ARTICLE

At the Crossroads of Control: The Intersection of Artificial Intelligence in Autonomous Weapon Systems with International Humanitarian Law

Alan L. Schuller[*]

[*] Judge Advocate, Lieutenant Colonel, U.S. Marine Corps. Associate Director, Stockton Center for the Study of International Law, U.S. Naval War College, and Fellow, Center on National Security and the Law, Georgetown University Law Center. The views herein should not be attributed to any of the author's institutional affiliates, to include the U.S. Department of Defense.

Abstract

This Article explores the interaction of artificial intelligence (AI) and machine learning with international humanitarian law (IHL) in autonomous weapon systems (AWS). Lawyers and scientists repeatedly express a need for practical and objective substantive guidance on the lawful development of autonomy in weapon systems. This Article proposes five foundational principles to enable development of responsible AWS policy. The findings emerged from a research project conducted by a team of military and civilian professors at the Stockton Center for the Study of International Law at the U.S. Naval War College. The study is informed by experts in computer sciences, government and military, non-governmental organizations, think tanks, and academia.

Advances in AI will likely produce AWS that are different *in kind* from existing weapon systems and thus require a fresh approach to evaluating IHL compliance. First, this Article describes the technological details pertinent to understanding the distinction between current and future systems. It argues that the technological evaluation of the spectrum of autonomy should focus on the combination of authorities granted to the computer that controls an AWS, while also taking into account the physical capabilities of the system. Second, it argues that a key issue bearing on IHL compliance is whether an AWS has been granted some combination of authorities and capabilities that *functionally* delegate the decision to kill from human to machine. Third, it posits that predictability must be at the core of an evaluation into whether a particular AWS breaches this delegation threshold and examines how AI handles uncertainty, a critical component of the predictability analysis. Finally, the Article proposes five foundational principles to guide the development of AWS policy.

Table of Contents

Introduction

Autonomous weapon systems (AWS) are the most militarily significant yet legally elusive challenge to international humanitarian law (IHL) since the proliferation of cyber operations. The modern debate over AWS ignited following the release of *Losing Humanity*,[1] a report co-authored by Harvard's International Human Rights Clinic and Human Rights Watch.[2] Since then, academics, government officials, non-government organizations (NGOs), and military leaders alike have struggled to address the myriad legal concerns potentially raised by AWS.[3]

In the years following *Losing Humanity*, it became apparent that the dilemmas presented by AWS would never be solved by any one professional field operating in isolation. Lawyers were hamstrung by a dearth of technical expertise. Scientists were hampered by a lack of legal acumen. Non-military personnel were confounded by their unfamiliarity with the likely battlefield application of AWS. And everybody was forced to sift through a significant amount of misinformation surrounding AWS.[4]

[1] HUMAN RIGHTS WATCH & INT'L HUMAN RIGHTS CLINIC, HARVARD LAW SCH., *Losing Humanity: The Case Against Killer Robots* (2012) [hereinafter *Losing Humanity*], http://www.hrw.org/sites/default/files/reports/arms1112_ForUpload.pdf; *see also* HUMAN RIGHTS WATCH & INT'L HUMAN RIGHTS CLINIC, HARVARD LAW SCH., *Advancing the Debate on Killer Robots: 12 Key Arguments for a Preemptive Ban on Fully Autonomous Weapons* (2014) [hereinafter *Advancing the Debate*], https://www.hrw.org/sites/default/files/related_material/Advancing%20the%20Debate_8May2014_Final.pdf.

[2] The debate that this report ignited is captured by a number of exchanges on Lawfare, involving, among others, Tom Malinowski, then-director of Human Rights Watch, and Ben Wittes of the Brookings Institution. *See* Benjamin Wittes, *Does Human Rights Watch Prefer Disproportionate and Indiscriminate Humans to Discriminating and Proportionate Robots?*, LAWFARE (Dec. 1, 2012), http://www.lawfareblog.com/2012/12/does-human-rights-watch-prefer-disproportionate-and-indiscriminate-humans-to-discriminating-and-proportionate-robots/ [hereinafter Wittes, *Does Human Rights Watch Prefer*]; *see also* Kenneth Anderson & Matthew Waxman, *Tom Malinowski Responds on Autonomous Lethal Systems*, LAWFARE (Dec. 5, 2012), http://www.lawfareblog.com/2012/12/tom-malinowski-responds-on-autonomous-lethal-systems/; Benjamin Wittes, *Tom Malinowski Responds on Lethal Autonomous Systems: Part II*, LAWFARE (Dec. 6, 2012), http://www.lawfareblog.com/2012/12/tom-malinowski-responds-on-lethal-autonomous-systems-part-ii/; Benjamin Wittes, *Tom Malinowski Ups the Game in Lawfare's Discussion of Killer Robots*, LAWFARE (Jan. 14, 2013), http://www.lawfareblog.com/2013/01/tom-malinowski-ups-the-game-in-lawfares-discussion-of-killer-robots/.

[3] Among the legal concerns, those regarding the application of IHL to AWS are arguably most pressing. Within the IHL context, questions arise primarily regarding: (1) how to ensure AWS comply with the general principles of IHL, and (2) how principles of accountability (such as command responsibility) will apply to AWS employment. This Article focuses on the technology of AWS and how it will intertwine with the application of IHL principles to AWS. There are also a multitude of non-legal concerns raised by AWS, including potential ethical dilemmas.

[4] *Compare, e.g.*, Bonnie Docherty, *Killer robots are 'quickly moving toward reality' and humanity only has a YEAR to ban them, expert warns*, DAILY MAIL (June 17, 2016), http://www.dailymail.co.uk/sciencetech/article-3647006/Killer-robots-quickly-moving-reality-humanity-YEAR-ban-expert-warns.html (stating erroneously that "[r]emoving humans from the targeting decision would create a dangerous world. Machines would make life-and-death

But after years of debate, we still continue to ponder: is there nothing to be concerned about, as some in government and industry would have us believe?[5] Or is humanity's war with machines imminent unless we take immediate action to ban AWS completely, as a few NGOs argue?[6] And, more to the point, why is applying IHL to AWS so difficult? After countless academic discussions,[7] law review articles,[8] conferences,[9] meetings of experts,[10] and consultations between state representatives,[11] the conversation about AWS has matured little. There remains a dearth of practical guidance on how states should regulate AWS development.[12] The root cause underlying this lack of progress is simpler than one might suspect: IHL is a sub-optimal tool for remedying our inability to predict the future.[13]

determinations outside of human control. The risk of disproportionate harm or erroneous targeting of civilians would increase. No person could be held responsible."), *with* Christopher P. Toscano, *"Friends of Humans": An Argument for Developing Autonomous Weapons Systems*, 8 J. NAT'L SEC. L. & POL'Y 189, 192 (2015) ("AWS use does not require a new legal paradigm because these machines are weapons systems at all times, not sentient beings.").

[5] Some government officials in the United States and abroad have privately expressed the view that we should not be overly concerned about AWS because, for a number of policy and practical reasons, governments will not develop unpredictable AWS that violate IHL because it would not be in their self-interest. This argument is unsatisfying in that entrusting weapons policy simply to self-interest invites deviation from policy when interests change instead of discouraging it.

[6] *See, e.g.*, *Learn*, CAMPAIGN TO STOP KILLER ROBOTS, https://www.stopkillerrobots.org/learn (last visited Apr. 20, 2017).

[7] *See, e.g.*, Benjamin Wittes, Lecture at Georgetown Univ. Law Center (Apr. 16, 2013) (arguing that we should not adopt a treaty *ex ante* to outlaw AWS); *but see* Tom Malinowski, Lecture at Georgetown Univ. Law Center (Apr. 16, 2013) (arguing in support of his Lawfare blog posts criticizing AWS).

[8] *See, e.g.*, Michael N. Schmitt & Jeffrey S. Thurnher, *"Out of the Loop": Autonomous Weapon Systems and the Law of Armed Conflict*, 4 HARV. NAT'L SEC. J. 231 (2013).

[9] *See, e.g.*, Workshop on Legal Implications of Autonomous Weapon Systems, Stockton Ctr. for the Study of Int'l Law, U.S. Naval War College (Feb. 6–7, 2014).

[10] *See, e.g.*, Expert Meeting on Autonomous Weapon Systems at the Int'l Comm. of the Red Cross, Versoix, Switz. (Mar. 15–16, 2016).

[11] *See, e.g.*, Meeting of Experts on Autonomous Weapon Systems at the Conference of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, Geneva, Switz. (Apr. 11–15, 2016) [hereinafter Meeting of Experts].

[12] The general sense from those charged with developing AWS policy is that there is little practical legal guidance available. *See* Workshop on Unmanned Systems held by Deputy Assistant Secretary of the Navy for Unmanned Systems & Navy Unmanned Warfare Systems Directorate (OPNAV N99), San Diego, Cal. (July 26–29, 2016).

[13] A failed attempt to outlaw the discharge of any weapons from aircraft is one particularly shortsighted example of this concept. *See* THE LAWS OF ARMED CONFLICTS: A COLLECTION OF CONVENTIONS, RESOLUTIONS, AND OTHER DOCUMENTS 309 (Dietrich Schindler & Jiri Toman eds., 4th ed. 2004) (describing the process); *see also* Declaration (XIV) Prohibiting the Discharge of Projectiles and Explosives from Balloons, Oct. 18, 1907, 36 Stat. 2439; Declaration (IV, 1) to Prohibit for the Term of Five Years, the Launching of Projectiles and Explosives from Balloons, and other Methods of Similar a Nature, July 29, 1899, 32 Stat. 1839.

IHL governs the conduct of parties to a conflict.[14] Its principles and rules generally serve to balance the practical realities of armed conflict with the desire to protect civilians from harm and combatants against unnecessary suffering.[15] Under IHL, attacks must serve a valid military purpose. A target may not be attacked unless it qualifies as a military objective and the commander must weigh the military advantage to be gained in attacking a target against expected collateral damage to ensure the collateral damage is not excessive in the circumstances.[16] The principles of IHL have developed as a function of treaty and customary international law distilled from the lessons of countless armed conflicts.[17]

It is tempting to think that we might use IHL to completely forestall some yet-undefined harmful technology. States toil tirelessly, however, to resolve legal problems that already exist;[18] when it comes to solving problems that are not yet realized, the task is nearly insurmountable. But why must we predict the future—could we not simply inventory all potential AWS and then make some general statements about how IHL might apply?[19] Probably not. An inventory approach would have to determine, as a preliminary matter, precisely what constitutes an AWS.[20] Are we referring to existing weapon systems that exhibit a degree of autonomy, such as the Phalanx close-in weapon system (CIWS) and Counter Rocket, Artillery, and Mortar (C-RAM) systems?[21] Or to likely near-future systems, such as unmanned sub-hunting ships[22] or swarming mini-drones?[23]

[14] *See generally* YORAM DINSTEIN, WAR, AGGRESSION AND SELF-DEFENCE 16 (5th ed. 2012) (discussing the scope of IHL).

[15] *See, e.g.*, COMMENTARY TO GENEVA CONVENTION I FOR THE AMELIORATION OF THE CONDITION OF THE WOUNDED AND SICK IN THE ARMED FORCES IN THE FIELD 39 (Jean Pictet ed., 1952) (Regarding the "origin and development of the idea" for standards of humane treatment embodied in Common Article 3, "[t]he principle of respect for human personality, which is at the root of all the Geneva Conventions, was not a product of the Conventions. It is older than they are and independent of them.").

[16] These principles are discussed in greater detail in Part IV.

[17] *See* 1 CUSTOMARY INTERNATIONAL HUMANITARIAN LAW ix (Jean-Marie Henckaerts & Louise Doswald-Beck eds., 2005) ("The laws of war were born of confrontation between armed forces on the battlefield.").

[18] *See, e.g.*, Workshop, Syria: Can International Law Cope?, Stockton Ctr. for the Study of Int'l Law, U.S. Naval War College (Nov. 16–18, 2015).

[19] This Article examines the particular problems of such an approach in Part II.B.1. Although the law generally attempts to use historical examples as a framework to approach future dilemmas, this approach proves unsatisfactory in the context of AWS.

[20] For an in-depth discussion of this matter, *see generally* Christopher M. Ford, *Autonomous Weapons and International Law*, 67 S.C. L. REV. (forthcoming 2017).

[21] *See Phalanx Close-in Weapons System*, RAYTHEON.COM, http://www.raytheon.com/capabilities/products/phalanx/ (last visited Mar. 30, 2017) ("At sea, the Phalanx close-in weapon system—a rapid-fire, computer-controlled, radar-guided gun system—is designed to defeat anti-ship missiles and other close-in air and surface threats. The land-based Phalanx weapon system is part of the U.S. Army's Counter Rocket, Artillery and Mortar systems used to detect and destroy incoming rounds in the air before they hit their ground targets.").

[22] *See* Sean Gallagher, *DARPA Robotic Sub-hunting Ship Will Set Sail This Spring*, ARS TECHNICA (Feb. 15, 2016), http://arstechnica.com/information-technology/2016/02/darpa-robotic-sub-hunting-ship-to-set-sail-this-spring/ (last visited Apr. 10, 2017).

Alternatively, are we referring to some fanciful, unrealized future concept such as killer robots?[24] The answer is that we simply do not know what to inventory because in many instances we would be trying to predict the development of systems that do not exist, are unlikely to exist in the foreseeable future, and might never exist. This is not how the law usually develops.

Law, and IHL in particular, is typically reactive.[25] Generally speaking, there are good reasons for this. Humans cannot predict the future. Attempting to devise complex legal schemes that will effectively anticipate actions and technologies that do not exist is vexing at best and likely counterproductive. As such, the law is more adept at fixing things that are broken after careful discussion and debate.

In the context of IHL practice, however, this means that a lot of people perish before a particular problem is addressed. As such, with the development of any weapon there is an unavoidable tension. On one hand, we may seek to prevent needless death and suffering by generally restricting the types of weapons produced and also by specifically tailoring the lethal effects of new weapons to fit a certain purpose. On the other hand, in the interest of protecting our national security we may seek to create new weapon systems that can eliminate enemy threats, which historically has meant that they were overwhelmingly destructive. From the second perspective, technological advantage is maintained by legal review of new systems on a case-by-case basis, placing concomitant limitations on use rather than broader restrictions on procurement.[26] There is no simple resolution to this dilemma—most parties involved in the development of IHL strive to strike a reasonable balance between these competing interests.

[23] *See, e.g.*, Paul Scharre, CTR. FOR A NEW AM. SEC., *Robotics on the Battlefield II: The Coming Swarm* (Oct. 2014), https://s3.amazonaws.com/files.cnas.org/documents/CNAS_TheComingSwarm_Scharre.pdf; Adam Clark Estes, *How 3D Printing Will Create On-Demand Swarms of Disposable Drones*, GIZMODO (Mar. 30, 2014), http://gizmodo.com/how-3d-printing-will-create-on-demand-swarms-of-disposa-1553933989; Joshua Steinman, *Imagine the Starling: Peak Fighter, the Swarm, and the Future of Air Combat*, WAR ON THE ROCKS (Feb. 17, 2016), http://warontherocks.com/2016/02/imagine-the-starling-peak-fighter-the-swarm-and-the-future-of-air-combat/; *see also* Press Release, Def. Advanced Research Projects Agency, Fast Lightweight Autonomy Program Takes Flight, (Feb. 12, 2016), http://www.darpa.mil/news-events/2016-02-12.

[24] *See Learn*, *supra* note 6 ("[F]ully autonomous weapons . . . would be able to choose and fire on targets on their own, without any human intervention.").

[25] In the U.S. federal courts, this concept is reflected in the justiciability doctrine, which holds in part that courts will not rule on cases or controversies that are not "ripe" for consideration. *See* Erwin Chemerinsky, *A Unified Approach to Justiciability*, 22 CONN. L. REV. 677, 677 (1990) ("Familiar and well-settled law requires that, in order for a federal court to hear a case, several justiciability doctrines must be met: the case must not present an advisory opinion; there must be standing; the case must be ripe; it must not be moot; and it must not present a political question.").

