133, at 365.

[136] Additional Protocol I, *supra* note 16, art. 57.

[137] *Id.*

human control, the better" underlying many MHC doctrines should therefore be reconsidered in light of the developing capabilities of machine learning technologies. The question should always be whether human supervision is actually contributing to minimizing civilian harm.

d. Schuller's Reasonable Predictability Theory

As discussed above, both Roff and Moyes's and Margulies's formulations of the MHC standard are unsatisfactory. By requiring ex ante human supervision, or the possibility of human supervision, they fail to accommodate the potential of Learning AWSs, which may require *insulation* from human control.

Schuller's variant of the MHC doctrine, on the other hand, moves away from the traditional notion of control as human supervision.[138] Instead, his test is based on whether an AWS can comply with IHL to a reasonable level of predictability (the "reasonable predictability" test).[139] If a human operator employing a future Learning AWS is satisfied that it passes the "reasonable predictability" test, there is no further requirement for human interaction with the AWS prior to lethal action.[140]

This Article endorses Schuller's movement away from "control as human intervention" and towards "control as predictability." Of the variants of MHC examined so far, Schuller's best accommodates future Learning AWSs. Unlike Margulies, Schuller would not object to a Learning AWS insulated from human control as long as it is reasonably predictable that the system will act in an IHL-compliant way.[141]

But Schuller's reasonable predictability standard remains incomplete. It focuses on the desired result, *i.e.*, reasonably predictable compliance with IHL. Yet he offers scant practical guidance on how the development of lawfully autonomous AWSs might be achieved.[142] Schuller sets out only a handful of relatively anodyne principles in this respect: (1) AWSs may be lawfully controlled through programming alone; (2) IHL does not require proximate *ex ante* human approval prior to lethal action; (3) reasonable predictability is not required over all aspects of the AWS's behavior, only over those relevant to IHL

[138] Schuller, *supra* note 14.

[139] *Id.* at 408–09. While Schuller formulates this test to ensure IHL compliance more generally, the following analysis will focus on how it might apply to the duty to take constant care.

[140] *Id.* at 420–21.

[141] *Id.* at 420–23.

[142] *Id.* at 415–25.

compliance; and (4) the destructive potential of an AWS can be limited through its physical capabilities and programming.[143] Finally, Schuller adds that a lethal targeting decision may never be "functionally delegated" to a computer such that a human is no longer able to ensure that the AWS complies with the reasonable predictability standard.[144]

These principles fail to address several key attributes of Learning AWSs. But despite its shortcomings, Schuller's doctrine provides the right starting point for defining how the duty to take constant care might apply to Learning AWSs. The following section will suggest a number of additional principles to guide the lawful development of Learning AWSs.

e. Beyond Meaningful Human Control

Schuller justifies his reasonable predictability standard on the grounds that Learning AWSs will inherently have some level of unpredictability.[145] However, he fails to explain how the standard might apply where control is shared between the AWS and human operator. He explains: "if we [the human operator] cannot reasonably predict whether the machine will comply with IHL, it may be unlawfully autonomous."[146] But where human and machine share control, the question of whether the machine will comply with IHL will depend on the human controller as well.

There may be a number of situations where such a "human-on-the-loop" scenario arises. In relation to self-driving cars, for example, SAE International, a standard-setting organization for engineering professionals,[147] has defined various levels of vehicle automation. At partial automation (level 2), the human operator monitors the driving environment and steers, accelerates, and decelerates only when necessary (as judged by the human operator).[148] In comparison, at conditional automation (level 3), the automated vehicle monitors the driving and traffic environment, and the human operator steers, accelerates, and decelerates only when prompted by the vehicle automation system.[149]

[143] *Id.* at 417–25.

[144] *Id.* at 415–25.

[145] *Id.* at 409–13.

[146] *Id.* at 409.

[147] Formerly the Society of Automotive and Aerospace Engineers.

[148] Bryant Walker Smith, *Lawyers and Engineers Should Speak the Same Robot Language*, *in* ROBOT LAW, 98 (Ryan Calo, A. Michael Froomkin, & Ian Kerr eds., 2016).

[149] *Id.* at 98.

