Dismantling False Assumptions about Autonomous Weapon Systems

Sarah Scheffler
Boston University Computer Science Department
sscheff@bu.edu

Jacob Ostling*
Boston University School of Law and Brown Rudnick LLP
jostling@brownrudnick.com

ABSTRACT

We respond to two claims within the larger debate over banning versus regulating Autonomous Weapon Systems (AWS). First, contrary to the claim that artificial intelligence (AI)-based AWS are legally distinct from traditionally-programmed AWS, we believe that the same high standard should apply to both types. Second, we caution against reasoning about future AI systems as if they will have the exact same limitations as today's AI. There is reason to believe the current limitations of AI performance and validation methods will be surpassed, especially with regard to the Law of Armed Conflict (LOAC) principle of distinction. We conclude with a discussion of the current state of "predictability" for autonomous systems, with consequences for adherence to the principles of LOAC.

ACM Reference Format:
Sarah Scheffler and Jacob Ostling. 2019. Dismantling False Assumptions about Autonomous Weapon Systems. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Autonomous Weapon Systems (AWS) are weapons which may autonomously choose their targets, and fire upon them without human intervention.[1] As artificial intelligence technology improves, a debate rages between proponents of a preemptive ban on AWS[2] and those who favor regulation of AWS, either via the existing Law of Armed Conflict (LOAC) or a new treaty.[3]

This debate is important, as autonomous weapons have the potential to completely change the way humans wage war. However, in this debate, we have observed two common assumptions that we believe to be faulty. In this paper we challenge these two assumptions and make some observations about the current state of autonomy.

Different standard for artificial intelligence. There is a split as to the exact difference between autonomous systems and automated systems.[4] In one view, artificial intelligence itself should be held to a different legal standard than other programming methods for AWS.[5] This view is based on the assumption of clean separation between traditional "automated" systems, and modern AI-based "autonomous" systems. However, in our view, the line between these categories is very blurry: many mere "automated" systems of the past use machine learning or artificial intelligence. In our view, *all* systems that make targeting and firing decisions without human intervention should be held to the same legal standard, regardless of programming technique.

A key issue for determining the legality of an AWS is the ability to predict its actions. Arguments for different legal standards often focus on the (correct) observation that AI behaves less predictably than traditional programs. However, we believe that the difference is a technical issue, not a legal one. While a traditional program is generally much easier than AI to analyze and test to ensure that it will behave predictably, the legal requirements of the system need not change: traditionally-programmed systems that can choose targets and fire without human intervention should still be considered autonomous.

No progress beyond today's technology. Second, we see many analyses of AWS legality that are based on the assumption that the technology for achieving autonomy will never progress far beyond the level it is now.[6] These analyses generally use the fact that AWS perform poorly at a task today as an argument that all future AWS will also fail at that task, often in favor of a blanket ban on AWS. We argue that there is not sufficient evidence to rule out the possibility that technology will progress to the point where an AWS could sufficiently perform the tasks at hand, at least in limited

[*] Work conducted while a student at Boston University School of Law

[1] This is a United States-centric definition of AWS. Other names for AWS include Lethal Autonomous Weapons (LAW) and Fully Autonomous Weapons Systems (FAW or FAWS). See note 7.

[2] See e.g. [27], [10], [44], [35], [39]

[3] See e.g. [78], [6], [18], [92], [9]

[4] See [5, p., 118], [16, p. 24]

[5] Papers that describe traditional programs and artificial intelligence as potentially legally different include Roff [72, p. 213] ("[AWS] that learn . . . would require a meaningful level of artificial intelligence. I reserve the word 'autonomous' for these weapons."); Williams [95, p. 32-33] (distinguishes an "*automated* system . . . programmed to logically follows pre-defined rules" from an "*autonomous* system . . . capable of understanding higher-level intent and direction" (emphasis added)); Kellenberger [47] ("A truly autonomous system would have artificial intelligence that would have to be capable of implementing IHL")

[6] Papers that make this assumption include Sloan [87, p. 117] (saying that lethal semi-autonomous and fully autonomous robots would not be suited for unconventional warfare); Asaro [10, p. 692] ("Because even 'artificially intelligent' autonomous systems must be pre-programmed, and will have only highly limited capabilities for learning and adaptation at best, it will be difficult or impossible to design systems capable of dealing with the fog and friction of war"); Sharkey [85] (describing how robots "do not have adequate sensory or vision processing systems for separating combatants from civilians," nor do they have the "situational awareness to make proportionality decisions," nor accountability, but without accounting for the possibility for these assessments to change in the future); Docherty [27, §IV] (saying that fully autonomous weapons "would be incapable of abiding by the key principles of international humanitarian law [and] unable to follow the rules of distinction, proportionality, and military necessity")

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM. . . $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

circumstances. Furthermore, it is prudent to plan for the possibility that the technology will progress to that point. We do not dispute that today's AI is limited in the ways they describe, but that does not preclude the possibility of future improvement.

We begin by providing background on autonomous weapon systems (section 2), and the legal requirements they must meet (section 3). Then, we provide our counterarguments for each of the above claims (sections 4 and 5). At the end, we describe the current state of affairs in testing and validation methods for autonomous systems, and identify areas in artificial intelligence testing and validation that are undergoing current research (section 6).

2 BACKGROUND ON AUTONOMOUS WEAPON SYSTEMS

We use the United States (U.S.) Department of Defense (DoD) term *Autonomous Weapon System* (AWS) to refer to weapons that can make targeting decisions and fire without intervention from a human.[7] One important feature of this definition is that we only consider "autonomy" with regard to targeting and firing decisions. Many weapon systems incorporate autonomy in a non-targeting application (for example, mobility or intelligence gathering)[8], but our definition of AWS only applies to systems that make targeting and firing decisions autonomously.

2.1 AWS in use or development

Limited autonomous weapon systems have been in military use for decades.[9] Automated target recognition (ATR) software, which automatically identifies targets that match a target profile, have been in use since at least the 1970s.[10] These were based on image processing and machine learning at least as early as 1986,[11] and used neural networks as early as 1990.[12] Generally, these systems were only developed to point out targets to human operators, who would pull the trigger.[13] The main exceptions are in domains where human reaction time is not sufficient to stop an enemy attack, such as missile defense.[14]

This is starting to change. Over the course of the 2000s, more decision-making was moved onboard remote systems, especially in the context of Unmanned Aircraft Systems (UASs) and Unmanned Maritime Systems (UMSs).[15] The U.S. has identified autonomy as a key component of its strategy, and is the dominant player in AWS.[16] Many other states are beginning to field or develop increasingly autonomous weapons, including Israel, China, Russia, the Republic of Korea, and Taiwan,[17] and non-state actors are also beginning to pursue unmanned and autonomous systems.[18] Many more states also have access to unmanned and autonomous weapons by buying them from countries that are developing them.[19]

Over the coming years, greater autonomy is expected to be added to existing systems, but only slowly. Aside from static and ship defense, and domains that require too quick a reaction time to involve a human, systems today generally use automation to point out targets, but not to fire upon them autonomously.[20] It is expected that the next stage will have a machine autonomously choose a target and fire unless a human vetoes the attack, as long as the machine is reasonably confident it is acting correctly and legally. As we will discuss in Section 6, these two phases can be used for more advanced testing and data-gathering. Eventually, if the technology progresses enough, AWS may become commonplace filling many different battlefield functionalities. For the meantime, however, the DoD intends to maintain "appropriate levels of human judgement over the use of force."[21]

2.2 Programming methods

We proceed to define three overlapping programming techniques: traditional programming, machine learning, and artificial intelligence. These can be placed along a spectrum of increasing "learning" done by the program.