[26] The manner in which one balances these important interests often hinges on the institutional biases inherent in one's profession. Those in the military acquisitions field are justifiably concerned that they obtain the greatest possible advantage over peer competitors. On the other hand, professionals who strive to enhance humanity during the conduct of hostilities rightly focus on alleviating unnecessary death, destruction, and suffering.

The potential impact of AWS technology on this balance is nebulous, but we must nevertheless glean meaningful practical guidance. We should not attempt, however, to solve intractable problems by debating capabilities that may not exist even 100 years from now;[27] those matters are simply too speculative to provide a foundation for meaningful debate. We should instead evaluate from a technical perspective the particularized challenges presented by autonomy in the reasonably foreseeable future. This Article seeks to craft broad but useful and substantively meaningful principles that further an understanding of IHL's application to AWS. In particular, the Article will dissect AI and its interplay with IHL.

This Article proceeds initially with a brief historical perspective on issues surrounding autonomy and then pinpoints the details pertinent to understanding how AWS technology intersects with IHL. After thus framing the discussion, it explains why an IHL analysis of AWS must focus primarily on the computer system. The Article next explains why future autonomy should be related logically to the human decision-making cycle, then explores how AI works and, in particular, how machines learn. It argues that in the context of AWS equipped with AI, an IHL evaluation must focus on predictability, and then examines how AI handles a primary challenge to predictability: uncertainty. After a brief review of the relevant IHL concepts, the Article proposes five principles to guide the development of AWS technology.

I. The Roots of Autonomy and Controversy

A. *Historical Perspective*

Automation is not new. Since the dawn of civilization, human beings have sought tools to mechanically assist them in completing a myriad of tasks.[28] In every endeavor across the spectrum of the human experience, mankind has developed implements designed to alleviate burdens previously borne by people. From agriculture[29] and industry[30] to national defense and warfare,[31] thousands of

[27] "We can only see a short distance ahead, but we can see plenty there that needs to be done." Alan Turing, *Computing Machinery and Intelligence*, 49 MIND 460 (1950). An intractable problem is defined as "not easily governed, managed, or directed." *Intractable*, MERRIAM-WEBSTER.COM, http://www.merriam-webster.com/dictionary/intractable (last visited Mar. 30, 2017).

[28] *See* Smithsonian Nat'l Museum of Natural History, *Early Stone Age Tools*, HUMANORIGINS.SI.EDU, http://humanorigins.si.edu/evidence/behavior/stone-tools/early-stone-age-tools (last visited May 9, 2017) ("The earliest stone toolmaking developed by at least 2.6 million years ago."); *see also* 2001: A SPACE ODYSSEY (MGM 1968) (in the opening scenes depicting the "dawn of man," human predecessors transition from using verbal and physical threats against a hostile tribe to employing improvised weapons against them.).

[29] The machine tractor is one example.

[30] Consider the advent of the machine assembly line.

[31] *See* WILLIAM H. BOOTHBY, CONFLICT LAW: THE INFLUENCE OF NEW TECHNOLOGY, HUMAN RIGHTS, AND EMERGING ACTORS 4–5 (2014) ("The digital revolution is only a part of the rapid

years of technological development have yielded devices that preceding generations could not have imagined.[32] Mechanization led to automation that helped us to accomplish our goals not only with greater ease, but also more quickly and efficiently than was previously possible. Unsurprisingly, humans have invariably adapted new technology to military applications when it might provide a warfighting advantage.[33]

The pace of automation has increased exponentially.[34] Machines have transitioned from simple automatic implements fashioned to assist us to autonomous computerized mechanisms sometimes capable of replacing us.[35] Konrad Zuse invented the first operational programmable computer in 1941.[36] Today—within the span of one human lifetime—automobiles can drive themselves.[37] Bursts of technological advancement have induced speculation about whether machines might eventually outperform their creators not only in routine physical tasks but also in behaviors previously viewed as uniquely

technological transformation of warfare that we have seen over the last couple of decades. . . . The ways in which war is conducted and the associated technology that is employed are . . . continually changing."); *see also* JAMES J. BUSUTTIL, NAVAL WEAPONS SYSTEMS AND THE CONTEMPORARY LAW OF WAR 15 (1998) ("The naval mine technology available today ranges from the simple to the esoteric. Unsophisticated mines from before World War I are used alongside microprocessor mines of the computer age.").

[32] *See* DEF. SCI. BD., *Summer Study on Autonomy* 5 (June 9, 2016), https://www.hsdl.org/?abstract&did=794641 ("Advances in AI are making it possible to cede to machines many tasks long regarded as impossible for machines to perform.").

[33] The Wright brothers' first flight at Kitty Hawk occurred on December 17, 1903. *See NOVA: Wright Brothers' Flying Machine* (PBS television broadcast Nov. 11, 2003). The first aerial combat victory occurred barely a decade later during World War I, on October 5, 1914. *See* Tony Reichhardt, *The First Aerial Combat Victory*, AIR & SPACE (Oct. 4, 2014), http://www.airspacemag.com/daily-planet/first-aerial-combat-victory-180952933/; *see also* WILLIAM H. BOOTHBY, WEAPONS AND THE LAW OF ARMED CONFLICT 363 (2009) ("If . . . technology may represent a significant military advantage to a state or states, the law can only make a difference if the states concerned can be persuaded to forego that advantage, and that may not be easy to achieve."). On the other hand, a multitude of technologies developed initially for military application later saw implementation for peaceful purposes. *See, e.g.*, *Biography of Dr. Wernher von Braun*, MARSHALL SPACE FLIGHT CTR. HISTORY OFFICE, https://history.msfc.nasa.gov/vonbraun/bio.html (last visited Apr. 10, 2017) ("[T]he V-2 rocket was the immediate antecedent of those used in space exploration programs in the United States and the Soviet Union.").

[34] *See* BOOTHBY, *supra* note 31, at 363 (describing technological advances). Might soldiers someday use their brainwaves to control weapons? *See* Stephen E. White, *Brave New World: Neurowarfare and the Limits of International Humanitarian Law*, 41 CORNELL INT'L L.J. 177, 177 (2008) ("[DARPA] has engaged in research on direct neurological control of weapon systems.").

[35] *Compare Robot History*, INT'L FED'N OF ROBOTICS, http://www.ifr.org/robot-history/ (last visited Mar. 29, 2017) (describing first industrial robot, developed in 1959, which "weighed two tons and was controlled by a program on a magnetic drum."), *with Worker Ant Robots Could Shape Production Lines in the Future*, UNMANNED SYSTEMS MISSION CRITICAL, Vol. 5, No. 2, May 2015, at 5 (describing 3D printed robot ants that operate as a swarm in order to communicate production needs).

[36] Zuse invented the Z-3 in Germany. *See* STUART RUSSELL & PETER NORVIG, ARTIFICIAL INTELLIGENCE: A MODERN APPROACH 14 (2010).

[37] The Waymo (formerly Google) self-driving car is one example. *See Technology*, WAYMO, https://waymo.com/tech/ (last visited Dec. 20, 2016).

human.[38] The core concern about autonomy is that we might eventually design machines so advanced that they slip out of the grasp of human decision-making and control.[39]

B. *Framing the Discussion*

Decisions regarding the use of force in any context are often quite controversial. But in order to have an informed debate about these matters in the context of AWS it is first necessary to examine carefully, from a technological perspective, what it means to "decide." In other words, what does it mean to say that a human decided that a machine would accomplish a given task? Conversely, when is control so attenuated that it could no longer reasonably be said that a human decided a machine would accomplish that task or the manner in which the machine should complete it?

One decision in particular—the decision to kill—lies at the heart of concerns over AWS.[40] The decision to kill inherently invokes analysis under IHL as to the lawfulness of a use of force. The burden of conducting this evaluation logically and necessarily must be borne by a human.[41] But the link between a human's decision to kill and the lethal kinetic action of a weapon continues to

[38] *See, e.g.*, EX MACHINA (Universal Pictures 2016) (android becomes self-aware and acts based on self-preservation); WARGAMES, (MGM/United Artists 1983) (computer is placed in control of America's nuclear arsenal and nearly causes World War III); 2001: A SPACE ODYSSEY, *supra* note 28 (computer kills the crew aboard a spacecraft because it determines that they are a threat).

[39] *See The Problem*, CAMPAIGN TO STOP KILLER ROBOTS, http://www.stopkillerrobots.org/the-problem/ (last visited Apr. 10, 2017) ("Allowing life or death decisions to be made by machines crosses a fundamental moral line . . . . [F]ully autonomous weapons would not meet the requirements of the laws of war."); *see also* FUTURE OF LIFE INST., *Autonomous Weapons: An Open Letter from AI & Robotics Researchers*, http://futureoflife.org/open-letter-autonomous-weapons/ (last visited Apr. 10, 2016) [hereinafter *Open Letter*] (arguing for "a ban on offensive autonomous weapons beyond meaningful human control.").

[40] The decision to take the life of another human being is one of the most difficult quandaries a rational human may ever face because it goes against the nature and upbringing of many cultures. *See, e.g.*, ROMANS 13:9 ("Thou shalt not kill."). A decision to kill, even when justified under the law, is intensely personal and for most humans fraught with indecision. *See* DAVE GROSSMAN, ON KILLING 4 (1995) ("[T]here is within most men an intense resistance to killing their fellow man . . . ."); *see also* THE THIN RED LINE (Twentieth Century Fox 1998) (During a lull in intense fighting, a battle-weary soldier screams, "Who decides who lives? Who decides who dies?"). Because of this conditioning, the idea that a non-human would be delegated the ultimate task of deciding—in the human sense—who lives and dies is too much for many to accept. But machines do not make decisions in the human sense. Thus, the key question from a technological standpoint is at what point has this burden been *functionally* delegated to a machine?

[41] *See* Interview with Paul Scharre, Senior Fellow and Director, Future of Warfare Initiative, Ctr. for a New Am. Sec., Washington, D.C. (Jan. 29, 2016) ("Humans are combatants; machines are not . . . . Humans have the obligation to comply with the laws of armed conflict."). Indeed, no commentators argue that the duty levied on parties in a conflict to comply with IHL norms is derogable in the case of AWS. Combatants are responsible for the reasonably foreseeable effects of the weapons they employ; this constant has held despite historical development of a multitude of weapon systems of varying predictability.

steadily degrade as a function of proximate cause,[42] and in particular as a temporal matter.[43] Thus, the appraisal of whether a human decided to kill is not a digression into philosophical inquiry; in the AWS context it is instead a technological assessment.[44] We must determine whether AWS technology could unlawfully dilute this causal link such that we could no longer say that a human *functionally* decided to kill.[45] Note, however, that this is not meant to imply that a human must provide an AWS input that is temporally proximate to lethal kinetic action.

As a corollary, we must scrutinize from a technological perspective the concept of what it means to "control" a machine. Before we could hope to answer the question of what amount or kind of human control over AWS is legally sufficient,[46] we must first possess a firm understanding of how humans control machines via programming. Indeed, if we misapprehend the manner in which machines are controlled, for example by layering legal significance onto the proximity of human interaction with the machine at the time of lethal kinetic action, then we risk misleading ourselves. Only after the technical underpinnings of machine decision-making[47] and control have been explored can we hope to parse out aspects that might prompt a legal objection. Thus, we must describe from a technological standpoint how machines "decide" and how humans control those "decisions."

[42] A more traditional model of proximate cause in this circumstance would be a soldier aiming a rifle at the enemy and firing. The link between the decision to kill and the death of the enemy is obvious. But if a weapon system is granted the ability to learn based on its environment and then to select from amongst a range of potential targets, the proximity of the decision to kill and the act of killing may be diluted.

[43] On the battlefields of even seventy-five years ago, death came soon after the firing of a rifle or lobbing of artillery rounds. Today, death may result minutes or even hours after the launch of a missile. In the future, lethal kinetic effects may follow days, weeks, months, or even years after the deployment of a weapon system with autonomous attributes.

[44] We could avoid functionally delegating the decision to kill, for example, through carefully tailored and tested computer programming. In the alternative, a control tether might suffice.

[45] Importantly, this does *not* mean that human involvement is temporally proximate to the moment of lethal kinetic action. This point is discussed in more detail in Part V.

[46] *Compare* Heather M. Roff & Richard Moyes, *Meaningful Human Control, Artificial Intelligence and Autonomous Weapons* (Apr. 8, 2016), http://www.article36.org/wp-content/uploads/2016/04/MHC-AI-and-AWS-FINAL.pdf (briefing paper prepared for Meeting of Experts, *supra* note 11) (arguing in favor of "meaningful human control" over AWS), *with* U.S. Dep't of Def. Directive 3000.09, Autonomy in Weapon Systems ¶ 4.a (2012) ("It is DoD policy that . . . [AWS] shall be designed to allow commanders and operators to exercise *appropriate levels of human judgment* over the use of force.") (emphasis added).

[47] The term "machine decision-making" is shorthand for instances in which machines are delegated by human programmers the authority to complete a task given certain inputs, variables, and/or algorithms.

II. Autonomy, Artificial Intelligence, and Machine Learning

A. *An IHL Analysis of AWS Should Focus on the Computer System*

As a point of departure, consider a conventional military unit: an artillery battery. In evaluating the ability of the battery to comply with IHL, one might start by investigating the performance characteristics of the cannons. Testing the accuracy of cannons is relatively easy.[48] But more to the point, the cannons do not decide where to aim themselves. A modern howitzer receives firing data from a computer system (a fire-control computer)[49] which has a straightforward task: given a set of geographical coordinates, calculate the proper elevation and deflection settings for the cannon while taking into account meteorological conditions and other measurable factors that will affect the trajectory of a round.[50] If the firing data caused the cannon to hit the target, it was correct.[51] If the round was off target, the data was incorrect. This is simple computer automation. The battery commander is responsible for ensuring that his or her weapons are employed lawfully. It would be inconceivable for a prosecutor to say, for example, that an IHL violation was the fault of the fire-control computer. The computer system is a very advanced calculator. But for AWS, the computer system plays a central role in the IHL analysis.

In the context of AWS, the primary focus of technological analysis necessarily shifts to the integrated system rather than the weapon component alone, for two reasons. First, computer-managed systems enable weapons to be removed from the continuous physical control of a human; a weapon that is not incorporated into such a system is therefore unremarkable from an IHL-compliance perspective because it ceases to be an AWS.[52] Second, it is a safe assumption that autonomy will continue to increase in modern weapons.[53] As

[48] *See* Elizabeth R. Dickinson, U.S. Army Materiel Command, Ballistic Research Laboratories, *The Production of Firing Tables for Cannon Artillery* (1967), http://www.dtic.mil/dtic/tr/fulltext/u2/826735.pdf (describing the process by which tabular firing table data is created for new cannons); *but see* Def. Sci. Bd., *supra* note 32, at 30 ("DoD's current testing methods and processes are inadequate for testing software that learns and adapts.").

[49] *See Advanced Field Artillery Tactical Data System*, RAYTHEON.COM, http://www.raytheon.com/capabilities/products/afatds/ (last visited Dec. 21, 2016) (AFATDS "provide[s] automated support for planning, coordinating, controlling and executing fires and effects.").

[50] *See* HEADQUARTERS, U.S. DEP'T OF THE ARMY, FM 6-40, TACTICS, TECHNIQUES, AND PROCEDURES FOR THE FIELD ARTILLERY MANUAL CANNON GUNNERY 1–3 (1996) (describing the process by which deflection and elevation data are calculated).

[51] Assuming the other requirements for accurate predicted fire were met. *See id.* at 1–3.

[52] And those systems that are not under continuous control, such as landmines, are so deterministic that the humans emplacing them can be certain of the result given a particular input.

[53] Relatively few broad generalizations regarding future AWS withstand careful analysis, but one can reasonably conclude that the computers onboard future AWS may need to have more discretion programmed into their computers than traditional computation-focused systems, such as an artillery computer.

such, the computer systems linked to weapons will play an increasingly significant role in how the weapon is employed. With respect to the system, we focus primarily on the computer that effects control of the machine. This is because, in large part, the level of autonomy a system enjoys is determined by the computer that effects control of it.