At the level of partial automation (level 2), there is the possibility of so-called "automation bias." This refers to a human's tendency to trust an automated system, in spite of evidence that the system is unreliable or wrong in a particular case. The concern is that users might abdicate too much responsibility to the automated system.[150]

Automation bias is already a risk with autonomous weapons and is not unique to Learning AWSs. Automation bias is routinely taken into account in the testing of auto-piloted planes, for example.[151] Indeed, the phenomenon can be present whenever machines assist human decision-making.[152] But machine learning systems may exacerbate the automation bias problem in two ways.[153]

First, an awareness of the sophistication of the learning algorithms, coupled with the inscrutability of the machine learning process, the aforementioned "dark secret heart of AI,"[154] could lead to an increased human tendency to trust the machine.[155] This automation bias may cause humans not to intervene even if the signs of system malfunction are obvious.[156]

The second element that should be taken into account in testing and verification is the "reverse automation bias" discussed above in relation to reinforcement learning.[157] This occurs when a human is *more* inclined to intervene because of a counterintuitive targeting decision made by a Learning AWS, where this human intervention might in fact lead to a worse IHL outcome.

The shortcoming of Schuller's reasonable predictability standard is its focus on the uncertainty arising from the Learning AWS's interactions with a complex battlefield environment.[158] In this way, he fails to examine the unpredictability that might emanate from the Learning AWS's interactions with its human operator. To fully reflect the very

[150] *See, e.g.*, Chantal Grut, *The Challenge of Autonomous Lethal Robotics to International Humanitarian Law*, 18 J. CONFLICT & SEC. L. 5, 14–15 (2013); Mary L. Cummings, *Automation and Accountability in Decision Support System Interface Design*, 32 J. OF TECH. STUD. 23 (2006).

[151] Kathleen Mosier et al., *Aircrews and Automation Bias: The Advantages of Teamwork?*, 11 INT'L J. AV. PSYCHOL. 1 (2001).

[152] *See, e.g.*, Grut, *supra* note 150, at 14–15 (discussing the 1988 *USS Vincennes incident*).

[153] Cosima Gretton, *The Dangers of AI in Healthcare: Risk homeostasis and automation bias*, TOWARDS DATA SCIENCE (June 24, 2017), https://towardsdatascience.com/the-dangers-of-ai-in-health-care-risk-homeostasis-and-automation-bias-148477a9080f.

[154] *See* discussion *supra* Section III.D.3.b.

[155] Grut, *supra* note 150, at 19.

[156] Kevin Neslage, *Does "Meaningful Human Control" Have Potential for the Regulation of Autonomous Weapon Systems?*, 6 NAT'L SEC. & ARMED CONFLICT L. REV. 151, 173–4 (2015).

[157] *See* discussion *supra* Section III.D.3.c.

[158] Schuller, *supra* note 14, at 409–13.

real risk of automation bias or reverse automation bias, a triangular approach is required. Unpredictability does not arise only in the interaction between the Learning AWS and its operating environment. A crucial third plane of analysis must be the interaction between the Learning AWS and the human operator.

At the design, testing and verification phase, therefore, a Learning AWS should not only train under different battlefield scenarios (e.g., various terrains, weather conditions, numbers and locations of civilians and combatants). It must also train with different human operators in each battlefield scenario to assess how they interact with the machine. This would help to ensure that any psychological nuances and potential unpredictable outcomes of the human-machine interaction are fully taken into account.

f. Beyond Reasonable Predictability

Schuller's test does not require near-certainty of IHL compliance, which is generally recognized as an "insurmountable goal."[159] But any standard lower than reasonably predictable compliance, he argues, would invite human operators to blame malfunctioning computers for violations.[160] Schuller further explains that the reasonableness standard has the advantage of being a well-established benchmark of performance and that anything higher would be unattainable "based on the complexity of computer programming magnified by the 'fog' of the modern battlefield."[161]

But this is only part of the picture. While "reasonable predictability" is the correct minimum benchmark for determining whether a Learning AWS is "lawfully autonomous," there should be a corresponding duty to optimize its compliance with IHL, where possible or "feasible." In this context, the duty to take constant care would not only require the Learning AWS to pursue the outcome where IHL compliance is reasonably predictable. It would require the Learning AWS, over and above that standard, to act in a way that *maximizes* the probability of compliance with IHL compliance, where feasible. Therefore, the reformulated standard should be "(feasible) optimal predictability, but at least reasonable predictability."