Traditional programming. This term refers to programs that use conventional logic and programming languages. The programmer specifies how to process the input via code that can be read as a logical flowchart. This category has also been referred to as "rule-based systems," [22] "classical programs,"[23] the "bottom-up" approach,[24] "if/then systems,"[25] and "handcraft programming."[26] Although traditional programs are often easier to understand than those written using machine learning or artificial intelligence, this is not always the case. Traditional programs are often incredibly difficult to understand, either because they are encoding a complicated algorithm or because they can be quite large. Moreover, such programs are not necessarily deterministic. Traditional programs sometimes use randomness to make decisions; this does not require machine learning or artificial intelligence.

Machine learning. Machine learning (ML) occupies the middle of the continuum between traditional programming and artificial intelligence. Machine learning programs attempt to discover or replicate patterns in datasets[27] using a combination of applied statistics and computing power.[28] Generally, the workflow of (supervised) ML is to *train* a model on an existing dataset where the results are known, *test* it using additional known data, and later *deploy* it on new unknown data. The ultimate goal of this process is to predict

[7] [19, p. 13]. For a summary of other proposed definitions of AWSs, see [49, p. 21-32] and [16, p. 8].

[8] [16, p. 21, 27]

[9] See generally [15, p. 401-402], [20, p. 82], [54, p. 5472], [51, p. 276-278], discussing the HAROP system developed by Israel Aerospace Industries, several missile defense systems (Israel's Iron Dome, the Netherlands' Goalkeeper, and Russia's Kashtan), anti-personnel sentries (the Republic of Korea's SGR-A1, Israel's Guardium, and the U.S.'s MDARS-E), as well as air, sea, and ground vehicles developed primarily by Israel, Russia, and the U.S.

[10] [16, p. 24]

[11] [14, p. 367], describing an ATR system that uses image processing and a k-nearest neighbors classifier

[12] [73]

[13] [50]

[14] [3] [34] [84], discussing CIWS and the Israeli Iron Dome system

[15] [26]

[16] [16, p. 58]

[17] [46] [59, p. 1] [96, p. 18-20]

[18] [96, p. 21]

[19] [96, p. 20]

[20] [7, p. 6-7] [16, p. 26]

[21] [19, §4.a]

[22] [37, p. 10]

[23] [53, p. 155]

[24] [42, p. 30]

[25] [81, p. 406-407]

[26] [16, p. 16]

[27] [83, p. 24-25]

[28] [83, ch. 2]

something about the new data. The model will base its predictions about the new data on patterns learned from the training data. Before deployment, the model is tested on the test data to measure its accuracy.

Artificial intelligence. Artificial intelligence (AI) is a broad term that includes machine learning. Beyond machine learning, AI also involves the creation of intelligent "agents" that learn about their surroundings and act upon them.[29] Some calls for banning or regulating AWS specifically focus on the issue of human-like or general AI.[30] We do not restrict our analysis of AI in this way, we also accept the possibility of AWS that use AI in a narrow domain, without any human-like intelligence, for example, an AI-based targeting system.[31]

An AWS could be built using traditional programming, ML, AI, or any combination thereof. The categories are not mutually exclusive, and most systems containing ML or AI also contain a fair amount of traditional logic. A program written in a completely conventional, non-machine learning, non-artificial intelligence manner could still be considered "autonomous" in the sense that it could choose and fire on a target without human action. Conversely, even a super-intelligent AI might not be able to engage targets it selects without human permission.

3 LAWS GOVERNING AUTONOMOUS WEAPON SYSTEMS

Before we address trends in the debate regarding AWS, we describe the existing legal requirements an AWS must meet. At the time of writing, no existing international law or treaty specifically references AWS[32] and the use of AWS is not sufficiently widespread for technology-specific customs to have developed through state practice.[33] The U.S. DoD has avoided claims about the legality of autonomous weapons that it uses in lethal settings,[34] presumably to avoid setting an unintended precedent. In the absence of treaty law, the baseline International Humanitarian Law (IHL) restrictions on weapon systems apply.[35]

Five broad principles limit the use of weapons under international law: (1) unnecessary or superfluous suffering; (2) military necessity; (3) proportionality; (4) distinction; and (5) command responsibility or "accountability".[36] Each of the five are recognized to some extent as customary international law ("CIL"), so they are binding on states to some extent regardless of treaty status.[37] Thus, if a weapon can never be used in a manner that comports with each standard, then it is *per se* unlawful.[38] There is nothing inherent to AWS that prevents them from abiding by each of these principles, but each principle does impose limits on their use *as applied.*[39]

Even if AWS as a class of weapon system are not *per se* illegal, the principles of IHL, especially distinction, proportionality, and accountability do impose substantive restrictions on the development and deployment of AWS.[40] We describe the restrictions (or lack thereof) on AWS for each of the five principles.

3.1 Suffering

The prohibition on unnecessary or superfluous suffering outlaws weapons which cause suffering to combatants with no military purpose.[41] The principle is codified in AP 1, Art.35(2), and as applied, is concerned with a weapon's effect on combatants (i.e. poisoning), not the platform used to deliver that weapon.[42] While an AWS has the potential mis-judge the amount of suffering it will cause, nothing in the use of an AWS as a delivery platform modifies the harm inflicted by a particular type of weapon.[43] Thus, AWS could comply with the rule by employing any traditionally legal weapon.

3.2 Military necessity

Military necessity requires that a weapon provide an advantage for legitimate military objectives.[44] This principle is augmented by the rule of precaution in attack, codified in Article 57 of Additional Protocol One, [45] but also reflecting CIL, which requires that attackers exercise "constant care … to spare the civilian population."[46] In particular, with respect to the means of warfare used, Article 57 requires that attackers use the means least likely to harm civilians, unless doing so would sacrifice some military advantage.[47] Consequently, to satisfy both rules, AWS must avoid civilian casualties

[29] [75, p. viii]

[30] See e.g. Docherty [27, p. 27]

[31] [73]

[32] [78] (describing principles of IHL and LOAC as the primary treaties affecting autonomous weapons in the absence of more specific treaties). See also [27, p. 1] (discussing the absence of any treaty specifically prohibiting autonomous weapon proliferation); [32, p. 4] (implicitly acknowledging absence of more specific treaty law by resorting to the Convention on Certain Weapons, and general principles of international humanitarian law to argue for legal limitations on autonomous weapons proliferation)

[33] The following compilations of customary international law do not reference any specific customary international law restricting autonomous weapons: [17, 41, 43, 62, 79, 94]. These were all reviewed by Schmitt [78, note 6], and did not make note of any customary international law imposing specific limitations on autonomous weapons, or indeed, any reference to autonomous weapons at all.

[34] [19, §4.a]

[35] [78, p. 32]

[36] [78, p. 8-9 and 33] (discussing specific restrictions of LOAC in the context of the underlying principles of distinction, proportionality, military necessity, unnecessary suffering, and accountability). See also, [56, §3.1.1, 3.1.2] (describing the LOAC principles that should be considered in the absence of expressly limiting treaty law and customary international law).