To be sure, the physical capabilities of the mechanical platform on which the computer is installed, as well as the characteristics of any accompanying weapon may play a significant role in describing the autonomy of the complete system. But if the computer effecting control of the machine sets limiting parameters on the system, the overall capabilities of the system may be moot. The converse is not true. By way of simple example, an unmanned aircraft might have the mechanical ability to carry a large unguided bomb. But if the sophisticated targeting computer onboard the aircraft only allows it to vector towards unpopulated areas in order to attack positively identified enemy tanks during an international armed conflict, our concerns over civilian casualties may be reduced. On the other hand, a similar platform with a different computer system that is granted unbridled discretion in attacking targets could raise significant legal concerns, almost without regard to the type of weapons on board. Indeed, it is difficult to imagine any weapon system that would not appear objectionable if given "a mind of its own" in the most human sense.

However, machines do not have minds of their own—they have computers programmed by humans. Generally, computers do what they are told to do. That said, computers of today are infinitely more complicated than they were even a few decades ago. Concepts such as machine learning and AI are quickly becoming the focus of the discussion on autonomy,[54] and factors like these greatly complicate the control analysis. Is it possible that advanced machines could eventually "decide" to do something other than what a human "told" them to do, or do something that we thought them incapable of doing?[55]

We must therefore delve deeper into the technological manner in which autonomy functions in machines and the role that machine learning and AI are likely to play in its future. Only with this understanding can we fashion a legal framework that adequately addresses concerns about control over machine decision-making.

[54] I argue in this Article that they should be at the forefront. *See also* DUSTIN A. LEWIS, GABRIELLA BLUM & NAZ K. MODIRZADEH, HARVARD LAW SCH. PROGRAM ON INT'L LAW AND ARMED CONFLICT, WAR-ALGORITHM ACCOUNTABILITY (2016) (identifying "algorithmically-derived 'choices' and 'decisions'" as a "central concern regarding technical autonomy in war").

[55] This is an intentional oversimplification of the question at this juncture. "An agent is anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators." RUSSELL & NORVIG, *supra* note 36, at 34. A simple reflex agent acts based solely on what it is able to perceive according to "if/then" condition-actions rules such as "if *car-in-front-is-braking* then *initiate-breaking*." *Id.* at 48–50. Technology advanced long ago past the point where actions taken by machines were the relatively predictable result of simple computers with reflex agents operating on "if/then" condition-action rules.

B. *Describing Autonomy from a Technical Perspective*

1. The Spectrum of Autonomy

The notion of autonomy in the context of machines is so broad as to defy simple definition.[56] It is accordingly helpful to think of autonomy as a spectrum or series of spectrums[57] rather than a singular concept.[58] And as described in the previous section, autonomy with respect to AWS must concentrate primarily on the computer system while also taking into account the physical capabilities of the system. As such, a description of the AWS spectrum should focus on the combination of authorities and capabilities that designers grant to computer systems running AWS. There is no flawless analytical construct to accomplish this task, but a helpful method for describing potential combinations is the Boyd Cycle, or "OODA Loop."

The OODA Loop is a simple way to evaluate human decision-making based on a continuous process: Observe, Orient, Decide, Act. As part of the cycle:

> a person first *observes* the world around her, gathering data about her environment through the array of human senses. Second, she *orients* herself, or interprets the information she has gathered. Third, she weighs the potential courses of action based on the knowledge she has

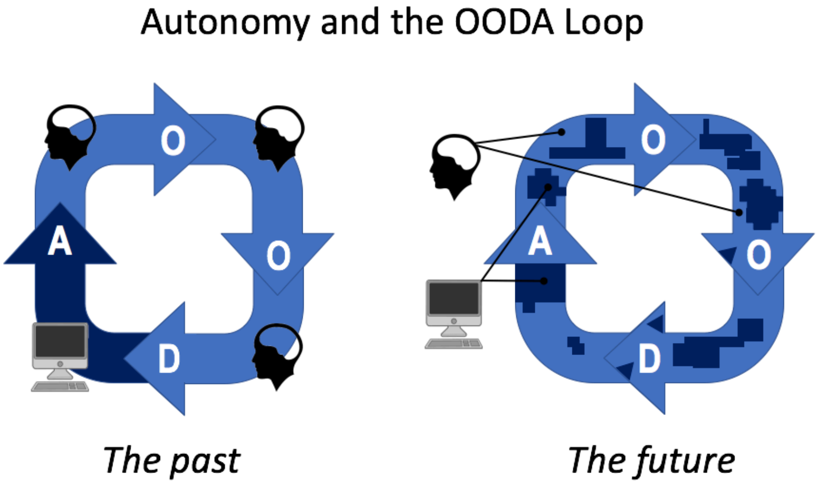
[56] *See* RUSSELL & NORVIG, *supra* note 36, at 28 ("What can AI do today? A concise answer is difficult because there are so many activities in so many subfields."). We should stop trying to draw a clear line between autonomous and automated. This is a futile effort that attempts to paint over infinite shades of gray with a façade of order. It is also likely a quest to know the unknowable. Most importantly, there is no legal tipping point inherent in these descriptions because they are non-linear at best and arbitrary at worst. *See* Interview with Paul Scharre, *supra* note 41 (arguing that it is not helpful to delineate overly generalized distinctions between gradations of autonomy). More automation does not always lead to autonomy or to legal challenges, and as such these categorizations are not useful in describing specific combinations of autonomy that are legally problematic.

[57] *See* Paul Scharre & Michael C. Horowitz, *An Introduction to Autonomy in Weapon Systems*, Working Paper (Washington, D.C., Ctr. for a New Am. Sec., Feb. 2015), 5, https://www.cnas.org/publications/reports/an-introduction-to-autonomy-in-weapon-systems, ("autonomy does not exist on merely one spectrum, but on three spectrums simultaneously").

[58] Attempts to comprehensively describe legal categories of autonomous systems *in toto* have borne relatively little fruit. The technology is too diverse to categorize succinctly yet comprehensively from a legal perspective. Whether a machine is allowed to "select and engage" a target may be useful in describing a segment of automation we should take a careful look at due to its operational significance, but it is less helpful in defining a category of automation that is legally objectionable. Instead of attempting to describe and formulate specific rules for the entire possible spectrum of autonomy, we should focus on broad governing principles for the responsible development of AWS.

> accumulated and *decides* how to act. Fourth and finally, she *acts*, or executes the decision she has made.[59]

Through this model, we can generally describe the types of authorities that might be granted to a machine that could supplant the human's role in fulfilling the requirements of the various points in a decision-making cycle. Importantly, in the context of controlling AWS, this Article does not refer to the moment where a human might be inserted into the cycle,[60] but instead to those fragments of the loop that have been delegated to computers. Granting certain portions of the OODA loop to machines may ultimately create issues with IHL compliance. This key distinction is illustrated in the following figure:



In this vision of the OODA loop, the puzzle pieces of tasks delegated to a computer in the future might consist of authority (e.g., in the precise programming or learning capacity of the computer) and/or physical capabilities (e.g., the ability of the host platform to loiter for long duration). Thus, the key issue bearing on IHL compliance is not whether the machine, for example, selects and engages targets (i.e., decides and acts) without human intervention.[61] Rather, the critical issue is whether designers have granted the machine some combination of tasks that *functionally* delegates the decision to kill from human to machine.[62]

[59] 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) (footnotes omitted).

[60] *See* Scharre & Horowitz, *supra* note 57, at 8 (describing the construct of human in, on, or out of the loop).

[61] *See, e.g.*, Meeting Report, Int'l Comm. of the Red Cross, *Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons*, Expert Meeting 8 (2016), https://www.icrc.org/en/publication/4283-autonomous-weapons-systems (defining an AWS as any system that can "select . . . and attack . . . targets without human intervention").

[62] This model is not a perfect fit insofar as it invites us to ascribe a "decision" to a machine, when in most situations this is misleading as the decision to behave in a certain way is, generally speaking, established by programming. Setting aside that issue, however, the OODA Loop is

2. The "Observe" Phase

Machines have historically performed the task of observing military operations and they continue to advance in their capability to do so. The first airplanes were used as platforms for humans to reconnoiter First World War battlefields.[63] Now, unmanned vehicles carry sensor suites that allow humans to observe the enemy from afar.[64] Ground-based sensor platforms guard demilitarized zones from intrusion and ship-mounted sensors can detect and destroy threats to warships.[65] Facial recognition technology and other sensors may soon advance to the point where machines are able to positively identify a person without additional input by a human at the time of observation.[66] Indeed, the range of sensors deployed in today's modern militaries is vast.

Nevertheless, machine observation and identification of objects and persons on the battlefield have generated little debate or concern from a legal perspective. Indeed, it would be beyond cavil to suggest that IHL restricts machine sensors that simply detect objects. Other fragments in the OODA loop, especially in combination with portions of the Observe phase, may however become troublesome.

3. The "Orient" Phase

In this phase, information gained during the Observe phase is analyzed in order to better understand the operating environment. In the context of human decision-making, a commander will consider the totality of the information at his or her disposal. Current intelligence estimates, sensor collection and battlefield reports are reviewed, the tactical and strategic implications weighed, as are countless other military and non-military considerations. The experiences of the commander play a key role. In the end, the human decision-maker may simply trust his or her "gut" feeling.

nevertheless a useful way of envisioning the spectrum of autonomy insofar as it is a generally accepted method of evaluating human decision-making in the employment of weapons, and thus the ways in which computers might supplant certain actions by humans.

[63] *See How did World War One's battle in the skies change warfare?*, BRITISH BROAD. CORP., http://www.bbc.co.uk/guides/zgxhpv4#z9pfyrd (last visited May 9, 2017).

[64] *See, e.g.*, *MQ-9 Reaper Fact Sheet*, AF.MIL, http://www.af.mil/AboutUs/FactSheets/Display/tabid/224/Article/104470/mq-9-reaper.aspx, (last visited Mar. 30, 2017).

[65] *See* Tim Hornyak, *Korean Machine-gun Robots Start DMZ Duty*, CNET (June 14, 2010), https://www.cnet.com/news/korean-machine-gun-robots-start-dmz-duty/ (describing the Samsung SGR-1 ground-based sensor platform) [hereinafter SGR-1]; *Phalanx Close-in Weapons System*, *supra* note 21 (ship-mounted sensor).

[66] Facial recognition technology with remarkable accuracy has already been implemented in social media. *See, e.g.*, Haje Jan Kamps, *Apple Introduces Facial and Object Recognition for Mobile Photographers*, TECHCRUNCH (June 13, 2016), https://techcrunch.com/2016/06/13/apple-image-and-facial-recognition/.

Machines do not have guts. They analyze data based on their programming. The process by which a computer-controlled machine completes this task depends on the attributes granted to it by a human. At the most basic level, machines analyze a rather narrow category of information against set decision-making processes.[67] This is essentially a flow chart, or an "if this, then that" method, of evaluating data. Broadly speaking, this process works well for simple machines completing basic tasks.

So far as we can reasonably tell at this time, machines in the foreseeable future will not have guts, hunches, or any of the other qualities unique to humans. They may, however, complete tasks in ways that are far more difficult for humans to predict. For example, the technology embedded even within many currently existing computers contains so many lines of code that testing out all defects can be impractical or even impossible.[68] Machine learning further complicates the task of predicting machine behavior.[69] While the equations employed in programming a machine may themselves be easy to understand, we may not be able to determine ahead of time the results that they will produce.[70] These are the computing methods that will likely be employed in future AWS.[71]

Machine learning and AI therefore pose difficult questions about how machines will perform in the future. Part of the dilemma in this regard exists because humans have proven rather inept at predicting technological advances, even in the short term.[72] With current technologies we can often predict at least

[67] *See* discussion *supra* note 52.

[68] *See* Interview with Alan C. Schultz, Director, Laboratory for Autonomous Systems Research, U.S. Naval Research Lab., Washington, D.C. (Jan. 28, 2016) (explaining this dilemma).

[69] An algorithm is "a procedure for solving a mathematical problem (as of finding the greatest common divisor) in a finite number of steps that frequently involves repetition of an operation; *broadly*: a step-by-step procedure for solving a problem or accomplishing some end especially by a computer." *Algorithm*, MERRIAM-WEBSTER.COM, http://www.merriam-webster.com/dictionary/algorithm (last visited Mar. 30, 2017).

[70] *See* Interview with Leslie Pack Kaelbling, Learning and Intelligent Systems Group, Computer Science and Artificial Intelligence Laboratory, Mass. Inst. of Tech., Cambridge, Mass. (Sept. 16, 2016) (explaining the critical role of training environments in machine learning); *see also* interview with Alan C. Schultz, *supra* note 68 (noting that it may be possible to reverse engineer why a system took a particular course of action); interview with Naval Undersea Warfare Ctr. autonomous systems specialists, Naval Sea Systems Command, U.S. Navy, Newport, R.I. (Oct. 2015) [hereinafter NUWC Interview] (describing the particular difficulties in this context with neural networks).

[71] *See* Conference Paper, William F. Bundy, *Future Maritime Forces: Unmanned, Autonomous, and Lethal*, EMC Chair Symposium: Maritime Strategy (Mar. 23–24, 2016), https://www.usnwc.edu/Academics/Faculty/Derek-Reveron/Workshops/Maritime-Strategy/working-papers/bundy.aspx (arguing for the development of "fully autonomous" and "intelligent" unmanned maritime systems); *see also* NUWC Interview, *supra* note 70 (supporting generally the premise that future systems will include learning technology).

[72] Due to a combination of "known unknowns" and "unknown unknowns." *See* Press Release, U.S. Dep't of Def., DoD News Briefing - Secretary Rumsfeld and Gen. Myers (Feb. 12, 2002), http://archive.defense.gov/Transcripts/Transcript.aspx?TranscriptID=2636; Kenneth Anderson & Matthew Waxman, *Human Rights Watch Report on Killer Robots, and Our Critique*, LAWFARE (Nov. 26, 2012), http://www.lawfareblog.com/2012/11/human-rights-watch-report-on-killer-

what processes a computer will use to analyze data or reverse engineer the process that brought a computer to a certain course of action based on its programming.[73] This capability, however, is rapidly eroding and may not generally be the case in the very near future.[74]

4. The "Decide" Phase

The most provocative and controversial aspect of task delegation to AWS is the Decide phase. This is because humans are generally uncomfortable with machines completing the "final" deliberative step in a sequence of events that will ultimately result in the death of a human. But it is a common error to posit that if a computer takes the last step in a process, the machine "decided" to kill.

It does not necessarily follow that a machine "decided" to kill simply because a machine selected and engaged a target based on certain narrow and pre-determined parameters.[75] Machines generally do as they are programmed to do, and humans make decisions to program machines a certain way. The programming could be based on immensely complex learning algorithms operating in stochastic environments[76] or it could be simple AI executing trivial tasks in a controlled situation. In the end, actions by the machine are the result of a human delegating a task to the machine for performance according to certain pre-established performance measures. Computers simply do not act on their own volition.

Nevertheless, this is an area that requires additional scrutiny because we must consider whether the link between the programming and lethal kinetic action might become so diluted that we cannot reasonably say a human decided to kill. Further, we must consider the possibility that a machine might be granted such advanced technology that it unexpectedly exceeds anticipated operating parameters or otherwise behaves incongruently with the discretion that humans

robots-and-our-critique/ ("[It is not] wise or even possible to decide today what targeting technology might or might not be able to do – say, a generation from now.").

[73] *See* Interview with Def. Advanced Res. Projects Agency robotics specialists, U.S. Dep't of Def., Arlington, Va. (Jan. 28, 2016) [hereinafter DARPA Interview]; interview with Alan C. Schultz, *supra* note 68.

[74] *See* Interview with Leslie Pack Kaelbling, *supra* note 70 (expressing skepticism that scientists will continue to be able to reverse engineer actions by a learning system in order to establish why it took a particular course of action); *see also* Will Knight, *The Dark Secret at the Heart of AI*, MIT TECH. REV., May/June 2017, at 56–60 (describing the difficulties associated with explaining why deep neural networks take certain actions).

[75] Indeed, no one would argue that a Tomahawk cruise missile "decided" to destroy an object when the GPS location of that target was programmed into the weapon by a human. Delays in temporal proximity between programming and a lethal kinetic event may prove legally significant when evaluating compliance with the principle of proportionality, for example, but they do not necessarily mean that a machine has made the decision to kill.