[159] Schuller, *supra* note 14, at 408; *see also* Prosecutor v. Delalić, Case No. IT-96-21-T, Judgment, ¶ 395 (Int'l Crim. Trib. for the Former Yugoslavia Nov. 16, 1998) ("[N]ecessary and reasonable measures" are "limited to such measures as are within someone's power, as no one can be obliged to perform the impossible.").

[160] Schuller, *supra* note 14, at 408.

[161] *Id.*

i. "Optimal Predictability" and Necessity

Schuller argues that states would not accept a higher standard than "reasonableness," which is the widely accepted and understood standard under IHL and "has balanced the competing interests of IHL for quite some time."[162]

But "optimal predictability" does not raise the current legal standard. It simply ensures that the duty to take constant care is correctly applied in conjunction with the other precautionary obligations that operationalize it, such as the principle of least expected collateral damage under Article 57(2)(a)(ii) and target selection under Article 57(3) of Additional Protocol 1 to the Geneva Conventions of 12 August 1949.[163]

Both rules reflect the fundamental IHL principle of military necessity, which permits only acts that are necessary to achieve a legitimate military purpose and not otherwise unlawful under IHL.[164] The military necessity principle provides the broad legal basis for the "optimal predictability" test because it prohibits any harm that goes beyond what is necessary to weaken the enemy. [165]

Furthermore, as mentioned above, the military necessity principle is one of the fundamental principles of IHL. This means that it underpins and informs the interpretation of the entire body of IHL.[166]

This Article argues, therefore, that applying the fundamental principle of military necessity to a Learning AWS would require it to be programmed to comply with IHL *as a whole* to an optimal level of predictability, as far as feasible.

ii. "Optimal Predictability" and Feasibility

Another potential counter-argument to the optimal predictability standard is that precautionary obligations require only what is feasible,

[162] *Id.* at 409.

[163] Additional Protocol I, *supra* note 16, arts. 57(2)(a)(ii), 57(3); *see also* Janina Dill, *Applying the Principle of Proportionality in Combat Operations*, POL'Y BRIEFING OF OXFORD INST. FOR ETHICS, L. AND ARMED CONFLICT 9 (2010).

[164] Chris af Jochnick & Roger Normand, *The Legitimation of Violence: a Critical History of the Laws of War*, 35 HARV. INT'L L.J. 49 (1994).

[165] *See, e.g.*, Hague Convention (IV) Respecting the Laws and Customs of War on Land and Its Annex: Regulations Concerning the Laws and Customs of War on Land, art. 23(g), Oct. 18, 1907, U.N.T.S. 539 (stating that enemy property cannot be seized or destroyed unless "imperatively demanded by the necessities of war").

[166] *See* Legality of the Threat or Use of Nuclear Weapons, Advisory Opinion, 1996 I.C.J. 226, 493 (July 8) (dissenting opinion of Weeramantry J.) (stating that the fundamental principles of IHL "provide both nourishment for the development of the law and an anchorage to the mores of the community").

and requiring anything more than "reasonable predictability" would go beyond that obligation.[167]

But let us recall the nature and purpose of the "feasibility" qualifier. The Tallinn Manual, in its restatement of the international law applicable to cyber operations, defines "feasibility" as what is "practicable or practically possible, taking into account all circumstances ruling at the time, including humanitarian and military considerations."[168] It can therefore be seen as "operational and interpretatory leeway"[169] that acknowledges it is not feasible for a human commander in the heat of battle to scrupulously pursue the *optimal* IHL outcome.[170] To demand any higher standard would "unduly shift risks towards one's own soldiers who in high-risk scenarios should not be burdened with the additional task of having to evaluate the availability of less harmful means."[171]

But in the AWS context a feasibility criterion could also raise the legal standard, since "an autonomous weapon could be a means to render certain precautions feasible which would not be so for a soldier."[172] For example, machine learning systems can be programmed according to utility theory to generate optimal outcomes in pursuit of a defined goal.[173] This theory works by instructing the machine to act based on the probability of certain outcomes as a function of the utility of such outcomes in achieving the desired goal. In other words, the rational decision of such a system ". . . depends on both the relative importance of various goals and the likelihood that, and degree to which, they will be achieved."[174] The machine will then pursue the course of action that provides the best outcome based on this calculus.[175]

Take a hypothetical where it is possible to program a Learning AWS to comply with IHL to a reasonable level of predictability. Based on the above, it would be entirely "feasible" to program a Learning AWS to seek the particular military objective or course of action that is *most* likely to comply with IHL. Additional algorithms would probably be

[167] *See* discussion *supra* Section III.B.