[37] [80, p. 231, 244, 251, 253, 259, 263] (discussing CIL status and relevant Additional Protocol One [1] provisions of the prohibition on unnecessary suffering, military necessity, proportionality, distinction, doubt, and command responsibility in respective sections throughout the paper)

[38] [78, p. 2]

[39] [78, p. 35] ("This article has demonstrated that autonomous weapons are not unlawful *per se*"); [80, p. 279] ("First, autonomous weapon systems are not unlawful per se. Their autonomy has no direct bearing on the probability they would cause unnecessary suffering or superfluous injury, does not preclude them from being directed at combatants and military objectives, and need not result in their having effects that an attacker cannot control. Individual systems could be developed that would violate these norms, but autonomous weapon systems are not prohibited on this basis as a category."); see also [7, p. 1105]

[40] [80, p. 279-281]

[41] [1, art. 35(2)]

[42] [78, p. 9]

[43] [80, p. 279] ("Autonomous systems would not automatically violate the prohibition on unnecessary suffering and superfluous injury because Article 35(2) only addresses a weapon system's effect on the targeted individual, not the manner of engagement (autonomous).")

[44] [80, p. 232] (explaining that military necessity is a principle that undergirds IHL rather than a strict requirement, and may be satisfied by "interest in maintaining a technological edge over potential adversaries, in particular by fielding systems that enable them to deliver lethal force while minimizing the risk to their own forces.")

[45] [1]

[46] [80, p. 259]

[47] [80, p. 259-261] (explaining that the key issue in the requirement to exercise precaution is feasibility; commanders are not required to sacrifice military advantage to mimimize harm to civilians)

at least as well as existing weapon systems *or* provide some other military advantage unavailable from those systems.

This is a low threshold in practice, because autonomy does offer numerous advantages. Autonomous systems reduce the need for communication between a system and a human pilot or operator.[48] This is especially useful compared to remote-controlled systems in environments where communication is denied or difficult,[49] but even in environments with assured communications, autonomy frees up communication bandwidth for other uses and allows reaction speeds quicker than communication latency would permit.[50] Unlike humans, who suffer cognitive fatigue after time has passed and suffer from stress, autonomous machines generally continue functioning at full potential for as long as they remain turned on.[51] They may directly use sensors more advanced than human senses, and they can integrate many different data sources effectively.[52] The quick reaction time of autonomous systems may be useful in applications where human reaction time is insufficient to address incoming threats.[53] Autonomous systems could further allow for reduction of expensive personnel such as pilots and data analysts.[54] Finally, unmanned autonomous weapons need not prioritize their self-preservation, enabling them to perform tasks that might be suicidal for manned systems.[55]

3.3 Proportionality

Proportionality, as codified in Article 51(5)(b) of Additional Protocol One, prohibits attacks "which may be expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated."[56] IHL does not offer any numerical definition for what constitutes an "excessive" ratio of civilian casualties, nor has any international consensus developed.[57] In practice, the standard has been treated as requiring that civilian casualties be "reasonable" in relation to military advantage.[58] This reasonableness standard may be satisfied by selecting targets based on assessments which use numerical values to weigh risks to civilians, such as the U.S. Collateral Damage Estimation Methodology.[59]

Today's AWS can only meet half of this standard. There is no technical barrier preventing an AWS from programatically determining acceptable distributions of collateral damage using existing frameworks, however, "military advantage" is considered a subjective case-by-case evaluation.[60] There are some initial attempts at technical methods for assessing proportionality from first principles, but they leave much to be desired both in terms of practicality and validity.[61]

In the near term, human commanders can perform this assessment and pre-specify conditions under which the AWS can act without violating proportionality.[62] Moreover, environments without civilians (e.g., undersea) offer venues where the proportionality assessment is likely to be fairly straightforward. Thus, AWS can satisfy proportionality.

3.4 Distinction

Distinction, codified in Article 48 of Additional Protocol One, requires that parties to a conflict "at all times distinguish between the civilian population and combatants and between civilian objects and military objectives."[63] Distinction only renders a weapon *per se* unlawful if it is *incapable* of being directed at a specific military objective, although such a weapon may be unlawful as applied for failure to distinguish between combatants and civilians during use.[64] Attackers are further required to err on the side of caution where there is doubt as to whether a target is a civilian or a combatant.[65] Today's AWS are are generally considered unable to distinguish between combatants and non-combatants in urban warfare conditions or other environments that mix combatants and non-combatants,[66] but even absent the ability to distinguish at this high level, AWS could legally be used in scenarios where there are no civilians present (e.g. undersea submarines or missile defense).[67]

3.5 Command accountability

Lastly, command responsibility, also known as "accountability," sets forth the requirement that superiors be held liable for war crimes committed by their subordinates, if they knew or should have known of the crimes, and failed to take reasonable measures to prevent those crimes.[68] Legal systems which impose responsibilities on superior officers to uphold the law are generally sufficient to satisfy the command responsibility requirement.[69] Human Rights Watch (HRW) has argued that AWS violate the principle of command responsibility "[s]ince there is no fair and effective way to assign legal responsibility for unlawful acts committed by fully autonomous weapons."[70] This is incorrect; humans decide how to deploy AWS, and may be properly held liable for failing to do so in accordance with the law.[71] As long as human commanders choose when and how to deploy AWS, responsibility for the weapon's actions rests with the commander, not the weapon itself.

IHL dictates a requirement that commanders must actively ensure that weapons they employ adhere to LOAC.[72] The DoD has also

[48] [68]
[49] [76, p. 39]
[50] [76, p. 43-44]
[51] [24, p. 15]
[52] [51, p. 280]
[53] [33, p. 16], [89, p. 96]
[54] [36, p. xii]
[55] [81, p. 406], [51, p. 280]
[56] [1]
[57] [80, p. 254]; [88, p. 316-318]
[58] [80, p. 256]
[59] [63]
[60] [78, p. 19-20] ("Because it is contextual, the military advantage element of the proportionality rule generally necessitates case-by-case determinations.")
[61] See e.g. [8, §6, §7] (describing a numerical model for making lethal decisions ethically and legally)
[62] [16, p. 74-75], citing [90, p. 189] and [78, p. 20-21]
[63] [1]
[64] [78, p. 10]
[65] [1, Art. 51(4)]
[66] [38, p. 129] ("until much more progress is made, we should not be sanguine about the advantages of robots in theaters on the noncombatant end of the spectrum")
[67] [7, p. 6] [78, p. 13]
[68] [1, Art 86 & 87] [78, p. 33-34].
[69] [78, p. 33-34]
[70] [27, p. 42].
[71] [78, p. 33] [29, p. 69]
[72] [81, note 121]

imposed this responsibility unambiguously, stating that "[p]ersons who authorize the use of . . . autonomous weapons must do so with appropriate care and in accordance with the law of war, applicable treaties, weapon system safety rules, and applicable rules of engagement (ROE)."[73] The threat of repercussions to those who would recklessly deploy AWS is a substantive restriction on their use.

4 RESPONSE TO THE "DIFFERENT STANDARD FOR AI" ARGUMENT

As mentioned in Section 2, automated targeting recognition (ATR) systems have been in military use since the 1970s, and ship defense systems that choose targets via ATR and fire upon them have been in use since the 1980s. To see one reason why the "higher standard" argument fails, we show that determining whether ATR should be considered "autonomous" goes beyond the question of whether they employ AI or traditional programs.

Some parties consider ATR systems with permission to fire merely *automated* rather than *autonomous*, since the criteria for targeting were pre-programmed by a human.[74] However, this line becomes very blurry very quickly. In current targeting systems going back at least as far as 1986, a sensor image is processed and evaluated against a ML model trained on images of enemy aircraft.[75] This is already more technologically "learned behavior" than many assume. But is it significantly different than a system that only used traditional programming, by comparing a processed image with pre-programmed specifications regarding the target's size and appearance? It is difficult to see why these should be considered different.