[76] That is to say, an environment that involves random variables—the opposite of which is a deterministic environment, where the operating environment remains stable and outcomes are determined solely by the actions of the machine in that environment. *See* RUSSELL & NORVIG, *supra* note 36, at 43 (contrasting stochastic and deterministic operating environments).

programmed it to demonstrate. These crucial questions of predictability will be discussed in greater detail below.

5. The "Act" Phase

The final phase in the cycle relates to the physical authority in time and space that designers have granted to the machine. Although this aspect is of secondary significance to the decision-making authority the machine has been granted, physical authorities may prove significant depending on the other aspects of the system. For example, to what kind of weapons is the system provided access? How long is the system able to loiter in the operating environment? How far is it able to travel? Can it detect the presence of nonmilitary objects and/or noncombatants? These and other questions may prove highly relevant to the evaluation of AWS. Then again, lack of discretion on the part of the machine to take advantage of these capabilities may obviate their significance.

In sum, our technological evaluation of the spectrum of autonomy should focus on the combination of authorities granted to the computer system that controls the machine while also taking into account the physical capabilities of the AWS. And the critical issue bearing on IHL compliance from a technological perspective is whether the AWS has been granted some combination of capabilities that functionally delegates the decision to kill from human to machine. This, however, is only the initial step in delineating the kind of autonomy that could prove legally objectionable under IHL. The next step is to more fully explain how autonomy functions from a technical standpoint so that we can fully understand machine decision-making.

C. *An IHL Analysis of the Computer System Should Focus on AI*

In order to discern the aspects unique to AWS that may interfere with IHL compliance, we must first identify the particular qualities of AI most relevant to the inquiry. In order to outpace sophisticated future adversaries, computers that run AWS may be equipped with the capacity to "learn" even after they are employed by a battlefield commander. As a result, at least certain aspects of future learning-equipped AWS will adapt in ways that we may be unable to predict. As such, historic examples of weapon systems that incorporated AI, which did not employ learning AI, are of little use in this context as future AI will be different in kind from past AI.

1. Surveying Current Technologies is Unhelpful

This Article does not attempt to conduct a comprehensive survey of existing technologies in an effort to craft generalized legal principles for AWS. It is tempting to select various weapons systems currently in existence and then try to rationalize them under some overarching rubric for AWS. One might reasonably question why we cannot simply construct a historically focused framework for systems that might exist in the future. This approach, while sensible at first glance, ultimately would be a fruitless endeavor.

As discussed in Part II.B.1, the breadth of legacy weapon systems—those already deemed lawful and fielded—that could be considered autonomous is vast. The question, for example, of whether or not a conventional land mine is autonomous is intriguing from a purely academic perspective, but the discussion has almost no practical value for those crafting regulations for future AWS development.[77] Other systems that exhibit some aspects of autonomy have also been reviewed and have enjoyed lawful status under IHL for quite some time.[78] As such, even those organizations most vehemently opposed to the development of AWS do not argue that legacy systems exhibiting elements of autonomy would be included in a ban.[79] Thus, an attempt to neatly categorize scores of legacy systems within comprehensive legal principles on AWS would do little to advance the conversation.[80]

More importantly, legacy technology will not significantly inform our evaluation of whether future AWS are lawful under IHL because future systems will be different in *kind*, not simply different in degree of autonomy.[81] Thus, past

[77] Land mines have been used as a means of warfare for many years and are already governed by specific legal regimes. *See* Convention on the Prohibition of the Use, Stockpiling, Production and Transfer of Anti-Personnel Mines and on Their Destruction, Sept. 18, 1997, 2056 U.N.T.S. 211.

[78] *See, e.g.*, *Phalanx Close-in Weapons System*, *supra* note 21. As an aside, the autonomy, or lack thereof, of the Phalanx is not dispositive on the issue of IHL compliance. *Compare* Press Release, U.S. Dep't of Def., DoD News Briefing (June 4, 1996), http://archive.defense.gov/Transcripts/Transcript.aspx?TranscriptID=527 (describing an incident in which a Japanese Navy CIWS accidentally shot down a U.S. warplane during joint exercises), *with* U.S. DEP'T OF DEF., FORMAL INVESTIGATION INTO THE CIRCUMSTANCES SURROUNDING THE DOWNING OF IRAN AIR FLIGHT 655 ON 3 JULY 1988 (1988) (describing the circumstances surrounding an incident where human operators on the USS VINCENNES in the Persian Gulf shot down a civilian airliner that they deemed a threat).

[79] *See* Interview with Mary Wareham and Bonnie Docherty, Human Rights Watch Arms Div., Washington, D.C. (Jan. 29, 2016) (expressing skepticism that legacy weapons systems would be included in a ban).

[80] Nevertheless, someone who accepts the proposition that a legal framework for future AWS need not revisit the legality of legacy systems might still argue that IHL architecture for AWS should include consideration of legacy systems' features. This is problematic because the model would be forced to contemplate the characteristics of legacy systems while abstaining completely from a fresh review of their legality.

[81] For example, the Samsung SGR-1 presents no vexing weapons-review issues because it is reasonably similar to legacy systems. The system is designed to be deployed in static positions along a demilitarized zone through which no persons are lawfully allowed to pass. *See* SGR-1, *supra* note 65. It is able to detect a human being and shoot them, although as currently fielded it requires human approval to engage. *See id.* It can also determine when a person raises their hands in surrender and hold fire. *See* IFSEC Global, *Intelligent Surveillance & Security System Samsung Techwin*, YOUTUBE (Mar. 10, 2009), https://www.youtube.com/watch?v=NevCAx6zWNU (showing mock attackers successfully surrender). But this operating environment is highly constrained and the discretion (shoot/do not shoot) provided to the weapon system is likewise narrow. When one compares SGR-1 to a legacy system such as an electric fence, the similarities are striking, assuming of course that trespassers are put on notice of the system's presence. The most pertinent difference weighs in favor of the legality of SGR-1: an electric fence cannot accept

analyses regarding the ability of legacy systems to comply with IHL will not be particularly useful in predicting whether future systems are lawful. It goes without saying that legacy systems are generally used as precedent for evaluating new systems insofar as they are reasonably similar. Weapons systems markedly different in kind from legacy systems, however, require novel approaches to applying the law. Weapon systems that are granted advanced AI and learning capability are so different from legacy systems that they require a fresh approach to applying IHL.[82] The focus must thus necessarily be on the systems that incorporate AI and machine learning in new and profoundly different ways.

2. Advanced AI will create autonomy that is different in *kind*

If left without practical guidance on the lawful development of AWS, future systems might be designed with AI that is so advanced that designers could not predict to a reasonable certainty how it would perform in an operational environment. To say that there is disagreement in the scientific community regarding what paths AI might take in the future would be an understatement. This Article does not enter the fray in this regard. Instead, we begin by establishing a baseline description for what direction—based on extensive research and interviews with leading experts in the field—AI will not take.

Contrary to the news headlines or Hollywood productions that one might encounter when researching AI,[83] the singularity is not near,[84] nor is "Skynet."[85]

surrender. As such, the system is an excellent example of one that varies in *degree* of automation from legacy systems but not in *kind*. That being said, if the SGR-1 was granted more advanced autonomy, such as machine learning, and/or provided with broader physical capabilities, an evaluation under IHL could easily become immensely complicated. For example, if the system was modified such that it could roam the countryside, but was not granted more sophisticated sensor and/or computational capabilities or limited by additional parameters, the evaluation of legality would be markedly different. Likewise, if the system were programmed to learn about human behaviors such as feigning surrender the analysis would be more difficult.

[82] To be sure, there may be some basic commonalities in at least the procedural application of weapons reviews between systems that are different in kind from each other—for example, a laser, a tank, and a cyber weapon. The law is of course the same and the requirements to conduct the reviews are identical. But the principles through which we apply the law to the systems may be quite distinct. In other words, the questions we must ask to inform our decision as to the lawfulness of a given system under IHL will be different based on the kind of system we evaluate. In much the same way that cyberwarfare forces us to re-evaluate how IHL applies, we must understand that autonomous systems are different *in kind* from legacy systems. *See generally* TALLINN MANUAL ON THE INTERNATIONAL LAW APPLICABLE TO CYBER WARFARE (Michael N. Schmitt ed., 2013) (providing a framework for evaluating cyber weapons that were different in kind from legacy weapons).

[83] News headlines concerning AI often lead with a picture from the "Terminator" movie series. *See, e.g.*, Docherty, *supra* note 4. The Terminator embodies and promotes popular misconceptions about AI. *See* Ardalan Raghian & Matthew Renda, *When Hollywood does AI, it's fun but far fetched*, CNET (June 30, 2016), http://www.cnet.com/news/hollywood-ai-artificial-intelligence-fun-but-far-fetched/ ("Least realistic [depiction of AI in a movie]: 'The Terminator.' This pop-culture touchstone is universally reviled by the AI community . . . [because it commits] the dual sin of overemphasizing the robotics aspect of AI and also vesting AI with human qualities like a hunger for power and an aptitude for murder.")

Even if such a technological tipping point might theoretically be reached,[86] IHL typically does not attempt to regulate technology that *might* exist in 30 years, 100 years, or may never exist.[87] Yet, the question remains whether there is anything that we should be concerned about when incorporating AI into AWS? The answer is yes.

Problems may arise with systems that are quite advanced but not smart *enough* to do what we demand of them. Specifically, AWS that either fail to meet performance standards under IHL or, more pertinent to the present discussion, whose performance cannot be adequately predicted due to their AI raise very real concerns.[88] With regard to the former, this is a relatively simple weapons testing question with which countries executing weapons reviews are quite familiar.[89] But with respect to AI in particular, a significant possibility exists that computers

[84] *See generally Special Report: The Singularity*, IEEE SPECTRUM (2016), http://spectrum.ieee.org/static/singularity. Most AI experts do not believe it is reasonably possible that machines will become sentient beings capable of outperforming humans at thinking and behaving in distinctly human ways. *See also* DARPA Interview, *supra* note 73; interview with Paul Scharre, *supra* note 41; interview with Alan C. Schultz, *supra* note 68; Alfred Nordmann, *Singular Simplicity*, IEEE SPECTRUM (June 1, 2008), http://spectrum.ieee.org/robotics/robotics-software/singular-simplicity (concluding that "there is nothing wrong with the singular simplicity of the singularitarian myth—unless you have something against sloppy reasoning, wishful thinking, and an invitation to irresponsibility"); *see also* DAVID A. MINDELL, OUR ROBOTS, OURSELVES 9 (2015) ("[The] myth of full autonomy" is "the utopian idea that robots, today or in the future, can operate entirely on their own."). We shall therefore set aside the possibility that computers might eventually become "self-aware." This is a fanciful suggestion and there is little debate that machines possessing self-awareness and human-like discretion would be legally problematic.

[85] *See* John Vogel, *Terminator 2: Judgment Day: Plot Summary*, INTERNET MOVIE DATABASE, http://www.imdb.com/title/tt0103064/plotsummary?ref_=tt_stry_pl (last visited Apr. 7, 2017) ("Skynet, the 21st century computer waging a losing war on humans sends a . . . terminator back in time to destroy the leader of the human resistance while he is still a boy.").

[86] But even "true believers" in the singularity do not envision it occurring inside of 30 years. *Special Report*, *supra* note 85.

[87] *See* W. Hays Parks, *Air War and the Law of War*, 32 A.F. L. REV. 1, 11 (1990) (citing a military officer who remarked long ago regarding proposed prohibitions on balloon warfare, "at present let us confine our action within the limits of our knowledge"); Kenneth Anderson & Matthew Waxman, *Law and Ethics for Autonomous Weapon Systems: Why a Ban Won't Work and How the Laws of War Can*, HOOVER INST. TASK FORCE ON NAT'L SEC. AND L.AW 27 (Apr. 9, 2013), http://www.hoover.org/research/law-and-ethics-autonomous-weapon-systems-why-ban-wont-work-and-how-laws-war-canhttp://www.hoover.org/publications/monographs/144241 (recommending reliance "on the gradual evolution and adaptation of long-standing law of armed conflict principles . . . to regulate what seems to many like a revolutionary technological and ethical predicament" in part because "the challenge of regulating apparently radical innovations in weaponry within a long-standing legal . . . framework is hardly novel").

[88] We should of course address AWS that are allocated pieces of the OODA loop but whose predictable execution of those tasks degrades performance of the system below that which is accepted by IHL. *See* Anderson & Waxman, *supra* note 72 (questioning "whether artificial intelligence and computer analytic systems could ever reach the point of satisfying the fundamental . . . legal principles of distinction and proportionality").

[89] *See generally* Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts art. 36, June 8, 1977, 1125 U.N.T.S. 3 [hereinafter API] (setting forth an obligation to review new weapons).

could advance to the point where we are unable to reasonably predict whether the weapon system will comply with IHL. In other words, a system could be so advanced that we would be unable to conduct testing sufficient to assure us to a reasonable degree of certainty about how it would perform in certain operational circumstances.[90] For this reason, we must understand how AI works and in what ways predictability and uncertainty manifest themselves in AI systems.

3. The Focus of AI in AWS: Performing to Rational Objective Standards

AI in AWS should be evaluated based upon how well it performs to rational and objective standards. Not all visions of AI follow this construct. Some models of AI are concerned with thought processes and reasoning by computers whereas others focus on behavior and performance.[91] How we define AI therefore hinges on whether we seek to have machines think like humans, act like humans, think rationally, or act rationally.[92] In approaching AWS from a legal standpoint, this Article's primary concern is the ability of a given system to comply with IHL. The focus of any evaluation under IHL is the lethal and destructive effects caused by the AWS.[93] Thus, the goal is to have the AWS act rationally by producing these effects in accordance with objective standards.

From early in the debate on AWS, most experts have agreed that aspiring to a human performance standard probably set the bar too low.[94] Indeed, humans sometimes demonstrate a sub-optimal track record in adhering to the standards of IHL.[95] And most agree that autonomous systems in non-military applications should also be required to outperform humans in rule compliance.[96] So we proceed from what is likely the uncontroversial proposition that AI in the AWS context should be designed and judged based upon how well the system is able to perform to rational and objective standards established by humans and informed

[90] *See* Interview with Leslie Pack Kaelbling, *supra* note 70 (expressing skepticism regarding our ability to predict the behavior of systems equipped with machine learning).

[91] *See* RUSSELL & NORVIG, *supra* note 36, at 1 ("[AI] attempts not just to understand [how we think] but also to build intelligent entities.") (emphasis omitted).

[92] The specific focus of any evaluation of AI depends largely on whether we are more concerned with a system's processes or outputs. *See* RUSSELL & NORVIG, *supra* note 36, at 27. And, with respect to processes and outputs, we must inquire as to whether we seek to mimic humans or instead to attain an objectively rational standard. *See id.* at 1 ("[Success can be measured] in terms of fidelity to *human* performance . . . [or instead] against an *ideal* performance measure, called rationality.").

[93] Whether or not a computer is able to mimic the way in which the human brain functions or whether its computational processes appear rational are of no concern in the context of AWS.

[94] *See* Wittes, *Does Human Rights Watch Prefer*, *supra* note 2 (calling for "the development and deployment of automated technologies in those instances in which they would perform better than people and not in those instances in which they would make things worse") (emphasis omitted).

[95] *See* Wittes, *supra* note 7 (arguing that we should not adopt a treaty which outlaws AWS in part because humans have a checkered record of complying with IHL, so we should allow for the possible development of machines that are better at this task).

[96] *See* John Pavlus, *What NASA Could Teach Tesla about Autopilot's Limits*, SCI. AM. (July 18, 2016), https://www.scientificamerican.com/article/what-nasa-could-teach-tesla-about-autopilot-s-limits/ (outlining the debate resulting from the first fatal accident involving Tesla's autopilot).

by IHL. The next question in the inquiry, then, is: What aspects of AI could help AWS achieve (or fall short) of these standards?

4. How AI Works in Practice

The science of AI that enables the design of systems that can behave rationally based on objective performance standards focuses on the programming of an intelligent "agent." A few terms of art must be explained at this juncture. "An agent is something that perceives and acts in an environment."[97] Think of this as the complete system—the physical platform mated with its computer hardware and software. The "agent function" describes what the agent will do in response to given inputs.[98] The function could range from simple "if/then" logic to non-deterministic functions based on complex algorithmic processes. "Agent programs" implement the agent function.[99] This is the software programmed into the system that is designed to achieve the desired (rational and objective) outcome. Programs are tailored to respond to the unique environment in which the agent operates. Agents can also be granted, via programming, the ability to "learn" through their perceptions by adjusting their behavior in order to better achieve assigned goals.