[168] TALLINN MANUAL 2.0, *supra* note 81, at 479.

[169] Bhuta, Beck & Geiss, *supra* note 90, at 373.

[170] Bhuta *et al.* make a similar argument in relation to an AWS applying the principle of distinction. *See id.* at 376.

[171] *Id.*

[172] ICRC Report, *supra* note 20, at 42.

[173] This is the essence of a so-called "deliberative system." BOULANIN & VERBRUGGEN, *supra* note 25, at 10–11.

[174] Schuller, *supra* note 14, at 411 (internal quotations omitted).

[175] *Id.*

required at the programming stage to implement utility theory, at negligible marginal cost and effort. Indeed, many machine learning systems, such as reinforcement learning, are already programmed to optimize outcomes in this way.[176] This seems to lie entirely within the realms of what is "practicable or practically possible."[177]

The potential of learning systems to carry out complex normative decisions that weigh up legal, ethical, and moral rules has been shown in the domain of bioethics.[178] A machine learning system has been developed that can produce coherent answers to some bioethical questions using the weighted utility theory outlined above. This Article agrees with Margulies when he argues that "[i]f we can formulate and implement such logical rules for the bioethics context, we should in theory be able to do the same for IHL."[179]

Unlike more basic automatic systems, the emerging ability of learning systems to carry out nuanced, strategic decisions opens up the potential for Learning AWSs to take over discretionary decision-making in the targeting cycle in the future. And the more the targeting process is mechanized, the more feasible it is to apply the higher "optimal predictability" standard for IHL compliance.

Furthermore, states would find it difficult to object to the "optimal predictability" standard because it simply seeks to apply the existing law in the right way. IHL's system of rules overlaid by fundamental principles has often been lauded as particularly adaptive to new technologies such as nuclear weapons. IHL should rightly flex to raise standards when applied to the technological paradigm shift that is the advent of machine learning weapons.[180] As Nehal Bhuta, Susanne Beck, and Robin Geiss argue, while the current geopolitical climate may militate against a "large-scale reconsideration of existing rules of [IHL], at least a progressive and dynamic interpretation of the existing rules, which takes into consideration the specificities of AWSs, should not be excluded."[181] This Article agrees and proposes the optimal predictability standard as one such progressive interpretation of existing rules.

[176] Magnus Stensmo & Terrence J. Sejnowski, *Learning Decision Theoretic Utilities Through Reinforcement Learning*, *in* ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 9 1061–1067 (Michael C. Mozer, Michael I. Jordan, & Thomas Petsche eds., 1996), http://dl.acm.org/citation.cfm?id=2998981.2999130 (last visited Dec. 21, 2018).

[177] CUSTOMARY INTERNATIONAL HUMANITARIAN LAW, *supra* note 71, at 54.

[178] Margulies, *supra* note 18, at 420.

[179] *Id.*

[180] Bhuta, Beck & Geiss, *supra* note 90, at 370.

[181] *Id.* at 375.

IV. Conclusion

The future implementation of machine learning techniques such as deep learning and reinforcement learning in AWSs demands a radical rethink of notions of "meaningful human control." Learning AWSs may in the future comply with IHL just as well, or even better, without human control. Or it may require human control to comply with IHL. The point is that a case-by-case analysis is required. Banning these weapons in all instances because they cannot be meaningfully controlled could mean losing a potential instrument for minimizing human suffering in future conflicts.

Instead, the paradigm of "the more human control, the better," currently favored by many scholars and members of the international community, should be reconsidered. Schuller provides the most workable framework in the context of Learning AWSs by shifting the focus from human control to predictability, *i.e.*, whether a Learning AWS can predictably comply with IHL. This Article argues that this also rightly shifts the focus to the development of prophylactic measures to ensure that machine learning weapons can comply with IHL rules in the first place. In a machine learning paradigm, it is at the stage of design, testing, and verification that human control and human supervision could be most meaningful.

AlphaGo Zero was able to greatly surpass human abilities because it was not prone to "the preconceived notions, rules of thumb, and conventional wisdom upon which most human decision-makers rely."[182] The AlphaGo Zero experience should inspire a fresh and more nuanced approach to the application of IHL to these new technologies, in order to fully leverage their potential to minimize human suffering in armed conflict.

[182] Scherer, *supra* note 133, at 365.