A similar question arises in the other direction: If these targeting systems were not autonomous before, do they become autonomous if their ML models are updated with new images of aircraft spotted during their use?

Moreover, if a very dedicated programmer examined the final code of an ML model or AI agent and rewrote it from scratch, surely it did not become traditionally programmed and therefore no longer autonomous?

A much more consistent and practical approach is to consider the traditionally-programmed weapon system to be autonomous as well, since they do select their own targets (albeit using pre-specified criteria) and fire upon them. A high standard should apply to all AWS, regardless of programming technique.

Including traditional programs as AWS also makes sense from a security perspective. Traditional programs are also vulnerable to attacks where they behave very differently than intended, or even fall under adversarial control. Modern security, authentication, and encryption mechanisms have been improved and refined over the years and it is commonly understood that the earlier these requirements are incorporated into the design of a system, the better that system will be (both in security and performance).[76] Traditionally-programmed weapons systems (autonomous or not) must defend against adversarial action as well.[77]

The "different standard for AI" argument also has consequences for the LOAC requirements of distinction, proportionality, and command accountability. (The requirements of suffering and military necessity are not strongly affected either way.) All three of these requirements are also affected by the problem described above, of the categorization being unclear. However, each of these also suffers from additional issues:

Distinction. The "autonomous" part of an AWS to refers only to the AWS' permission to choose its targets and decide whether to fire. If AI were held to a higher legal standard than traditional programming then a question arises as to whether the "autonomous" part is in the identification of the target, or the decision to fire.

Imagine an AWS that "learns" its targets using AI, and uses traditional programming to compare identified targets to pre-specified profiles and make a firing decision. AI would be responsible for choosing targets, but not the decision of when to fire. If the AWS fires on a non-combatant, was it due to the AI, or the traditional program? In our view, the failure was in the whole system (since it could have been addressed by improvements in either one), and thus, it makes sense to treat this system as any other AWS, even though the firing decision logic was pre-programmed.

There is one other issue with treating traditional programs as non-autonomous regarding distinction. It is generally accepted that the pace of battle and warfare is speeding up.[78] So far, autonomous targeting decisions have only been required in especially fast-paced domains, such as missile defense. But as this changes, more machine decisions will replace human decisions. Leaving traditionally-programmed AWS out of the discussion would be a mistake.

Proportionality. Since attempts at having AWS "learn" how to balance military advantage against civilian casualties are still in their early stages[79] (and may indeed never be achievable), several parties have proposed that human values could be pre-programmed into an AWS instead.[80] This puts proportionality decisions in the realm of traditional programming rather than AI. But if a weapon system is in a position to need to make a proportionality judgement in the first place (using pre-preogrammed values or not), then that should be autonomous enough for the purposes of being considered an AWS, even if the system contains no AI at all.

Command accountability. A common argument for a ban on AWS is that AI is too unpredictable to be governed under the command accountability standard.[81] Commanders would be unable to judge what AI is likely to do, and therefore could not be held accountable for the AI's actions. Technically, this is not exactly a "higher standard" argument – the objection is that while traditional

[73] [19, §4.b]

[74] [16, p. 24] ("There is an open debate over whether it is appropriate to discuss autonomy in the area of targeting because the software technology that existing weapon systems use to find and attack targets is, from a technical standpoint, closer to basic automation than autonomy."

[75] [14]

[76] [69]

[77] See e.g. [40]

[78] [86, p. 63] (quoting an unnamed army colonel: "The trend towards the future will be robots reacting to robot attack, especially when operating at technologic speed . . . As the loop gets shorter and shorter, there won't be any time in it for humans.")

[79] [8], [64]

[80] [16, p. 74-75], citing [90, p. 189] and [78, p. 90]

[81] See e.g. [28] ("Would fully autonomous weapons be predictable enough to provide commanders with the requisite notice of potential risk? Would liability depend on a particular commander's individual understanding of the complexities of programming and autonomy?")

programming can meet this single standard, artificial intelligence cannot. However, the core of the argument stands, since as a practical consequence of this requirement, the single standard would restrict AI more than traditional programs.

The core of the argument is based in fact – AI is notorious for failing unpredictably, in contrast to traditional programs.[82] As we will discuss further in Section 6, testing methods for AI perform poorly compared to those for traditional programs, partly because of the relative age of the fields and partly due to the nature of the problem itself. However, we need not develop perfectly predictable AI, we need only develop AI that is capable of meeting the reasonableness standard.[83] The difference between restrictions on AI and traditional programs would lessen in the presence of better testing mechanisms for AI, rather than remaining artificially inflated as a result of a separate standard.

5 RESPONSE TO THE "AI WILL NEVER IMPROVE" ARGUMENT

As we have described above, although there are constrained environments where AWS can meet the requirements of LOAC, there are some tasks that AWS struggle to do with current technology.

There is reason to believe that artificial intelligence will see major changes in the coming years. Beyond the DoD's push toward autonomy, AI and ML are also growing in the commercial sector. Developments in autonomous vehicles are likely to be especially helpful for military applications. 37 states have enacted legislation or passed an executive order to encourage the testing of autonomous vehicles.[84] However, the progress in the commercial sector may not be enough – requirements on AWS are likely to be tougher than anything average companies have to deal with.[85] We proceed to specifically address how developments in AI could impact distinction, proportionality, and command accountability.

Distinction. Those who doubt AI will ever be able to distinguish combatants from non-combatants have a point – targeting itself is already a difficult task on the battlefield, without even getting to determining the target's validity. Current targeting practices are severely limited and suffer from too many false alarms and the inability to recognize "clutter" in a sensor image.[86] They must also deal with dynamic targets which not only look different in different environments, but also actively mask themselves or create decoys.[87] Presently, human+machine combinations seem to outperform both humans alone (in time and accuracy)[88] and machines alone (in accuracy).[89] Improving the performance of targeting software (using any and all programming techniques) is an area of ongoing research.[90]

There have been attempts to create AWS that can directly satisfy distinction in the context of recognizing surrender. The Samsung SGR-A1, an AWS developed for use (though never used autonomously) in the demilitarized zone between South and North Korea, recognized arms held high above the head as a sign of surrender.[91] However, so far, these attempts have not been promising enough to survive pilot programs.[92]

However, none of this is to say that AWS will *never* be able to distinguish adequately between civilians and combatants. Given the rapid progress in other technological fields, it seems prudent to at least allow for the possibility that distinguishing AWS could be built. Indeed, algorithms could ultimately become better than humans at distinguishing.[93]

Proportionality. The impact on proportionality is similar to the impact on distinction, though the current trends in the technology show less potential for improvement.[94] It has already been demonstrated that AWS could use human-programmed values to aid in the proportionality determination, but what about fully artificial reasoning?

Existing methods to technologically learn the principle of proportionality essentially attempt to encode numerical and logical systems of ethics.[95] A similar idea is to encode guidelines found in army manuals.[96] A third option is to have the machine learn from human decisions.[97] All of these attempts focus on a narrow domain where the AWS must make decisions. They impose restrictions on the AWS's actions, requiring it to adhere to ethical or legal principles.

For proportionality more than distinction, newer developments are needed before an AWS should try to directly automate the proportionality decisions. As AWS become more popular, this will become a more important endeavor.

Command accountability. Finally, the command accountability requirement dictates that AI be sufficiently predictable for commanders to be reasonably held accountable for their actions.[98] As we said in Section 4, however, we need not predict AI perfectly. We simply need to predict it sufficiently well for command accountability to hold. AI testing is an active area of research, and we believe it is possible for AI to be predictable *enough* for the command accountability property to be met in the future. In the meantime, it would be prudent to hold commanders accountable for even unpredicted AWS actions, to discourage over-reliance on the AI.