Consider the simple case of the Roomba.[100] The Roomba is a small vacuum that is designed to clean floors without direct physical manipulation by a human. This agent uses a suite of sensors that helps the agent function determine how to navigate and locate objects so as to most efficiently vacuum, then "remembers" where those obstacles were and avoids them in the future. The agent "knows" when its battery is low and the agent program guides the Roomba back to a charging station. The Roomba also can detect whether it is cleaning carpets or hardwood and adjust power output accordingly. With these capabilities one could argue that the Roomba is a fully autonomous and artificially intelligent learning robot.

That said, it is important to avoid the temptation of ascribing broad and unrealistic capabilities to AI simply because a system is quite advanced in certain narrow respects. Careful attention must be paid to what the AI *cannot* do. Despite some impressive features, the Roomba is constrained in what it can achieve due to its limited agent platform, sensors, and agent program.[101] In other words, if you

[97] *See* RUSSELL & NORVIG, *supra* note 36, at 59.

[98] *See id.* at 35 ("Mathematically speaking, we say that an agent's behavior is described by the agent function that maps any given percept sequence to an action.").

[99] *See id.* at 59 ("The agent program implements the agent function.").

[100] *See generally Roomba Robot Vacuum*, IROBOT.COM, https://www.irobot.com/For-the-Home/Vacuuming/Roomba.aspx (last visited Mar. 29, 2017) (describing the functions of the device).

[101] The Roomba is limited not only by its physical platform but also by its software. It employs deterministic software that controls its actions based on "if/then" rationale. For example, the Roomba's software tells it, in essence, that "if I have already vacuumed a location, I do not

expect it to do anything other than vacuum the floor, you will be disappointed. More advanced but nevertheless relatively narrow AI-enabled machines are more illustrative and logically relatable to AWS. For example, computers have beaten the best human opponents in complex games such as chess,[102] *Go*,[103] and the television game-show *Jeopardy*.[104] While these achievements are impressive, it is important to note the narrow operating environment in which these agents functioned. Simply because an agent is adept at playing a very complex game it does not necessarily follow that its AI is sophisticated enough to conduct high-level reasoning comparable to humans.[105]

As a result, some who criticize the introduction of AI into weapon systems argue that such narrowness in current AI means that it could never be sophisticated enough to handle the complexities of the modern battlefield.[106] Others argue that future AWS will need to cope with decision-making speeds that are beyond human capacity because future wars will be fought at "machine speed."[107] Many agree, however, that regardless of the resolution of these issues,

vacuum it again" and "if my battery reaches *X* percent, I return to my charging station." The vacuum cannot perceive or respond to environments outside of its programming.

[102] *See Icons of Progress*, *Deep Blue*, IBM, http://www-03.ibm.com/ibm/history/ibm100/us/en/icons/deepblue/ (last visited Mar. 29, 2017) ("On May 11, 1997, an IBM computer called IBM Deep Blue beat the world chess champion after a six-game match.").

[103] *See AlphaGo*, DEEPMIND.COM, https://deepmind.com/research/alphago/ (last visited Mar. 29, 2017) ("In March 2016 AlphaGo won 4-1 against . . . the top Go player in the world over the past decade.").

[104] *See* John Markoff, *Computer Wins on 'Jeopardy!': Trivial, It's Not*, N.Y. TIMES (Feb. 16, 2011), http://www.nytimes.com/2011/02/17/science/17jeopardy-watson.html (describing the win); *see also Watson*, IBM, https://www.ibm.com/watson/ (last visited Mar. 29, 2017) (describing the current industry applications for Watson AI technology).

[105] Certain games such as chess are relatively complex yet have a finite number of possible moves. This type of game plays to the strengths of an AI because it can evaluate all possible moves and counter-moves. Other games do not have a finite set of actions. For example, in no-limit Texas hold 'em poker, a participant can bet any amount of money. Still, AI can compensate by learning how to bluff in ways that are unpredictable to a human and thereby prevail. *See* Avery Thompson, *An AI Just Crushed Poker Pros at Texas Hold 'Em*, POPULAR MECHANICS (Jan. 31, 2017), http://www.popularmechanics.com/technology/a24989/ai-wins-texas-hold-em/. But in all of these games, the AI uses experience to refine its actions in order to achieve optimal results. It has no higher order appreciation for the behavior of its opponents outside the narrow parameters of the game itself.

[106] *See Losing Humanity*, *supra* note 1, at 30 (concluding, for example, that yet unrealized "fully autonomous weapons. . .would appear to be incapable of abiding by the key principles of international humanitarian law").

[107] *See* Cheryl Pellerin, *Work: Human-Machine Teaming Represents Defense Technology Future*, U.S. DEP'T OF DEF.: DOD NEWS (Nov. 8, 2015), https://www.defense.gov/News/Article/Article/628154/work-human-machine-teaming-represents-defense-technology-future (quoting Deputy U.S. Defense Secretary Bob Work: "Learning machines . . . literally will operate at the speed of light. So when you're operating against . . . [rapidly developing attacks], you [need] . . . a learning machine that helps you solve that problem right away."); *see also* U.S. DEP'T OF DEF., 14-S-0553, UNMANNED SYSTEMS INTEGRATED ROADMAP FY 2013-2038 67 (2012), http://archive.defense.gov/pubs/DOD-USRM-2013.pdf ("Autonomy in unmanned systems will be critical to future conflicts that will be fought and won

the sophistication of AI will need to advance in order for AWS to account for or mitigate these realities. The implication of this debate is that in the future, AWS will likely be called upon to "learn" in order to handle complex and changing environments.[108]

5. How Machines Learn

Agents learn by being provided data sets from which an onboard algorithm can be programmed to attain rational goals. The agent may then be placed in an unknown environment in which it will draw upon its learning data sets and endeavor to achieve optimal results. A learning agent will continue to refine its behavior based on the results that it achieves in real-world operating environments as compared to established goals.[109]

Learning agents are generally comprised of a performance element and a learning element. The performance element senses the environment of the agent and determines a course of action,[110] while the learning element employs feedback from the system "on how the agent is doing and determines how the performance element should be modified to do better in the future."[111] Machine learning in artificially intelligent "agents can be summarized as a process of modification of each component of the agent to bring the components into closer agreement with the available feedback information, thereby improving the overall performance of the agent."[112]

By way of basic example, consider a hypothetical stealth drone equipped with learning AI that helps the system avoid detection by enemy radar. The performance element of the drone determines the heading, speed, and altitude at which it will fly. Suppose the drone begins its journey into enemy territory by flying at a high altitude. Enemy radar then detects the craft despite its stealthy design. A "critic"[113] will inform the learning element that the enemy radar is

with technology."); David Ignatius, *The exotic new weapons the Pentagon wants to deter Russia and China*, WASH. POST (Feb. 23, 2016), https://www.washingtonpost.com/opinions/the-exotic-new-weapons-the-pentagon-wants-to-deter-russia-and-china/2016/02/23/b2621602-da7a-11e5-925f-1d10062cc82d_story.html?utm_term=.6c8434d4ab06 (describing the DoD's "Third Offset" strategy to use technology as a strategic deterrent); Missy Cummings, Professor, Dept. of Mech. Eng'g and Materials Sci., Duke Univ., Comment at the Georgetown Univ. Law Center's Nat'l Sec. Law Soc'y Panel Discussion: Legal and Ethical Implications of Autonomous Weapons (Apr. 4, 2013) (describing "human neuromuscular lag" and why humans could never be as fast as robots in completing some tasks).

[108] AWS that fight at machine speed and can adapt their behavior quicker than either humans or enemy AWS will arguably be at a natural advantage in future conflict.

[109] *See* RUSSELL & NORVIG, *supra* note 36, at 55 ("Learning has another advantage . . . it allows the agent to operate in initially unknown environments and to become more competent than its initial knowledge alone might allow.").

[110] *See id.* ("The performance element . . . takes in percepts and decides on actions.").

[111] *Id.*

[112] *Id.* at 57.

[113] The "critic" is a portion of the agent that provides feedback to the learning element. *See id.* at 55. The system may also employ a "problem generator" in order to devise novel means by which

targeting the drone, and the learning element may then inform the performance element that flying at a high altitude is not optimal. The performance element flies the drone to a lower altitude, and it does not again return to higher altitudes. The machine has learned.

In order to fully grasp how intelligent systems learn about the world into which they are placed, it is also necessary to appreciate how agents are programmed to view their environment. The most basic agents are only able to observe their environment in binary terms, in what is referred to as an "atomic" representation of the world.[114] For example, to return to the Roomba hypothetical, assume that the agent is only able to ascertain two states of the world, where the floors are either "clean" or "dirty." The agent does not consider other factors, such as whether the homeowner might be annoyed by the presence of the Roomba. If the floor is dirty, the Roomba will endeavor to make it clean. There are no other factors for the Roomba to weigh in the calculus.

But suppose we change the hypothetical such that the Roomba is able to measure the tone of the homeowner's voice. Also, what if it knew that houseguests were visiting who might not appreciate its presence? There are a litany of other factors that might bear on the Roomba's ultimate "to clean or not to clean" decision. If the agent were able to sense and weigh these factors, it would be considering either a "factored" or "structured" representation of its environment. When an agent uses a factored representation of its world, it can consider a range of variables and attributes that have discernible values.[115] In a structured representation, the agent is able not only to consider these variables but also to weigh the inter-relationship between them.[116]

A machine's ability to learn using factored and structured representations of the world may be crucial to the application of AI to AWS for two reasons. First, AWS that are fielded will often need to consider multiple variables regarding their operating environment. In certain circumstances, the AWS will need to be able not only to weigh numerous variables that it senses, but it might also need to consider the way in which those variables relate to others. These considerations will continue to impact a wide swath of actions by AWS, such as navigation, object recognition, and fire-control solutions. Second, and more

the drone can achieve its goal of remaining undetected. *See id.* at 56. The "problem generator" is "responsible for suggesting actions that will lead to new and informative experiences." *Id.* It forces the agent to consider courses of action that might appear sub-optimal in the near term but are more successful in attaining long-term goals. *Id.*

[114] *See id.* at 57 ("In an atomic representation each state of the world is indivisible—it has no internal structure . . . [it is a] state of the world . . . whose only discernible property is that of being identical to or different from another [state of the world].").

[115] This means, for example, that the Roomba could consider the presence of guests and the attitude of its owner in determining whether or not it would set about its mission. *See id.* at 56 ("A factored representation splits up each state into a fixed set of variables or attributes, each of which can have a value.").

[116] In other words, the Roomba would be able to ascertain whether the presence of the guests affected the mood of the owner and vice versa.

importantly with regard to future AWS, an agent's ability to effectively represent its environment and account for complex interrelated variables while learning may be pivotal in handling uncertainty, as well as ensuring predictability.

It is important, however, not to overstate this point. While future AWS will likely need to be programmed to consider factored or structured representations of their operating environment, it does not follow that they will always need to consider *all* of the factors a human would consider or be able to make fine-grained judgments about how each relevant variable in the environment relates to the others. This was the fatal assumption of *Losing Humanity*,[117] as it envisioned a world where AWS would necessarily be called upon to directly substitute for human soldiers in the most complex battlespaces.[118] The simple resolution to this issue has three aspects. First, for most scientists experienced in AI, it is nearly impossible to rationally envision that machines will possess the technological capability to make such complex and subjective decisions, even if equipped with highly sophisticated AI.[119] Second, machines are not compelled to act under the same time constraints as humans because they have no self-preservation instinct.[120] Third, if a military commander is not reasonably certain that a weapon system will comply with IHL as employed, he is under a positive obligation not to use it.[121] This is true for any weapon or system, past, present, and future.

That said, learning in intelligent systems that are provided lethal capability nevertheless raises significant issues that must be carefully scrutinized. This is because the further we move from deterministic ("if/then") systems and towards

[117] *Losing Humanity*, *supra* note 1.

[118] By way of an example from *Losing Humanity*, soldiers kicking down doors to residential homes must make split-second "shoot/no shoot" decisions based on determinations of hostile intent. A soldier may use lethal force to counter a hostile act, which is a use of force directed at the soldier or his unit and which is already underway. Hostile intent is inherently more subjective than a hostile act because, in essence, one is trying to predict something that is *about* to happen. With respect to U.S. military forces, hostile intent is defined in relevant part as the "[t]he threat of imminent use of force against . . . U.S. forces." Chairman of the Joint Chiefs of Staff Instruction 3121.01B, Standing Rules of Engagement/Standing Rules for the Use of Force for US Forces app. A-3 (June 13, 2005). The approach of *Losing Humanity* falls into a logical trap described below. *See infra* note 134.

[119] *See* DARPA Interview, *supra* note 73; interview with Leslie Pack Kaelbling, *supra* note 70; NUWC Interview, *supra* note 70; interview with Paul Scharre, *supra* note 41; interview with Alan C. Schultz, *supra* note 68.

[120] Unless of course they are programmed to protect themselves. An AWS could be limited by programming, however, only to use lethal force in response to a hostile act, as opposed to attempting to discern intent. Indeed, this would be a more reasonable requirement for a robot because it does not fear death. The primary reason we allow combatants to use deadly force in response to demonstrated hostile intent is to preserve life, and secondarily to preserve military assets. With an autonomous system, the primary concern is alleviated.

[121] *See* API, *supra* note 89, at art. 57 ("[T]hose who plan or decide upon an attack shall . . . take all feasible precautions in the choice of means and methods of attack with a view to avoiding, and in any event minimizing, incidental loss of civilian life, injury to civilians and damage to civilian objects.").

complex learning systems, the less reasonably we may be able to predict how the system will arrive at given solutions.[122] This may or may not present issues with IHL compliance; it depends on which puzzle-shaped pieces of the OODA loop the system has been granted. Each time that a new AWS is desgined, we must evaluate whether the learning capacity that it is granted will prevent the AWS from being employed in conformance with IHL.

One might reasonably ask, then, why we would grant an AWS any learning capacity in the first place. Why not simply design the system to account for all of the difficulties that it might encounter in its operating environment? Depending on the environment and the mission of the system, this goal may be impossible. The reasons for this are threefold:

> First, the designers cannot anticipate all possible situations that the agent might find itself in. For example, a robot designed to navigate mazes must learn the layout of each new maze it encounters. Second, the designers cannot anticipate all changes over time; a program designed to predict tomorrow's stock market prices must learn to adapt when conditions change from boom to bust. Third, sometimes human programmers have no idea how to program a solution themselves. For example, most people are good at recognizing the faces of family members, but even the best

[122] This is especially so in more advanced learning systems such as artificial neural networks, which in their most basic sense attempt to mimic the ways in which biological neurons in the human brain function. *See A Basic Introduction to Neural Networks*, UNIVERSITY OF WISCONSIN-MADISON COMPUTER SCIENCE DEP'T, http://pages.cs.wisc.edu/~bolo/shipyard/neural/local.html (last visited Dec. 23, 2016); *see also Unsupervised Feature Learning and Deep Learning Tutorial: Convolutional Neural Network*, STANFORD UNIVERSITY, http://ufldl.stanford.edu/tutorial/supervised/ConvolutionalNeuralNetwork/ (last visited Dec. 23, 2016) (modeled after the visual cortex of animals, these networks are "comprised of one or more convolutional layers . . . and then followed by one or more fully connected layers as in a standard multilayer neural network"). Artificial neural networks "remain one of the most popular and effective forms of learning system." RUSSELL & NORVIG, *supra* note 36, at 728. Other cutting-edge technologies may further complicate the question of predictability. For example, polymorphic networks modify their own code while keeping algorithms intact. *See Glossary: Polymorphic*, SYMANTEC, https://www.symantec.com/security_response/glossary/define.jsp?letter=p&word=polymorphic (last visited Dec. 27, 2016) ("Polymorphic malicious code generates functionally equivalent but distinct copies of itself when it replicates, in the hopes that pattern matching security tools won't be capable of detecting it, as there is little or no stable pattern of code to match against."). Transfer learning may also create seams in predictability. *See, e.g.*, "DEFCON 24" Convention, Las Vegas, Nev. (Aug. 4–7, 2016) ("Common machine learning algorithms . . . traditionally address isolated tasks. Transfer learning attempts to change this by developing methods to transfer knowledge learned in one or more source tasks and use it to improve learning in a related target task.") (notes on file with author). When machine learning is exposed to adversarial environments, additional issues arise with respect to predictable performance. *See* Pavel Laskov & Richard Lippmann, *Machine Learning in Adversarial Environments*, MIT LINCOLN LABORATORY (June 28, 2010), http://llwww.ll.mit.edu/mission/cybersec/publications/publication-files/full_papers/2010_10_25_Lippmann_MLJ_FP.pdf (noting that when machine learning is employed in hostile environments, "adversaries consciously act to limit or prevent accurate performance").

> programmers are unable to program a computer to accomplish that task, except by using learning algorithms.[123]

In essence, then, if we develop AWS that will be able to fight at machine speed in future conflicts by leveraging learning technologies in AI, then we must accept that some aspects of these systems will adapt in ways that we may be unable to predict. As such, we must delineate the implications of this increased uncertainty.

D. *Within the Learning-enabled AI, the Analytical Focus is on Predictability*

Given the concerns over AI and machine learning, how do we prevent ourselves from functionally delegating those decisions (e.g., the decision to kill) that we may not delegate? Predictability is the key.[124] The analysis here focuses on aspects of a system that might, in combination, affect our ability to reasonably predict its compliance with IHL. It cannot be overstated, however, that not all aspects of the system must be predictable. There is of course great potential military advantage to be gained by providing advanced machine learning, for example, to aspects of a machine that either do not bear on IHL compliance or do not combine with other autonomous features to functionally delegate the decision to kill.

Like most IHL requirements, our ability to predict the actions of the machine must be based on a reasonableness standard. The test of reasonableness emanates from the recognition under IHL that attaining even near-certainty in armed conflict is usually an insurmountable goal.[125] From a practical standpoint, a lower standard would encourage noncompliance with IHL by inviting humans simply to blame erratic computers for violations. A higher standard would likely be unattainable based on the complexity of computer programming magnified by the "fog" of the modern battlefield. For a variety of policy reasons, states may

[123] RUSSELL & NORVIG, *supra* note 36, at 693.

[124] *See* Alan L. Schuller, *Focusing the debate on autonomous weapon systems: A new approach to linking technology and IHL*, *in* Meeting Report, Int'l Comm. of the Red Cross, *supra* note 61, at 27.

[125] *See* CUSTOMARY INTERNATIONAL HUMANITARIAN LAW, *supra* note 17, at 558 (suggesting that commanders are responsible for war crimes committed by subordinates if they do not take "necessary and reasonable measures in their power to prevent their commission"); *see also id.* at 563 ("[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.") (citing Prosecutor v. Delalić et al., Case No. IT-96-21-T, Judgment, ¶ 395 (Int'l Crim. Trib. for the former Yugoslavia Nov. 16, 1998)); *see also* API, *supra* note 89, at art. 57(4) ("In the conduct of military operations at sea or in the air, each Party to the conflict shall . . . take all reasonable precautions to avoid losses of civilian lives and damage to civilian objects."); Prosecutor v. Stanislav Galic, Case No. IT-98-29-T, Judgment and Opinion, ¶ 58 (Int'l Crim. Trib. for the former Yugoslavia Dec. 5, 2003) ("In determining whether an attack was proportionate it is necessary to examine whether a reasonably well-informed person in the circumstances of the actual perpetrator, making reasonable use of the information available to him or her, could have expected excessive civilian casualties to result from the attack.").

want to set higher bars for their own introduction of certain AWS into operation.[126] As a matter of law, however, the reasonableness standard is a well-established benchmark of performance that has balanced the competing interests of IHL for quite some time.

Predictability cannot diminish past the point where we can reasonably say a human was in control of compliance with IHL. Importantly, this is not the same standard as physical human control over the actions of the machine itself at the time of lethal kinetic action.[127] It also does not mean that a human made a call on IHL compliance that was temporally proximate to a lethal attack. Rather, it means that we can reasonably predict what action the system will take and that we are reasonably certain that the system will comply with IHL. If we can reasonably predict compliance, then we maintain control no matter the level or type of our interaction with the machine at the instant of lethal action. But if we cannot reasonably predict whether the machine will comply with IHL, it may be unlawfully autonomous.

III. Machine Learning and Predictability

A. *Uncertainty in AI and the Pitfalls of Unpredictability*

AWS must be designed to account for at least some of the uncertainty deriving from complex operating environments. In order to evaluate the ability of a system to meet this requirement we must understand from a technical perspective how uncertainty is handled. Generally speaking, AI accounts for uncertainty by weighing the probability of certain outcomes against the desirability of such outcomes. The ability of the system to meet performance goals in the face of uncertainty is a key variable in these calculations.

When we ask whether a particular system equipped with AI is lawful *per se* and whether it can comply with IHL as employed on the battlefield, we must necessarily inquire into whether it can meet technical performance standards.[128] We know that these performance standards must be based on rationally described goals or endstates.[129] In controlled laboratory environments, rational performance

[126] For example, states may require predictability to a "near certainty." *See generally* THE WHITE HOUSE, REPORT ON THE LEGAL AND POLICY FRAMEWORKS GUIDING THE UNITED STATES' USE OF MILITARY FORCE AND RELATED NATIONAL SECURITY OPERATIONS 25 (2016), https://www.justsecurity.org/wp-content/uploads/2016/12/framework.Report_Final.pdf (explaining that, in the context of targeting, "the United States must have 'near certainty' that the terrorist target is present and that non-combatants will not be injured or killed" before lethal action is permitted).

[127] To be clear, this Article does not argue that a human must be "in the loop" and either affirmatively approve or choose not to override the system's decision just prior to lethal kinetic action. *Contra* Malinowski, *supra* at note 7 (arguing that while there may be advantages to automation, humans should always be in the decision-making loop in order to approve or deny lethal decisions).

[128] *See supra* Part II.A.

[129] *See supra* Part II.B.3.

standards might be easy to describe. The environment in which the AI is called upon to operate could be quite simplistic. This makes the task of designing and testing that sort of AI easier. But what about on a battlefield?

Even the most straightforward combat scenarios are often immensely more complex than a laboratory environment.[130] The challenges computer programmers face in designing AI that can handle uncertain environments are significant. A program that is designed to handle all possible eventualities presented even in environments of relatively limited uncertainty could require impossibly large data sets.[131] This has led some groups to conclude that AI could never function effectively in combat while still adhering to IHL standards.[132] The weakness of such critiques, however, is that they assume too much about how AI will function on the future battlefield. In fact, most people—regardless of their ideological or institutional biases—assume far too much about what roles machines might fulfill and how they might go about fulfilling them.[133] The root cause of these assumptions is our collective concern over how future AI will handle uncertainty because we worry that machines will perform unpredictably.

But science fiction aside, even the most advanced computers and cutting edge AI should perform to some level of predictability. If we carefully instruct computers how to account for and respond to uncertainty, we should in theory be able to predict IHL compliance (or lack thereof) to a reasonable certainty.[134] Again, this will depend heavily on the pieces of the decision-making loop that are delegated to computers, precisely which AI technology is incorporated, and which physical capabilities are granted to the system.[135]

The uncertainty that AI-enabled AWS must be able to handle derives from operating environments that are only partially observable and/or

[130] One of the challenges AI developers face is establishing training and test data that will adequately replicate the experiences faced by an agent outside of the laboratory. This is true even in comparatively simple applications, such as driverless cars. *See* Interview with Leslie Pack Kaelbling, *supra* note 70 (describing the challenges). It is difficult to overstate the complexities of a traditional battlefield, and even more so when contemplating future conflict areas.

[131] *See* RUSSELL & NORVIG, *supra* note 36, at 480 ("When interpreting partial sensor information, a logical agent must consider *every logically possible* explanation for the observations, no matter how unlikely. This leads to impossibly large and complex belief-state representations.").

[132] *See Losing Humanity*, *supra* note 1; *see also Advancing the Debate*, *supra* note 1.

[133] *See* Interview with Alan C. Schultz, *supra* note 68 (taking the position that machines will more likely *team* with humans on future battlefields than replace them). For example, if we imagine robots in the future playing a game of soccer, most people conjure images of what amounts to metal humanoids on a traditional pitch. But robots might fare better on a hard surface rather than a grass field. Also, a lower center of gravity than a human could be advantageous. Instead of legs, the robots could use flappers like a pinball machine. They might be painted instead of wearing uniforms. The ball might be hard instead of soft. If the machines were programmed properly, referees might not be needed to adjudge penalties. In short, machines will continue to change the way "the game" is played.

[134] *See* Interview with Alan C. Schultz, *supra* note 68 (supporting this proposition).

[135] These general concepts are explored via hypotheticals in Part V, *infra*.

nondeterministic.[136] Programming a computer with vast data sets that can account for every possible outcome to even simple problems is often mathematically infeasible, factually impossible, or undesirable due to other mission constraints or restraints.[137] Thus, AI in the context of AWS must necessarily account for potentially immense uncertainty while achieving a desired end state. Uncertainty in AI manifests itself primarily in two ways: first, it is sometimes infeasible or impossible to establish exceptionless rules for the system to follow; second, the system may be ignorant regarding some aspects of its operating environment.[138]

So how does AI account for this uncertainty? It does so by linking a computer's decisions to the probability of certain outcomes and the utility of such outcomes. "Probability provides a way of summarizing the uncertainty that comes from our laziness and ignorance."[139] Utility theory establishes "preferences between the different possible outcomes of the various plans."[140] "The right thing to do—the rational decision—therefore depends on both the relative importance of various goals and the likelihood that, and degree to which, they will be achieved."[141] An AI is considered to make rational decisions "if and only if it chooses the action that yields the highest expected utility, averaged over all the possible outcomes of the action."[142] Simply put, humans must give computers decision-making priorities based on levels of certainty. We must program the computer to achieve the "best" possible outcome under the circumstances.

The notion that AI must acknowledge and account for uncertainty in its programming is important in the evaluation of AWS. The sophistication of the programming in its ability to account for uncertainty on the battlefield determines whether the system can comply with IHL. The AI might possess crude and unsophisticated deterministic software, a more sophisticated Bayesian network,[143]

[136] RUSSELL & NORVIG, *supra* note 36, at 480.

[137] *See id.*

[138] The context in which AWS are likely to operate amplifies uncertainty for a variety of reasons:

> Trying to use logic to cope with [complex domains] . . . fails . . . [due to] . . . Laziness: It is too much work to list the complete set of antecedents or consequents needed to ensure an exceptionless rule and too hard to use such rules; Theoretical ignorance: . . . science has no complete theory for the domain; Practical ignorance: Even if we know all the rules, we might be uncertain [about a given circumstance due to our inability to sense or account for all variables].

Id. at 481.

[139] *Id.* at 482. Laziness is the inability to define an exceptionless rule. Ignorance is the inability to know all particular applications of a rule. *Id.*

[140] *Id.*

[141] *Id.*

[142] *Id.* at 483. "Expected" refers to the average or statistical mean of the outcomes, weighted by the probability of the outcome. *Id.*

[143] Bayesian means "[b]eing, relating to, or involving statistical methods that assign probabilities or distributions to events (as rain tomorrow) or parameters (as a population mean) based on experience or best guesses before experimentation and data collection and that apply Bayes' theorem to revise the probabilities and distributions after obtaining experimental data." *Bayesian*,

or some unforeseen technology that surpasses what we can currently envision. The bottom line is that we must be able to account for uncertainty in the programming of the system such that we can reasonably predict it will comply with IHL despite the inherent complexities of combat. Depending on the specific characteristics of the system, this prediction could be simple, impossible, or somewhere in between.

B. *Predictability and Uncertainty in AWS: Tracing Decisions to Kill Back to a Human*

Robots seem frightening because, despite their potential power, they inherently lack the context from which to discern the appropriate application of force. They do not know *anything*, except what they are told through programming. Suppose a human being was born at age 21, fully grown and strong but lacking the knowledge that the average adult would have gained during their two decades of life. That person would know nothing about social norms, rules, or consequences. Babies are not threatening because they are powerless to act on their irrational and undeveloped thoughts. A full-grown person with the temperament of a toddler, however, would be very dangerous. Then again, even the "newborn adult" described above could process the emotions of other humans and begin adapting its behavior. The same is not necessarily so for a machine.

But this conundrum is mitigated by the fact that, unlike humans, machines do not possess free will. As discussed in Part II.B.5, they generally do as they are programmed. They handle uncertainty in the way we tell them to handle it. Their learning is bounded by the ways we tell them to learn. We decide if they have a neural network, no network, or a *Speak & Spell* for a processor.[144] As such, the responsibility lies with those designing AWS to account for uncertainty, both internal to the system and in handling its external environment, in a responsible manner that ensures we can reasonably predict IHL compliance.

With respect to environmental uncertainty, we may safely assume that AWS will not be able to sense and consider all of the variables present on the battlefield that humans might consider.[145] But AI within AWS will not necessarily *need* to consider all the same factors that a human might in order to be lawful. Instead, AWS may be able to compensate for a lack of situational awareness in certain respects through other capabilities and/or limitations. For example, a

MERRIAM-WEBSTER.COM, http://www.merriam-webster.com/dictionary/Bayesian (last visited Dec. 22, 2016).

[144] *See History of Innovation*, TEX. INSTRUMENTS http://www.ti.com/corp/docs/company/history/timeline/eps/1970/docs/78-speak-spell_introduced.htm (last visited Mar. 29, 2016) (describing the *Speak & Spell* as the first time the "human vocal tract had been electronically duplicated on a single chip of silicon").

[145] *See* DARPA Interview, *supra* note 73; interview with Paul Scharre, *supra* note 41; interview with Alan C. Schultz, *supra* note 68; interview with Lincoln Laboratories autonomous systems specialists, Mass. Inst. of Tech., Lexington, Mass. (Sept. 16, 2016) [hereinafter Lincoln Laboratories Interview].

system may not need to complete the complex task of discerning hostile intent because it need not act out of self-preservation.[146] If the AWS is either expendable or well hardened,[147] the system could be restricted to responding only to hostile acts, or forbidden entirely from acting in its own defense. The evaluation of uncertainty and how it is handled by the system should not be confused with the ways in which a human would conduct the process.

Our technological inquiry into whether an AWS handles uncertainty acceptably hinges on whether the actions of the system that result in the loss of human life are predictable enough to be traced back to a human decision to attack a target or class of targets. There are infinite ways in which puzzle pieces from the OODA loop could be stitched together to demonstrate that the decision to kill was reasonably made by a human.[148] Removal or insertion of any given piece could make the difference in determining whether or not a human retained control of this decision and a concurrent IHL evaluation. We must therefore establish broad principles that will allow us to avoid the development of unlawful autonomy in weapon systems. A brief review of the basic IHL principles is first required.

IV. Overview of International Humanitarian Law

The principles of IHL guide the conduct of belligerent parties at all times.[149] They constitute the foundation for how military forces prosecute lethal attacks.[150] The principles are general in nature, and they have withstood the test of time and endured through the development of innumerable means and methods of destruction.[151] The IHL principles are: military necessity, distinction, proportionality, and preventing unnecessary suffering.