[82] [60]; see also [11] on synthesizing inputs which cause these failures and [48] on surprising results during AI development

[83] [29, p. 69] ("a commander must have a reasonable understanding of the AWS and how it will work before deploying it in a particular situation"). See also "dynamic diligence" [52]

[84] [2]

[85] [25, p. 12-13] (Most commercial applications of autonomy benefit from a simplified environment, greater ability to use humans for difficult steps in the process to be made autonomous, and a lack of direct adversaries that are attempting to defeat the commercial system.); See also Feldman et al. [31, p. 6], describing how resiliency is not a priority in commercial AI, but that development of military AI may aid the commercial sector.

[86] [70, p. 7]

[87] [70, p. 7]

[88] [70, p. 7] [45] [71] [50]

[89] [70, p. 7,9] [14]

[90] [23]

[91] [66], [16, p. 25]

[92] [64]

[93] [16, p. 74], citing [51, p. 273-315] and [9, p. 1-9]; [92]

[94] Contrary to the technological attempts to improve distinction, there is only one major attempt at having a machine perform the entirety of the proportionality computation [8, §5-6] and to our knowledge, it has not been incorporated into any military programs for testing.

[95] [8, §5-6]

[96] [63]

[97] [64]

[98] See note 83

6 OVERCOMING THE PREDICTABILITY BARRIER

A key issue involved in determining whether or not an AWS meets the LOAC criteria involves how predictably it will behave.[99] The idea that AI is too unpredictable for use in AWS is behind both arguments we disputed earlier. The worry about holding AI to a higher standard is only raised because we are not good at predicting what AI will do. The belief that AI will always fail at certain tasks incorporates the assumption that we will never be able to better predict AI behavior. But, even for humans, perfect adherence to IHL is an impossible goal.[100] We do not need to be perfect at predicting AI behavior, we just need to be sure that our predictions are good enough to comply with LOAC.

Both assumptions are also based on the correct premise that today's technology is not good at predicting AI actions, compared to traditional program testing. Due to the nature of AI itself, we will likely never be as good at predicting AI as we are at predicting the behavior of traditional programs (though traditional programs have bugs and exploitable flaws that demonstrate that we are not perfect at predicting the behavior of those either). But there is a real possibility that we will get good enough at predicting AI that LOAC compliance in more complex environments is possible.

Absent a blanket ban, we must ensure that International Humanitarian Law apply to all AWS in a coherent and consistent manner. In this section, we describe the current state of testing and evaluation for traditional programs, machine learning, and artificial intelligence.

The DoD (and the commercial domain) has mature testing methods for traditional programming, but these are not sufficient for the requirements on autonomous weapon systems.[101] Autonomous weapons also have high security requirements to mitigate damage if an autonomous device is compromised. Even with the limited forms of autonomy in use today, misjudgements in software and malfunctions in hardware have been lethal.[102] The human challenges brought about by AWSs are no less daunting – humans must understand how and when to use autonomous systems. The human-computer interface problem has caused instances of friendly fire in the past because even the humans trained in the use of a system misunderstood some of the system's properties,[103] or came to be overly-reliant on its judgements.[104] Human commanders must learn and understand when an autonomous system is a good alternative to using a non-autonomous system, and ensure that the personnel that monitor them have been sufficiently trained in how to do so.[105]

6.1 Testing methods for traditional programming are mature

One advantage of traditional programming over ML and AI in the context of AWS is that the testing techniques for traditional programming are much more mature than those for ML and AI.

Over decades of developing programming languages and techniques, robust methods and frameworks have been developed for unit testing individual components of a program and integration testing to ensure that the components work together properly. These include static code analysis (analyzing the program's source or binary without running it), dynamic analysis (running the program on some test inputs), fuzz testing (running the program on random inputs), and a number of design principles to help ensure that the tests are covering as many edge cases as possible.[106] Tests are sometimes accompanied by formal verification, which seeks to generate a proof of correct behavior in the presence of good inputs and some assumptions.[107]

One important takeaway from traditional testing methods is the idea of deliberately inducing failure. It is not sufficient to ensure that the program works correctly on expected inputs; programs must also fail gracefully on unexpected inputs – say, by outputting an error instead of crashing or exhibiting undefined behavior.

6.2 Traditional programs often have catastrophic bugs and vulnerabilities

Despite all the testing methods for traditional programs, software is still notoriously buggy and rife with exploitable flaws. In 2015, a pair of hackers demonstrated their ability to remotely hack and control a car.[108] This would clearly be a bad outcome for an AWS, and this vulnerability arose solely from traditional programming methods (though hopefully, a military would have had more stringent security requirements than the car manufacturer). This is a different problem than the unpredictability of AI, but it demonstrates that we also sometimes struggle to predict the consequences of adversarial inputs in traditional programming. In addition, AWS based on any method would also need to be resistant to physical tampering and would need sufficient security and encryption to prevent secret information from falling into enemy hands.

6.3 Testing methods for machine learning and artificial intelligence are limited

By contrast, testing methods for machine learning and artificial intelligence are much less developed.[109] This is partly due to the younger age of the field; we should expect testing methods for ML and AI to improve over the coming decades.

Supervised[110] machine learning models are typically evaluated based on their accuracy on the test data. At minimum, the model is run against test data (with known "truth") to measure its accuracy. Sometimes, multiple parameters in the model are tweaked to find the best tradeoff between false alarms and false misses. Calibration[111] is also a common tool to ensure that the algorithm knows how likely a specific output is to be correct.

[99] [19, enclosure 2]
[100] [81, p. 408]
[101] [24, p. 62] [59, p. iv] [21, p. 5] [30, p. 23]
[102] [82]
[103] [67]
[104] [86, p. 125]
[105] [24, p. 65]
[106] For an introduction to topics in software testing, see [58], [22]
[107] See e.g. [57]
[108] [55]
[109] [26, app. D]
[110] Supervised ML uses example input/output pairs to let the model learn what output to return on new inputs [37, p. 103]. In unsupervised ML, there are no known examples, rather, the algorithm looks for patterns in data.
[111] [13]

Machine learning also struggles if the algorithm does not match the dataset.[112] But this is not simple: ML and AI programs are often used to solve problems that the programmers cannot verify directly. It is virtually impossible to generate the kind of data needed to train military-combat AI in a laboratory setting[113] and the space of possible inputs/outputs to test becomes so large that it is impossible to test them exhaustively.[114] The program will be either "too abstract to be properly computable, or too specific to cover all situations."[115] But the program must be given some form of "correctness" (in the form of labeled training data, a utility function, or observing decisions) to be able to learn what to do.

There are very few mature methods for testing machine learning or AI beyond the accuracy approaches, but there are many candidates under current research. See e.g. [65, 91, 93] just to name a few. The accuracy approach is useful for making the algorithm behave well on inputs it (or the programmer) expects, but are ineffective at helping the algorithm deal with unexpected input.

Methods to test AI robustness have been identified as a key research area for AI in 2015.[116] We need testing methods that measure AI's *verification* and *validity*. These two measures are different. Verification asks how good the system is at achieving its goal, while validity asks whether its goal was good in the first place. ML and AI are also vulnerable to attacks where adversarially-chosen input can consistently yield incorrect results.[117]

For the meantime, we do not have mature tools to estimate the likelihood of failure under adverse conditions, though this is an active area of research in both the military and commercial sectors. However, we only need a rudimentary estimate of the probability of failure in order to take steps to reduce the consequences of failure, should one occur.

6.4 Restricting the consequences of failure

Aside from the question of how likely an AWS is to fail, we must also address the consequences of failure. The capability of the AWS and the environment in which it is deployed both affect the consequences of a malfunction.[118] Furthermore, the speed with which the AWS acts could affect the damage done by a misbehaving AWS: If a human operator is present or the AWS automatically stands down after a certain amount of time has passed, then the damage potential is limited by that timespan and the speed of the AWS.[119] A human supervisor is unlikely to prevent failure from occurring in the first place, but there are many scenarios where the human can prevent a failure that already happened from getting worse.[120] This has design consequences: if we are to rely on human intervention to limit the consequences of a bad decision made by an AWS, then the decisions made within the timeframe of a few seconds should have minimal irreversable consequences.

Temporal or hardware restrictions can also reduce the risk of misuse. An analogy may be drawn to mines.[121] Under the Hague Convention, unanchored automatic contact mines must become harmless within an hour after they leave control of whoever laid them.[122] AWS can apply the same principle and deactivate (or some other appropriate action) if they are away from human control for longer than expected. Putting these restrictions in hardware and rendering the AWS physically useless after this period will also lower the risk of adversaries controlling the AWS.

7 CONCLUSION

We began this work by highlighting two frequent claims in the debate over whether to ban AWS, both of which arise due to the unpredictability of AI systems using today's technology. We rebut the first claim, that AI-based autonomous weapons should be held to a different legal standard than traditionally programmed AWS, by showing how blurry the line is between different technologies in AWS, and by specifically addressing consequences of analyzing compliance under IHL. We fully believe that AI-based AWS should be held to a high standard to ensure that it complies with extant law, but we believe that all AWS should be held to the same standard, regardless of their programming technique. The second claim is that AI will never improve significantly beyond today's technology. Though none of us know for sure what the future will bring, there is enough evidence of rapid improvement in AI that we think it prudent to allow for the possibility that AI will improve greatly over the coming years. Both of these claims were partly based on the unpredictability of today's artificial intelligence. We provide a feel for the current landscape of testing for traditional programs, ML, and AI, and identify methods for mitigating bad consequences of AWS in the short term, while our understanding of AI improves.

REFERENCES

[1] 1977. *Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts*. Vol. 1125 U.N.T.S. 3.

[2] 2019. Autonomous Vehicles | Self-Driving Vehicles Enacted Legislation. http://www.ncsl.org/research/transportation/autonomous-vehicles-self-driving-vehicles-enacted-legislation.aspx, last accessed 3 June, 2019.

[3] Thomas K Adams. 2001. Future warfare and the decline of human decisionmaking. *Parameters* 31, 4 (2001), 57–71.

[4] Naveed Akhtar and Ajmal Mian. 2018. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access* 6 (2018), 14410–14430.

[5] JÃijrgen Altmann and Frank Sauer. 2017. Autonomous Weapon Systems and Strategic Stability. *Survival* 59, 5 (2017), 117–142. https://doi.org/10.1080/00396338.2017.1375263 arXiv:https://doi.org/10.1080/00396338.2017.1375263

[6] Kenneth Anderson and Matthew C Waxman. 2013. Law and ethics for autonomous weapon systems: Why a ban won't work and how the laws of war can. (2013).

[7] Kenneth Anderson and Matthew C Waxman. 2017. Debating Autonomous Weapon Systems, their Ethics, and their Regulation under international law. (2017).

[8] Ronald Arkin. 2009. *Governing lethal behavior in autonomous robots*. Chapman and Hall/CRC.

[9] Ronald Arkin. 2018. Lethal Autonomous Systems and the Plight of the Non-combatant. In *The Political Economy of Robots*. Springer, 317–326.

[112] [37, p. 105], (âĂIJThe algorithm can return bad results if the underlying dataset violated assumptions of the algorithm – for example, if the data should have been represented by a curve, rather than a line. And the higher quality the input data, the higher quality the result")

[113] [81, note 130] (citing an interview with Leslie Pack Kaelbling at note 70)

[114] [16, p. 70], [30, p. 23]

[115] [64]

[116] [74, p. 107]

[117] [60] [11] [4] [12, p. 144]

[118] [77, p. 9] ("The consequences of failure with an autonomous car are far more potentially sever than a toaster failing to properly cook bread . . . [t]he hazard associated with an autonomous car driving on a closed-circuit track is much less severe than one driving through crowded city streets with pedestrians."

[119] [77, p. 10]

[120] [77, p. 11]

[121] Contact mines can be thought of as very crude AWS which will fire upon anyone that makes contact with them.

[122] [61, XIII, Art. 1]; see also [51, p. 282-283]

[10] Peter Asaro. 2012. On banning autonomous weapon systems: human rights, automation, and the dehumanization of lethal decision-making. *International Review of the Red Cross* 94, 886 (2012), 687–709.

[11] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. 2017. Synthesizing robust adversarial examples. *arXiv preprint arXiv:1707.07397* (2017).

[12] Marco Barreno, Blaine Nelson, Anthony D Joseph, and J Doug Tygar. 2010. The security of machine learning. *Machine Learning* 81, 2 (2010), 121–148.

[13] Antonio Bella, Cèsar Ferri, José Hernández-Orallo, and María José Ramírez-Quintana. 2010. Calibration of machine learning models. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*. IGI Global, 128–146.

[14] Bir Bhanu. 1986. Automatic target recognition: State of the art survey. *IEEE transactions on aerospace and electronic systems* 4 (1986), 364–379.

[15] Ingvild Bode and Hendrik Huelss. 2018. Autonomous weapons systems and changing norms in international relations. *Review of International Studies* 44, 3 (2018), 393âĂŞ413. https://doi.org/10.1017/S0260210517000614

[16] Vincent Boulanin and Maaike Verbruggen. 2017. Mapping the development of autonomy in weapon systems. *SIPRI Report. Accessed November* 14, 2018 (2017), 2017–11.

[17] C Bruderlein. 2013. Manual on International Law Applicable to Air and Missile Warfare. *The Program on Humanitarian Policy and Conflict Research at Harvard University* (2013).

[18] Michael W Byrnes. 2014. *Nightfall: Machine autonomy in air-to-air combat.* Technical Report. AIR UNIV MAXWELL AFB AL AIR FORCE RESEARCH INST.

[19] Ashton B. Carter. 2012. *Directive 3000.09, Autonomy in Weapon Systems* (updated 8 may, 2017 ed.). Department of Defense. https://fas.org/irp/doddir/dod/d3000_09.pdf, last accessed 29 May, 2019.

[20] Jean-Lou Chameau, William F Ballhaus, and Herbert Lin. 2014. *Emerging and readily available technologies and national security: A framework for addressing ethical, legal, and societal issues.* National Academies Press Washington, DC.

[21] Matthew Clark, Jim Alley, Paul Deal, Jeffrey DePriest, Eric Hansen, Connie Heitmeyer, Richard Nameth, Marc Steinberg, Craig Turner, Stuart Young, et al. 2015. *Autonomy Community of Interest (COI) Test and Evaluation, Verification and Validation (TEVV) Working Group: Technology Investment Strategy 2015-2018.* Technical Report. Office of the Assistant Secretary of Defense (Research and Engineering).

[22] Lee Copeland. 2004. *A practitioner's guide to software test design.* Artech House.

[23] Defense Advanced Research Projects Agency (DARPA). [n. d.]. Target Recognition and Adaption in Contested Environments (TRACE). https://www.darpa.mil/program/trace, last accessed 6 June, 2019.