The principle of military necessity holds that if a target is "indispensable for securing the complete submission of the enemy as soon as possible,"[152] and an

[146] *See* discussion *supra* note 120.
[147] With, for example, protective armor.
[148] Again, the decision need not occur in temporal proximity to lethal action by an AWS.
[149] The law of war principles have attained the status of customary international law. *See generally* CUSTOMARY INTERNATIONAL HUMANITARIAN LAW, *supra* note 17 (describing the customary IHL principles)*; see also* GARY D. SOLIS, THE LAW OF ARMED CONFLICT 250 (Cambridge Univ. Press 2010) ("[T]he core LOAC/IHL principles . . . bind every armed force.").
[150] *See, e.g.*, U.S. Dep't of Def. Directive 2311.01E, DoD Law of War Program ¶ 4.1 (May 9, 2006, incorporating Change 1, Nov. 15, 2010) ("Members of the DoD Components comply with the law of war during all armed conflicts, however such conflicts are characterized, and in all other military operations."); *see also* U.S. DEP'T OF ARMY, FIELD MANUAL 27-10, THE LAW OF LAND WARFARE app. A-1 (1956) (IHL seeks to "diminish the evils of war by: *a.* Protecting both combatants and noncombatants from unnecessary suffering; *b.* Safeguarding certain fundamental human rights of persons who fall into the hands of the enemy, particularly prisoners of war, the wounded and sick, and civilians; and *c.* Facilitating the restoration of peace").
[151] *See* SOLIS, *supra* note 150, at 250 ("Despite the codification of much customary law into treaty form during the last one hundred years, four fundamental principles still underlie the law of armed conflict.") (internal citations omitted).
[152] FIELD MANUAL 27-10, *supra* note 150, at 4.

attack upon it is not otherwise illegal, then it is a valid target.[153] Targets are persons and objects "which by their nature, location, purpose, or use make an effective contribution to military action" and whose destruction or neutralization "offers a definite military advantage."[154] Military necessity does not justify targeting something that is otherwise illegal, and it is not a defense to a violation of IHL.[155]

The principle of distinction holds that belligerents may only attack targets that are valid military objectives.[156] As a subset of the distinction principle, the requirement to take precautions in the attack mandates that a belligerent take active steps to determine whether persons and groups are civilians or combatants and to direct operations only against combatants.[157] This is an affirmative duty on the part of the belligerents. Civilians and their property are of course generally protected from attack.[158]

Proportionality under IHL means that the anticipated loss of civilian life and damage to property incidental to attacks must not be "excessive in relation to the concrete and direct military advantage anticipated."[159] Thus, collateral damage to civilian personnel and property incurred while attacking a military objective may not be disproportionate to the advantage gained.[160] Military advantage is not

[153] Attacks that would produce a "concrete and direct military advantage," and are not otherwise unlawful, are not prohibited. *See* API, *supra* note 89, at art. 51(5)(b); CUSTOMARY INTERNATIONAL HUMANITARIAN LAW, *supra* note 17, at R. 14.

[154] API, *supra* note 89, at art. 52(2); *see also* CUSTOMARY INTERNATIONAL HUMANITARIAN LAW, *supra* note 17, at R. 8.

[155] *See* MARCO SASSÒLI ET AL., HOW DOES LAW PROTECT IN WAR? 387 (3d ed. 2014) ("Necessity . . . and self-defence are not circumstances precluding the wrongfulness of IHL violations."); *see also* THE U.S. ARMY JUDGE ADVOCATE'S LEGAL CENTER AND SCHOOL, OPERATIONAL LAW HANDBOOK 11 (David H. Lee ed., 2015) [hereinafter OPERATIONAL LAW HANDBOOK] (noting that military necessity is "not a criminal defense" and does not justify otherwise unlawful acts, but rather "must be applied in conjunction with other LOAC principles"). As such, it is more helpful to think of military necessity as a restraint rather than a permissive concept. In other words, because of the principle of military necessity, a target may not be attacked unless there is a concrete military reason to do so. And, even if there is a military reason to attack it, the other IHL requirements must still be complied with.

[156] *See* API, *supra* note 89, at art. 52(2) ("Attacks shall be limited strictly to military objectives."); CUSTOMARY INTERNATIONAL HUMANITARIAN LAW, *supra* note 17, at R. 7; *see also* SOLIS, *supra* note 150, at 258 (noting that the distinction principle is considered customary international law).

[157] *See* API, *supra* note 89, at art. 57; CUSTOMARY INTERNATIONAL HUMANITARIAN LAW, *supra* note 17, at R. 15–21; *see also* OPERATIONAL LAW HANDBOOK, *supra* note 156, at 12 ("[P]arties to a conflict must direct their operations only against combatants and military objectives.").

[158] *See* API, *supra* note 89, at art. 51–52. However, there are some exceptions to this rule. For example, civilians may lose their protected status based on their actions, such as taking *direct part in hostilities*, in which case they may be targeted for such time as they do so. *See id.* at art. 51(3); *see also* CUSTOMARY INTERNATIONAL HUMANITARIAN LAW, *supra* note 17, at R. 6.

[159] API, *supra* note 89, at art. 51(5)(b); *see also* CUSTOMARY INTERNATIONAL HUMANITARIAN LAW, *supra* note 17, at R. 14.

[160] *See* Rome Statute of the International Criminal Court art. 8(2)(b)(iv), July 17, 1998, 2187 U.N.T.S. 90 ("Intentionally launching an attack in the knowledge that such attack will cause incidental loss of life or injury to civilians or damage to civilian objects . . . which would be clearly excessive in relation to the concrete and direct overall military advantage anticipated" is a

weighed based on tactical gains alone. The "expected advantage should be seen in relation to the attack as a whole,"[161] and it is linked to the full strategic and operational context of the attack.

With respect to combatants in a conflict, the prohibition against causing unnecessary suffering prohibits the use of weapons that *by their nature* cause unnecessary suffering and the use of lawful weapons in a manner that is intended to cause unnecessary suffering.[162] There is no simple test to determine whether the use of a weapon would constitute unnecessary suffering.[163] Nonetheless, as a matter of policy, the U.S. Department of Defense (DoD) reviews every weapon in its inventory to ensure that it does not by its nature cause unnecessary suffering or otherwise violate IHL or other laws.[164] Many other countries do the same, either voluntarily or out of treaty obligations.[165] With these IHL tenets in mind, we may now describe principles that inform the responsible development of AWS.

V. Five Principles: Avoiding Unlawful Autonomy

A. *Principle 1: The decision to kill may never be functionally delegated to a computer*

Machines will not develop human-like cognitive qualities any time soon. As such, the discussion about machine decision-making must be focused on the potential for *functional* delegation of the decision to kill. The question of whether or not a target or class of targets can be attacked under the given conflict rubric is inherently a human burden. For this reason, humans must retain control over adequate fragments of the OODA loop. We cannot predict the weapons technology that might exist in the future, but we must ensure that within the OODA construct, as applied to targeting decisions, every future weapon system retains sufficient human input such that the decision to kill is not functionally delegated. This is admittedly the most nebulous aspect of these principles. It is

war crime under the statute.). During American military operations under offensive rules of engagement, the collateral damage estimation (CDE) methodology ensures that attacks conform with this principle. *See generally* Chairman of the Joint Chiefs of Staff Instruction 3160.01, No-Strike and the Collateral Damage Estimation Methodology (Feb. 13, 2009) (establishing the CDE process). Proportionality in the IHL context, as opposed to the self-defense context of proportionality of force, does not mean that an attacker must use the same type of weapons or force as the enemy. If a target is a lawful one, it may be attacked with any weapon in the military inventory, provided the attack is otherwise lawful. *See* SOLIS, *supra* note 150, at 280 (describing common misunderstandings about proportionality).

[161] COMMENTARY ON THE ADDITIONAL PROTOCOLS OF 8 JUNE 1977 TO THE GENEVA CONVENTIONS OF 12 AUGUST 1949 ¶ 2217 n.15 (Yves Sandoz, Christophe Swinarski & Bruno Zimmermann eds., 1987).

[162] *See* API, *supra* note 89, at art. 35(2)–(3); Legality of the Threat or Use of Nuclear Weapons, Advisory Opinion, 1996 I.C.J. 226, ¶ 78 (July 8); CUSTOMARY INTERNATIONAL HUMANITARIAN LAW, *supra* note 17, at R. 70.

[163] *See* SOLIS, *supra* note 149, at 271–72.

[164] *See* U.S. Dep't of Def. Directive 5000.01, The Defense Acquisition System ¶ E1.1.15 (May 12, 2003) (requiring legal review of all of weapons systems intended for acquisition).

[165] *See* API, *supra* note 89, at art. 36.

also, however, the most significant. If we cannot predict the weapons that will emerge in the future, we must endeavor to describe in more detail what it means to functionally delegate a decision.

The primary inquiry in this context is how confidently we can establish in advance that a weapon system will kill the intended people or classes of people and destroy the intended objects or classes of objects. Targets may be attacked only if they are legitimate military objectives. Whether a target or class of targets are valid military objectives is a decision that must be left to a human. Thus, we must be able to establish through design and testing that an AWS is reasonably expected to attack only those targets or categories of targets that a human has determined to be valid. It is important to note that this principal does not imply that a human must provide input to an AWS that is temporally proximate to lethal kinetic action, a point that will be further explained in subsequent principles.

By way of example, suppose that during an international armed conflict an unmanned submarine is allowed to loiter in international waters and that it is programmed to destroy any and all enemy warships that it identifies.[166] Assume that, through testing, we can ensure to a reasonable certainty that it will attack only these warships and no other kind of ships. After deployment, the submersible spends a great deal of time ignoring ships it cannot yet positively identify. After some time, perhaps months, it locates a convoy of enemy warships and attacks each of them. In this circumstance it would be inaccurate to say that a machine "decided" to attack the enemy ships. Although it enjoyed a significant degree of autonomy, the decision to designate the enemy ships as valid military targets was made by a human during the design and programming of the submersible.

But what if an AWS was granted far broader discretion—if certain pieces of the OODA loop were delegated that obscured our ability to reasonably conclude that a human being made a decision to kill? Suppose that the above hypothetical is modified so that technology has not advanced to the point where we could assure that the submersible would attack only enemy warships. Assume also that the submersible is granted broader discretion over where it may roam the seas. Under these conditions, we could no longer reasonably expect the system to comply with IHL or the law of neutrality.[167] This is not to say, of course, that such a system would be illegal *per se*. But our concerns over whether the system could comply with IHL, and in what circumstances it could be lawfully employed, are

[166] The enemy warships are targetable in international waters at any time due to their status as enemy combatants under the Law of Naval Warfare. *See* U.S. NAVY, U.S. MARINE CORPS & U.S. COAST GUARD, THE COMMANDER'S HANDBOOK ON THE LAW OF NAVAL OPERATIONS NWP 1-14M/MCWP 5-12/COMDTPUB P5800.7A, para. 8.6.1 (2007) ("Enemy warships . . . are subject to attack, destruction, or capture anywhere beyond neutral territory.").

[167] *See* Hague Convention No. V Respecting the Rights and Duties of Neutral Powers and Persons in Case of War on Land, Oct. 18, 1907, 36 Stat. 2310, art. 1 ("The territory of neutral Powers is inviolable."); United Nations Convention on the Law of the Sea, Dec. 10, 1982, 1833 U.N.T.S. 397, art. 2 ("The sovereignty of a coastal State extends, beyond its land territory and internal waters . . . to an adjacent belt of sea, described as the territorial sea.").

greatly increased. This is because we are no longer reasonably certain that the system will attack only valid military objectives. Notably, the submersible is not "deciding" to attack targets; we simply cannot determine which ones it will attack. In this scenario, too many pieces of the OODA loop may have been ceded to the system to be able to state that a human decided to attack a target or class of targets. In this hypothetical, the decision to kill may have been *functionally* delegated to the AWS.

The inquiry therefore hinges on the combination of capabilities and authorities granted to the AWS. Capabilities are a relatively familiar concept: what platform is the system deployed on and what weapons is it armed with? As described in Part II.B.5, these factors are subordinate in the analysis to the authorities granted through computer programming. Authorities relate to questions such as, is the system deterministic and, as such, inherently more predictable? Or has it been granted learning AI, which may make the inquiry far more complex? Can the learning be bounded in a way that alleviates concerns about IHL compliance? These issues are discussed below in Principle 2.

B. *Principle 2: AWS may be lawfully controlled through programming alone*

Developers of AWS can ensure that these systems comply with IHL through their programming, even in cases where the system is granted advanced learning AI. Compliance with IHL was more straightforward in the case of legacy weapon systems, which employed only simple and deterministic AI. Essentially, the prediction regarding what a machine would do in given circumstances was akin to a flowchart of "if/then" determinations. But future AWS that possess learning capacity may still comply with IHL, depending on which puzzle-shaped pieces of the OODA loop they are granted and, more importantly in the learning context, on how the system is bounded.[168]

The challenge of this principle lies in the performance of AI in non-deterministic or partially observable environments, the so-called "fog of war." The dilemma is how to develop algorithms, establish training data,[169] and equip AWS to learn in a battlefield environment, which is one of the most confusing and chaotic experiences thinkable. Computer scientists currently struggle to assure reasonable predictability for AI in rigidly controlled laboratories.[170] How

[168] A simple example is a system whose learning capacity is limited to actions that do not bear on IHL compliance, such as the route an unmanned aircraft will take to the objective based on local weather, assuming air traffic control schemes are accounted for. The more complicated but still tractable problem relates to systems where a system's learning capacity bears directly on lethal actions, such as an unmanned aircraft that uses data from previous strikes to refine its future strike criterion.

[169] *See* Hope Reese, *Why Microsoft's 'Tay' AI Bot Went Wrong*, TECHREPUBLIC (Mar. 24, 2016), http://www.techrepublic.com/article/why-microsofts-tay-ai-bot-went-wrong/ ("One needs to explicitly teach a system about what is not appropriate, like we do with children.").

[170] *See* Interview with Leslie Pack Kaelbling, *supra* note 70.

can we envision AI that could operate with reasonable predictability in armed conflict? The answer, at least in the near term, lies in what tasks we can reasonably expect machines to perform and which combinations of OODA loop fragments cannot safely be delegated to them.[171]

It is safe to assume that in the near term AWS will not be able to perceive, process, and act upon all of the factors humans consider before employing lethal force.[172] However, it does not follow that AWS cannot comply with IHL, even on complex battlefields, while employing learning AI. The inquiry will hinge on the aspects of the uncertain environment for which the AWS must account and, in turn, the range of available responses provided to the system through its authorities and capabilities.

Since AWS will likely not be able to account for all variables and associated uncertainty on a battlefield, we must inquire into which aspects of the environment are particularly relevant to an IHL compliance inquiry. The first step in this process is to catalogue the range of relevant variables we expect the AWS will encounter.[173] Within this range, the second step is to establish which variables we expect the AWS to be able to observe. The third step is to define which variables within the observable range could bear on IHL compliance. The final step is to determine which of the observable IHL-relevant variables that we expect the AWS to encounter will be affected by learning AI granted to the system. After narrowing our category of inquiry in this manner we may arrive at a much smaller and more manageable set of variables.[174]

From this point, the AWS can be programmed to evaluate the probability of certain outcomes as compared to the expected utility of particular actions. Perhaps future systems will need to incorporate highly advanced structured representations of the environment, but not necessarily. The key will be to carefully delineate what the AWS can sense, what it must consider, and therefore the level of sophistication required of the onboard AI.

Suppose that facial recognition software and optical technology advances to the point where airborne drones can affirmatively discern the identity of an individual from afar. The aircraft also have improved efficiency in multiple systems that enable them to loiter for months. Such a system hypothetically conducts a grid-style search for a senior leader from a designated-hostile terrorist group with instructions to attack once it obtains positive identification. After searching for nearly a month, the drone locates a man it believes with 49 percent

[171] *See* discussion *supra* note 127.

[172] Again, the critical failure of *Losing Humanity* was that it assumed AWS would necessarily be called upon to perform all of the functions a human would in a time-compressed lethal decision-making cycle.

[173] This range may be unwieldy, and if those evaluating the system do not carefully define the relevant variables, the range could become intractable.

[174] Care must be taken, however, to map with particularity the ways in which these subsets of variables interact and therefore gain relevance, when they might on their face have none.

certainty is the terrorist leader; it takes no action and continues to track the man. The system continues to refine its facial recognition analysis so that by the next day, the drone determines it is 97 percent certain that the man identified is the proper target. Having studied the leader's habits, the system already knows that the terrorist leader goes for a walk alone at about 0800 every morning. As the terrorist leader strolls alone the following day, and without any input from a human, the drone fires a low-collateral damage projectile that has no explosive charge. The weapon guides itself precisely into the forehead of the terrorist leader, killing him instantly.