[24] Defense Science Board, Department of Defense 2012. *Task Force Report: The Role of Autonomy in DoD Systems.* Defense Science Board, Department of Defense. https://fas.org/irp/agency/dod/dsb/autonomy.pdf, last accessed 30 May, 2019.

[25] Defense Science Board, Department of Defense 2016. *Report of the Defense Science Board Summer Study on Autonomy.* Defense Science Board, Department of Defense. https://fas.org/irp/agency/dod/dsb/autonomy-ss.pdf, last accessed 30 May, 2019.

[26] Department of Defense [n. d.]. *Unmanned Systems Integrated Roadmap: FY2013-2038* (reference number 14-s-0553 ed.). Department of Defense. https://archive.defense.gov/pubs/DOD-USRM-2013.pdf, last accessed May 7, 2018.

[27] Bonnie Docherty. 2012. *Losing humanity: The case against killer robots.* https://www.hrw.org/sites/default/files/reports/arms1112ForUpload_0_0.pdf, last accessed 28 May, 2019.

[28] Bonnie Lynn Docherty. 2015. *Mind the gap: The lack of accountability for killer robots.* Human Rights Watch.

[29] Charles J Dunlap Jr. 2016. Accountability and autonomous weapons: Much ado about nothing. *Temp. Int'l & Comp. LJ* 30 (2016), 63.

[30] Mica R Endsley. 2015. Autonomous Horizons: System Autonomy in the Air Force-A Path to the Future. *United States Air Force Office of the Chief Scientist, AF/ST TR* (2015), 15–01.

[31] Philip Feldman, Aaron Dant, and Aaron Massey. 2019. Integrating Artificial Intelligence into Weapon Systems. *arXiv preprint arXiv:1905.03899* (2019).

[32] United Nations Institute for Disarmament Research. 2017. *The Weaponization of Increasingly Autonomous Technologies: Concerns, Characteristics, and Definitional Approaches.* https://www.unog.ch/80256EE600585943/(httpPages)/8FA3C2562A60FF81C1257CE600393DF6?OpenDocument, last accessed 30 May, 2019.

[33] United States Air Force. 2009. *Unmanned Aircraft Systems Flight Plan 2009-2047.* https://fas.org/irp/program/collect/uas_2009.pdf, last accessed 5 June, 2019.

[34] Norman Friedman. 2006. *The naval institute guide to world naval weapon systems.* Naval Institute Press.

[35] Denise Garcia. 2015. Killer robots: Why the US should lead the ban. *Global Policy* 6, 1 (2015), 57–63.

[36] Daniel Gonzales and Sarah Harting. 2014. *Designing unmanned systems with greater autonomy: using a federated, partially open systems architecture approach.* Technical Report. RAND National Defense Research Institute.

[37] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning.* MIT press.

[38] Marcello Guarini and Paul Bello. 2012. Robotic warfare: some challenges in moving from noncivilian to civilian theaters. *Robot ethics: The ethical and social implications of robotics* 129 (2012), 136.

[39] Mark Gubrud. 2016. Why Should We Ban Autonomous Weapons? To Survive. *IEEE Spectrum* (2016).

[40] Charan Gudla, Md Shohel Rana, and Andrew H Sung. 2018. Defense Techniques Against Cyber Attacks on Unmanned Aerial Vehicles. In *Proceedings of the International Conference on Embedded Systems, Cyber-physical Systems, and Applications (ESCS).* The Steering Committee of The World Congress in Computer Science, Computer âĂę, 110–116.

[41] Jean-Marie Henckaerts and Louise Doswald-Beck. 2005. *Customary international humanitarian law.* Vol. 1. Cambridge University Press.

[42] Maziar Homayounnejad. 2018. The Lawful Use of Autonomous Weapon Systems for Targeted Strikes (Part 1): Concepts, Advantages and Technologies. (2018).

[43] International Committee of the Red Cross [n. d.]. *Customary IHL Database.* International Committee of the Red Cross.

[44] Aaron M Johnson and Sidney Axinn. 2013. The morality of autonomous robots. *Journal of Military Ethics* 12, 2 (2013), 129–141.

[45] K. Kalyanam, M. Pachter, M. Patzek, C. Rothwell, and S. Darbha. 2016. Optimal Human-Machine Teaming for a Sequential Inspection Operation. *IEEE Transactions on Human-Machine Systems* 46, 4 (Aug 2016), 557–568. https://doi.org/10.1109/THMS.2016.2519603

[46] Elsa Kania. 2018. China's AI Talent 'Arms Race'. *Australian Strategic Policy Institute: The Strategist* (April 2018). https://www.aspistrategist.org.au/chinas-ai-talent-arms-race/, last accessed 1 May, 2018.

[47] Jakob Kellenberger. 2011. International humanitarian law and new weapon technologies. *34th Round Table on current issues of international humanitarian law, San Remo* (2011), 8–10.

[48] Joel Lehman, Jeff Clune, Dusan Misevic, Christoph Adami, Lee Altenberg, Julie Beaulieu, Peter J Bentley, Samuel Bernard, Guillaume Beslon, David M Bryson, et al. 2018. The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities. *arXiv preprint arXiv:1803.03453* (2018).

[49] Dustin A Lewis, Gabriella Blum, and Naz K Modirzadeh. 2016. War-Algorithm Accountability. *Available at SSRN 2832734* (2016).

[50] Jean MacMillan, Eileen B Entin, and Daniel Serfaty. 1994. Operator reliance on automated support for target recognition. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 38. SAGE Publications Sage CA: Los Angeles, CA, 1285–1289.

[51] Gary E Marchant, Braden Allenby, Ronald C Arkin, Jason Borenstein, Lyn M Gaudet, Orde Kittrie, Patrick Lin, George R Lucas, Richard O'Meara, and Jared Silberman. 2015. International governance of autonomous military robots. *Handbook of Unmanned Aerial Vehicles* (2015), 2879–2910.

[52] Peter Margulies. 2016. Making Autonomous Weapons Accountable: Command Responsibility for Computer-Guided Lethal Force in Armed Conflicts. *Research Handbook on Remote Warfare, Edward Elgar Press, Jens David Ohlin ed* (2016).

[53] Per Martin-Löf. 1982. Constructive mathematics and computer programming. In *Studies in Logic and the Foundations of Mathematics.* Vol. 104. Elsevier, 153–175.

[54] SS Mehta, William MacKunis, and JW Curtis. 2011. Adaptive vision-based missile guidance in the presence of evasive target maneuvers. *IFAC Proceedings Volumes* 44, 1 (2011), 5471–5476.

[55] Charlie Miller and Chris Valasek. 2015. Remote exploitation of an unaltered passenger vehicle. *Black Hat USA* 2015 (2015), 91.

[56] Judge Advocate General (Col. Craig Miller). 2011. *Legal Reviews of Weapons and Cyber Capabilities* (af151-402 ed.). U.S. Air Force.

[57] Toby Murray and Paul van Oorschot. 2018. BP: Formal Proofs, the Fine Print and Side Effects. In *2018 IEEE Cybersecurity Development (SecDev).* IEEE, 1–10.

[58] Glenford J Myers, Tom Badgett, Todd M Thomas, and Corey Sandler. 2004. *The art of software testing.* Vol. 2. Wiley Online Library.

[59] Naval Research Advisory Committee 2017. *Autonomous and Unmanned Systems in the Department of the Navy.* Naval Research Advisory Committee. https://fas.org/irp/agency/navy/nrac-autonomous.pdf, last accessed 30 May, 2019.

[60] Anh Nguyen, Jason Yosinski, and Jeff Clune. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 427–436.

[61] Martinus Nijhoff. 1915. Convention Relative to the Laying of Automatic Submarine Contact Mines. In *Conventions and Declarations.* Springer, 46–48.

[62] International Institute of Humanitarian Law, Michael N Schmitt, Yoram Dinstein, and Charles HB Garraway. 2006. *The Manual on the law of non-international armed conflict, with commentary.* International Institute of Humanitarian Law.

[63] Chairman of the Joint Chiefs of Staff. 2009. *No-strike and the Collateral Damage Estimation Methodology* (cjcsi 3160.01 ed.). https://www.aclu.org/files/dronefoia/dod/drone_dod_3160_01.pdf, last accessed 30 May, 2019.

[64] Simon Parkin. 2015. Killer robots: The soldiers that never sleep. *BBC* (July 2015). http://www.bbc.com/future/story/20150715-killer-robots-the-soldiers-

that-never-sleep, last accessed 7 June, 2019.

[65] Kexin Pei, Yinzhi Cao, Junfeng Yang, and Suman Jana. 2017. DeepXplore: Automated Whitebox Testing of Deep Learning Systems. In *Proceedings of the 26th Symposium on Operating Systems Principles (SOSP '17)*. ACM, New York, NY, USA, 1–18. https://doi.org/10.1145/3132747.3132785

[66] John Pike. 2011. Samsung Techwin SGR-A1 Snetry Guard Robot. https://www.globalsecurity.org/military/world/rok/sgr-a1.htm, last accessed 6 June, 2019.

[67] Charles Piller. 2003. Vaunted Patriot Missile Has a 'Friendly Fire' Failing. (April 2003). http://articles.latimes.com/2003/apr/21/news/war-patriot21, last accessed 30 May, 2019.

[68] Mark Pomerleau. 2015. Army Wants Autonomous UAS for GPS-Denied Environments. *Defense Systems* (October 2015). https://defensesystems.com/articles/2015/10/22/army-autonomous-uas-gps-denied.aspx, last accessed 5 June, 2019.

[69] Aakanksha Rastogi and Kendall E Nygard. 2017. Cybersecurity Practices from a Software Engineering Perspective. In *Proceedings of the International Conference on Software Engineering Research and Practice (SERP)*. The Steering Committee of The World Congress in Computer Science, Computer âĂę, 51–55.

[70] James A Ratches. 2011. Review of current aided/automatic target acquisition technology for military target acquisition tasks. *Optical Engineering* 50, 7 (2011), 072001.

[71] Adam J. Reiner, Justin G. Hollands, and Greg A. Jamieson. 2017. Target Detection and Identification Performance Using an Automatic Target Detection System. *Human Factors* 59, 2 (2017), 242–258. https://doi.org/10.1177/0018720816670768 arXiv:https://doi.org/10.1177/0018720816670768 PMID: 27738280.

[72] Heather M Roff. 2014. The strategic robot problem: Lethal autonomous weapons in war. *Journal of Military Ethics* 13, 3 (2014), 211–227.

[73] Michael W Roth. 1990. Survey of neural network technology for automatic target recognition. *IEEE Transactions on neural networks* 1, 1 (1990), 28–43.

[74] Stuart Russell, Daniel Dewey, and Max Tegmark. 2015. Research priorities for robust and beneficial artificial intelligence. *Ai Magazine* 36, 4 (2015), 105–114.

[75] Stuart J Russell and Peter Norvig. 2016. *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,.

[76] Scott Savitz, Irv Blickstein, Peter Buryk, Robert W Button, Paul DeLuca, James Dryden, Jason Mastbaum, Jan Osburg, Philip Padilla, and Amy Potter. 2013. *US Navy employment options for unmanned surface vehicles (USVs)*. Technical Report. RAND National Defense Research Institute.

[77] Paul Scharre. 2016. *Autonomous weapons and operational risk*. Center for a New American Security Washington, DC.

[78] Michael N Schmitt. 2012. Autonomous weapon systems and international humanitarian law: a reply to the critics. *Harvard National Security Journal Feature (2013)* (2012). http://harvardnsj.org/wp-content/uploads/2013/02/Schmitt-Autonomous-Weapon-Systems-and-IHL-Final.pdf, last accessed 1 May, 2018.

[79] Michael N Schmitt. 2013. *Tallinn manual on the international law applicable to cyber warfare*. Cambridge University Press.

[80] Michael N Schmitt and Jeffrey S Thurnher. 2012. Out of the loop: autonomous weapon systems and the law of armed conflict. *Harv. Nat'l Sec. J.* 4 (2012), 231.

[81] Alan L Schuller. 2017. At the crossroads of control: The intersection of artificial intelligence in autonomous weapon systems with international humanitarian law. *Harv. Nat'l Sec. J.* 8 (2017), 379.

[82] Noah Shachtman. 2007. Robot cannon kills 9, wounds 14. *Wired.com* 18 (2007).

[83] Shai Shalev-Shwartz and Shai Ben-David. 2014. *Understanding machine learning: From theory to algorithms*. Cambridge university press.

[84] Yiftah Shapir. 2013. Lessons from the iron dome. *Military and Strategic Affairs* 5, 1 (2013), 81–94.

[85] Noel E Sharkey. 2012. The evitability of autonomous robot warfare. *International Review of the Red Cross* 94, 886 (2012), 787–799.

[86] Peter Warren Singer. 2009. *Wired for war: The robotics revolution and conflict in the 21st century*. Penguin.

[87] Elinor Sloan. 2015. Robotics at war. *Survival* 57, 5 (2015), 107–120. https://doi.org/10.1080/00396338.2015.1090133, last accessed 28 May, 2019.

[88] Robert D Sloane. 2015. Puzzles of Proportion and the Reasonable Military Commander: Reflections on the Law, Ethics, and Geopolitics of Proportionality. *Harv. Nat'l Sec. J.* 6 (2015), 299.

[89] Robert Sparrow. 2016. Robots and respect: Assessing the case against autonomous weapon systems. *Ethics & International Affairs* 30, 1 (2016), 93–116.

[90] Jeffrey S Thurnher. 2016. Means and Methods of the Future: Autonomous Systems. In *Targeting: The Challenges of Modern Warfare*. Springer, 177–199.

[91] Yuchi Tian, Kexin Pei, Suman Jana, and Baishakhi Ray. 2018. DeepTest: Automated Testing of Deep-neural-network-driven Autonomous Cars. In *Proceedings of the 40th International Conference on Software Engineering (ICSE '18)*. ACM, New York, NY, USA, 303–314. https://doi.org/10.1145/3180155.3180220

[92] Christopher P Toscano. 2015. Friend of Humans: An Argument for Developing Autonomous Weapons Systems. *J. Nat'l Sec. L. & Pol'y* 8 (2015), 189.

[93] Cumhur Erkan Tuncali, Georgios Fainekos, Hisahiro Ito, and James Kapinski. 2018. Simulation-based adversarial test generation for autonomous vehicles with machine learning components. In *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 1555–1562.

[94] United Nations 1998. *Rome Statute of the International Criminal Court*. Vol. 2187, U.N.T.S. 90. United Nations.

[95] Andrew Williams. 2015. Defining Autonomy in Systems: Challenges and Solutions. *Issues for Defence Policymakers* (2015), 27.

[96] Micah Zenko. 2013. *Reforming US drone strike policies*. Number 65. Council on Foreign Relations.