Consider first the distinction evaluation in this hypothetical. The drone located what it suspected was the target but was not confidant even to a "more likely than not" standard that it had positive identification. In this case, autonomy decreases the likelihood that an innocent person will be killed. The drone, unconstrained by limited time on station and unaffected by the natural human tendency to jump at the opportunity in spite of uncertainty, simply waited. The bias of humans who must decide whether to act during these windows of opportunity is unarguably in favor of attacking.[175] The drone knows nothing of wanting to "win" or please its boss. It only knows, so to speak, that the criteria for attack have not been satisfied.

Next consider the IHL requirement to take precautions in the attack in order to minimize "incidental loss of civilian life, injury to civilians and damage to civilian objects."[176] In this hypothetical, the drone dispassionately evaluated the situation until it determined the optimal moment when it could achieve the mission and satisfy these requirements. If on the other hand a human pilot was asked to make a shoot/no-shoot decision, he or she would currently have limited time on station to make this call. The human pilot might not have time to establish pattern of life. A pilot would feel pressure to attack for fear that the window of opportunity might close.[177] As such, a human would be more likely to accept greater potential for civilian harm. Although arguably not as pressing as piloted aircraft, a remotely piloted aircraft[178] is also controlled by a human who feels the same kind of pressure.

In this circumstance, IHL compliance was assured through programming alone. The human emotions, which distract from achieving an optimal result under IHL, were eliminated. The AWS was not called upon to do more than could reasonably be sensed, processed, or acted upon by a machine. Through programming, we may therefore leverage the strengths of machine learning while

[175] *See* Alan L. Schuller, *Inimical Inceptions of Imminence: A New Approach to Anticipatory Self-Defense Under the Law of Armed Conflict*, 18 UCLA J. INT'L L. & FOR. AFF. 161, 190–94 (2014) (evaluating *ad bellum* and *in bello* concepts of imminent threats).

[176] API, *supra* note 89, at art. 57.

[177] Based on extensive personal experience and interactions with combat aircrew by the author.

[178] For example, an MQ-9 Reaper. *See MQ-9 Reaper Fact Sheet*, *supra* note 64.

avoiding the pitfalls associated with trusting machines to complete tasks where their performance would be unacceptably unpredictable.

C. *Principle 3: IHL does not require temporally proximate human interaction with an AWS prior to lethal kinetic action*

As a direct corollary to Principle 2, there is no requirement based in IHL that a human must interact with an AWS at or near the time lethal action is taken.[179] As a policy matter, human involvement that is proximate to the point of lethal action might be a good idea under some circumstances,[180] and in others it might not.[181] But there is no legal requirement for it. This flows from the fact described in Principle 2 that AWS are capable of being lawfully controlled through programming alone.

Some argue that having a human involved in the "Act" phase of the lethal decision-making process is always desirable because this will inherently improve the overall performance of weapon systems.[182] This argument fails for three main reasons. First, alluded to previously, is the significant possibility of a future conflict with a peer competitor. Systems that require a human to approve final lethal kinetic actions will likely be incapable of competing at machine speed with sophisticated peer-competitor opponents in the future. Other countries already claim to be developing systems that could potentially take action independent of human approval processes.[183] In order to be postured to meet such threats, the United States and its allies may need to possess systems that can respond at machine speed. Otherwise, enemy systems massed at critical locations on the battlefield could overwhelm the ability of U.S. forces to react.

[179] A detailed discussion regarding whether or not humans *should* interact with a given system at the time of lethal action is beyond the scope of this Article. From a technological standpoint, however, it is likely unwise to try answering this question conclusively when uncertainty exists about what AWS may or may not exist in the future. Each new system must be carefully reviewed to ensure it complies with IHL. The question of whether or not humans will be *required* to interact with the system will depend on the capabilities and authorities granted to that particular AWS.

[180] For example, when involvement by a human would not negatively impact either mission accomplishment or legal compliance to an unacceptable level, such as during deliberate targeting of an objective whose location is fixed.

[181] One example is based on operational necessity, such as when fighting an international armed conflict against a peer competitor in which AWS must engage enemy AWS at such speed and in such numbers that humans would be unable to effectively react. Another example is based on legal compliance, such as when the involvement of a human (who may act based out of fear, anger, or self-preservation) would reasonably be expected to degrade conformity with IHL to an unacceptable level.

[182] *See* Malinowski, *supra* note 7 (arguing that a human should remain "in the loop" for any AWS).

[183] *See, e.g.*, Jason Koebler, *Report: Chinese Drone 'Swarms' Designed to Attack American Aircraft Carriers*, USNEWS.COM (Mar. 14, 2013), http://www.usnews.com/news/articles/2013/03/14/report-chinese-drone-swarms-designed-to-attack-american-aircraft-carriers.

Second, manned systems are a significant drain on personnel, training, and budget. It takes years of instruction and millions of dollars, for example, to train one human pilot in the U.S. military.[184] Once employed, humans tend to wear out if subjected to constant combat operations, even if operating aircraft remotely.[185] We simply may not be able to keep pace with the production and employment of AWS by potential adversaries if we continue to rely on direct and proximate human involvement in the lethal decision-making cycle. Again, this is highly contextual. The same argument would not be nearly as convincing if the particular AWS was designed for any operations short of international armed conflict (IAC) with a peer competitor. Then again, the idea that conflict status and intensity might inform the employment of certain weapon systems is not a novel concept.

Third, human involvement is not always helpful with respect to IHL compliance. Some groups have called for "meaningful" human control of AWS.[186] This Article does not engage directly with that particular debate. However, the conception of AWS presented here suggests that the mere fact that a human pushed a button to either approve or disapprove lethal kinetic action does not necessarily correlate to improved IHL compliance.

The issue of whether or not a human must be involved just prior to lethal action by an AWS is a hotly contested matter. The answer will depend on the specific design and intended use of the particular AWS contemplated. There is no *per se* requirement from a legal standpoint that a human be involved at or near the point of lethal kinetic action.

D. *Principle 4: Reasonable predictability is required only with respect to IHL compliance, but will hinge on the specific fragments of the OODA loop granted to the AWS*

Our ability to predict the actions of future AWS must be based on a reasonableness standard. This is the standard by which we have historically judged weapons systems in determining whether or not they could comply with IHL.[187] But the predictability of the AWS must, from a legal standpoint, be reasonable only as it bears on our ability to understand whether the system will comply with the law. This means that the system may in fact be lawfully unpredictable in certain ways. So long as the ways in which the system is

[184] *See* U.S. GOV'T ACCOUNTABILITY OFF., GAO/NSIAD-99-211, MILITARY PERSONNEL: ACTIONS NEEDED TO BETTER DEFINE PILOT REQUIREMENTS AND PROMOTE RETENTION 18 (1999), http://www.gao.gov/archive/1999/ns99211.pdf ("[T]he cost to train each military pilot through basic flight training is about $1 million, and the cost to fully train a pilot . . . can be more than $9 million.").

[185] *See* James Dao, *Drone Pilots Are Found to Get Stress Disorders Much as Those in Combat Do*, N.Y. TIMES, Feb. 22, 2013, http://www.nytimes.com/2013/02/23/us/drone-pilots-found-to-get-stress-disorders-much-as-those-in-combat-do.html.

[186] *See, e.g.*, *Open Letter*, *supra* note 39 (arguing for "a ban on offensive autonomous weapons beyond meaningful human control").

[187] *See supra* Part II.C.

unpredictable are reasonably unlikely to render an AWS action unlawful, the system may be lawful.

That being said, it would be ill advised to assume that simply because an AWS was predictable in the manner by which it "selected and engaged"[188] targets that it would thereby be lawful. Nor in assessing a system's conformity with IHL should we become overly focused on any particular "critical function."[189] These factors may bear significantly on the analysis of whether a future AWS is *per se* unlawful or not. They will not, however, be dispositive. Those reviewing future AWS for compliance with IHL will need to carefully scrutinize which specific pieces of the OODA loop have been granted to the AWS and, in particular, how machine learning is inserted into this process. Two examples illustrate this point.

First, suppose a legacy drone system such as the MQ-9 Reaper[190] was retrofitted with technology that allowed it to use machine learning to select its route to and from the target area, but was otherwise identical to the system as currently fielded. A remotely stationed pilot would still be required to make the final decision on whether to strike a target. As such, the system could not "select and engage" any target without human approval. Also, as currently posited, the drone would not have any autonomy in its "critical functions." If left unbounded, however, the relatively simple machine learning granted to this system could violate IHL. For example, the system might decide that the quickest way to the battlefield is a straight line—directly through an air traffic control scheme in neutral airspace. Though the remedy to this dilemma is straightforward, the point still stands that potential threats to IHL compliance linger in other combinations of pieces from the OODA loop than simply the moment of kinetic action.

Next, consider the hypothetical unmanned submarine described in the discussion of Principle 1. The system is able to positively identify and attack targets without human intervention. If it is able to positively identify an opposing belligerent's warships, it sinks them without obtaining permission from a human. The submarine is plainly able to "select and engage" targets and has a great deal of autonomy in its "critical functions," but its IHL compliance is not in dispute. This is, of course, a very narrow hypothetical. If the facts were changed to take the example out of the IAC construct of status-based targeting under the laws of naval warfare, then the evaluation of the system would take on a different form. Clear regulations on the employment of such a system may need to be issued along with its fielding to the military. It provides an example, however, of an AWS with sufficient autonomy to conduct the targeting OODA cycle without temporally proximate human input, but which would nevertheless cause little concern from an IHL compliance standpoint.

[188] Meeting Report, Int'l Comm. of the Red Cross, *supra* note 60, at 8.
[189] *Id.*
[190] *See* discussion *supra* note 64.

Thus, the only aspects of future AWS that must be reasonably predictable are those that bear on their ability to comply with IHL. Which particular pieces of the puzzle these are will vary widely based on the specifics of the system being developed. It is impossible to predict the future and, as such, impossible to delineate ahead of time which aspects of all AWS must be scrutinized in order to ensure compliance with IHL. We must not attempt to create law that solves intractable problems within systems that do not yet exist. In much the same way, we must also temper the temptation to focus on the capabilities of developing systems by also inquiring into the limitations of their capabilities.

E. *Principle 5: Limitations imposed on an AWS may compensate for performance shortfalls*

While it is true that future AWS may need advanced AI and machine learning in order to match peer-competitors, it does not necessarily follow that this will lead to the development of sentient "killer robots." To assume so ascribes too much capability from a technological standpoint to systems that are advanced only in narrow, bounded ways. This assumption also fails to recognize that machines will not necessarily be called upon simply to substitute for humans. The focus of the IHL inquiry should therefore delve not only into the capabilities of a given system, but also scrutinize the ways in which the capacity of the system could be bounded through limitations on authorities and capabilities.

By way of example, *Watson*[191] is an AI that is currently quite adept at defeating human opponents at answering trivia questions during the game show *Jeopardy*.[192] This seems highly sophisticated until one reflects on all the things that *Watson* cannot do. To wit, *Watson* cannot go anywhere or kill anybody. Even if *Watson* were outfitted with weapons and mobility, it would have no ability to use them because it was never programmed to do so. Simply put, *Watson* is in this context quite good at only one thing and that is trivia.[193]

At first, this seems somewhat trivial, but the point is by no means glib. Even if we reassure ourselves that the singularity is not near, we should not thereby be satisfied that AWS do not pose a significant threat to IHL compliance. The problem is that some researchers evaluate the technology backwards. Instead of focusing on all the interesting things *Watson* can do, we should instead be asking what it cannot do. For it is only after we satisfy ourselves that a certain system will not be able to complete a certain required task which might affect IHL compliance that we can adequately compensate for the shortfall. Again, the real dilemma with the ability of future AWS to comply with IHL is not in the production of machines that are smarter than us, it lies in the development of

[191] *See Watson*, *supra* note 104.

[192] *See* Markoff, *supra* note 104.

[193] It must be noted that *Watson* technology is being expanded to uses in other industries, but the analogy still applies because the computer will continue to be bounded in innumerable ways.

systems which are quite smart, but not smart enough. There are three primary ways in which technology can be designed to account for this threat.

First, we must not assume that machines will simply substitute for humans in any mission set. Instead, we must recognize that machines will likely team with humans in ways that leverage the strengths and weaknesses of both.[194] Autonomy may simply augment human actions rather than replace them.[195] There may be situations in which AWS can and should be deployed with the authority and capability to take lethal action without temporally proximate human input, such as the future conflict with a peer competitor described above. This will not always be the case, however, and so we must carefully evaluate the unique aspects of how a machine could team with a human rather than whether it would be able to replace a human. The answer will depend heavily on the mission set.

Second, AWS can be limited by restricting their platforms and available weapons. We might not entrust the future drone hypothesized in the discussion of Principle 2 with a missile like those currently carried by remotely piloted aircraft. We would have concerns in that instance over collateral damage.[196] We may instead require it to employ extremely-low-collateral-damage weapons such as a high velocity non-explosive bolt. But it would be erroneous to conclude that because an advanced AI might be unable to process and consider all the complexities of the battlefield that we could not incorporate advanced AI and machine learning into AWS. The destructive potential of any system can be limited through its physical capabilities.

Third, as described by Principle 2, AWS may be bounded through the authorities that they are granted in their programming. Deterministic systems will remain predictable enough that we can ensure IHL compliance through simple "if/then" type programming. But even more advanced agents may comport with IHL if we are able to reasonably predict how the agent function and program will respond in certain environments. Future AWS may need to incorporate highly complicated factored and structured representations of the environment in order to account for the complexities of the battlefield. This is not to say, however, that an AWS must account for all factors that a human might consider in arriving at a decision to employ lethal force. By carefully delineating the variables that the agent must observe and assess, we can establish ahead of time whether the agent program and function will collectively be reasonably likely to meet rational

[194] *See* Interview with Alan C. Schultz, *supra* note 68; *see also* Def. Sci. Bd., *supra* note 32, at 17.

[195] One excellent example of this is autopilot systems in military and passenger aircraft. These systems do not replace pilots, but they do an excellentjob of compensating for areas in which humans are known to perform poorly, such as recognizing the flight profile of the plane and correlating this with required control inputs during emergency situations. *See* Interview with Alan C. Schultz, *supra* note 68.

[196] *See* Paul Scharre, *Autonomous Weapons and Operational Risk*, CTR. FOR A NEW AM. SEC. 18–19 (Feb. 2016), http://s3.amazonaws.com/files.cnas.org/documents/CNAS_Autonomous-weapons-operational-risk.pdf (suggesting that damage potential is an important variable in assessing the risk of unintended engagement by an AWS).

objective IHL standards. If we cannot predict precisely how the machine will learn, then the agent program could be provided with hard rules to bound its behavior.[197] If the machine was still unpredictable in ways that adversely impact IHL compliance to unacceptable levels, it would need to be re-designed.

Conclusion

Autonomy in weapon systems will likely continue to evolve as technology advances. It is incumbent upon those responsible for the legal review, policy, acquisitions, and systems design of future AWS to ensure compliance with IHL. The discussion regarding potential ramifications of increasing autonomy in weapon systems must likewise evolve from theoretical to practical. The principles described in this Article are intended to serve as the foundation for guidance that will help ensure the lawfully responsible development of autonomy in weapon systems.

The U.S. DoD should take the lead in this regard by incorporating practical legal guidance for the responsible development of autonomy in weapon systems into policy. Policies that currently exist should be extended to delineate the specific, substantive areas of focus for those who seek to develop lawful AWS. The principles proposed by this Article seek to reconcile the need for practical guidance with the perils of crafting rules that are either too broad or unduly narrow.

The question of whether creating any particular AWS is a wise policy decision must likewise be carefully scrutinized. There exists a myriad of non-legal concerns that must be addressed. The balance between developing systems that will facilitate the security of free civilization or instead usher in avoidable death and suffering is a delicate one.

[197] For example, if we are not sure what the system will do in all situations, we could instruct it that, “whatever you do, do not do *X*.” While it is relatively simple to build basic constraints, more complicated and subjective restraints may prove far more vexing. *See* Interview with Leslie Pack Kaelbling, *supra* note 70. This is especially so in complex operating environments. *See* Lincoln Laboratories Interview, *supra* note 145. The complexity of the programming dilemma could be reduced in some circumstances by greatly limiting the available options of the AWS. *See* Interview with Ronald C. Arkin, Associate Dean, Coll. of Computing, Georgia Inst. of Tech., Newport, R.I. (Sept. 22, 2015).