2017

Debating Autonomous Weapon Systems, Their Ethics, and Their Regulation Under International Law

Kenneth Anderson
ANDERS@WCL.AMERICAN.EDU

Matthew C. Waxman
Columbia Law School, mwaxma@law.columbia.edu

American University Washington College of Law

Washington College of Law Research Paper No. 2017-21

DEBATING AUTONOMOUS WEAPON SYSTEMS, THEIR ETHICS, AND THEIR REGULATION UNDER INTERNATIONAL LAW

Kenneth Anderson
Matthew C. Waxman

CHAPTER 45

DEBATING AUTONOMOUS WEAPON SYSTEMS, THEIR ETHICS, AND THEIR REGULATION UNDER INTERNATIONAL LAW

KENNETH ANDERSON
AND MATTHEW C. WAXMAN

1. Introduction

In November 2012, a high-profile public debate over the law and ethics of autonomous weapon systems (AWS) was kicked off by the release of two quite different documents by two quite different organizations.

The first of these is a policy memorandum on AWS issued by the US Department of Defense (DOD), under signature of then-Deputy Secretary of Defense (today Secretary of Defense) Ashton B Carter: the *DOD Directive: Autonomy in Weapon Systems*) (DOD Directive 2012). The Directive's fundamental purposes are, first, to establish DOD policy regarding the 'development and use of autonomous and semi-autonomous functions in weapon systems' and, second, to establish DOD

'guidelines designed to minimize the probability and consequences of failures in autonomous and semi-autonomous weapon systems that could lead to unintended engagements' (DOD Directive 2012: 1).

The Directive defines terms of art, and in particular the meaning of 'autonomous' and 'semi-autonomous' with respect to weapons and targeting in the international law of armed conflict (LOAC)—the body of international law, also known as international humanitarian law, regulating the conduct of warfare (DOD Directive 2012: 13–15). As a policy directive, it provides special requirements for AWS that might now or in the future be in development. But its substance draws upon long-standing DOD understandings of policy, law, and regulation of weapons development—understandings premised, in the Directive's language, on the requirement that AWS be designed to 'allow commanders and operators to exercise appropriate levels of human judgment over the use of force' (DOD Directive 2012: 2).

The gradual increase in the automation of weapon systems by the US military (taking the long historical view) stretches back at least to World War II and the early development of crude, mechanical feedback devices to improve the aim of anti-aircraft guns. Efforts to increase weapon automation are nothing new for the United States or the military establishments of other leading states. The Directive represents (for DOD, at least) an incremental step in policy guidance with respect to the processes for incorporating automation technologies of many kinds into weapon systems, including concerns about legality in particular battlefield uses, and training to ensure proper and effective use by its human operators. But the Directive's fundamental assumption (indeed DOD's fundamental assumption about *all* US military technologies) is that, in general, automation technologies will, and should, continue to be built into many new and existing weapon systems. While the Directive emphasizes practical and evolving policies to minimize risks and contingencies that any particular system might pose in any particular setting, it takes for granted that *of course* advancing automation, even to the point of 'autonomy' in some circumstances, is a legitimate aim in weapons design.

That assumption, however, is precisely what comes under challenge by a second high-profile document. It is a report and public call to action (also issued in November 2012) by the well-known international human rights organization, Human Rights Watch (HRW), *Losing Humanity: The Case against Killer Robots*. Its release was coordinated with, and the basis for, the launch of an international NGO campaign under the name Stop Killer Robots (2013). This new campaign draws on the now familiar model of the 1990s campaign to ban antipersonnel landmines. The Stop Killer Robots coalition, with HRW at its core and *Losing Humanity* as its manifesto, called in the most sweeping terms for a complete, pre-emptive ban on the development, production, transfer or sale, or use of any 'fully autonomous' AWS. It called for an international treaty to enact this sweeping, pre-emptive ban.

Losing Humanity is thus not primarily about debating DOD over the optimal prudent policies and legal interpretations to ensure that today's emerging weapon systems would be lawful in one battlefield setting or another. Rather (as this chapter discusses in sections 3 and 4), *Losing Humanity* asserts flatly that on its initial assessment, AWS—now or in the future, and no matter how advanced artificial intelligence (AI) might one day become—will not be able to comply with the requirements of LOAC. It is a remarkable claim, as critics of the report (including the present authors) have noted, because it contains sweeping assumptions about what technology will be capable of far into the future.

Today's international advocacy campaign, seeking a total, pre-emptive ban treaty, paints a dire picture of future warfare if current trends toward automation and artificial intelligence in weapon systems are not nipped in the bud today. Advocates make bold claims, implicitly or explicitly, about the future capabilities and limits of technology. And, deploying tropes from popular culture and science fiction (the catch-phrase 'Killer Robots,' to start with), this public advocacy urges that the way to prevent a future in which Killer Robots slip beyond human control is to enact today a complete ban on AWS.

Largely as a result of the *Losing Humanity* report and the coalition to Stop Killer Robots campaign, AWS and debates over its normative regulation, whether by a ban or something else, have been taken up by some states and United Nations officials at various UN forums. Beginning in 2013, several expert meetings on AWS have been convened under the aegis of the UN Convention on Certain Conventional Weapons (CCW 1980). Debate over the appropriate application of international law to AWS is far from static, however, and it is likely that positions and views by one actor or another in the international community that loom large today will have shifted even by the time this chapter reaches print.

The two foundational documents from 2012, viewed together, represent two main positions in today's debate over AWS: regulate AWS in ways already required in LOAC, on the one hand, or enact a complete ban on them, on the other. While other, more nuanced positions are emerging in the CCW meetings, these two represent major, fundamental legal alternatives. Yet the debate between these two has a certain 'ships passing in the night' quality to it; the DOD Directive is about practical, current technological R&D, while HRW's call for a total pre-emptive ban is grounded in considerable part on predictions about the long run. The 'risks' that each position sees in AWS are thus very different from each other, and likewise are the forms of norms and regulation that each side believes addresses those risks. Although some intellectual leaders of the debate have gone some distance over the last three years in bridging these conceptual gaps, at some fundamental level gaps are likely always to remain. It bears noting, however—in a Handbook about not just weapons and war, but about emerging technologies and their regulation more broadly—that many aspects of the AWS debate arise in other debates, over other technologies of automation, autonomy, and artificial intelligence.

The aim of this chapter is to provide a basic overview of the current normative debates over AWS, as well as the processes through which these debates are taking place at national and international levels.

2. What is an AWS and Why Are They Militarily Useful?

The DOD Directive defines an AWS as a 'weapon system that, once activated, can select and engage targets without further intervention by a human operator'. The Directive goes on to define a 'semi-autonomous weapon system' as one that 'once activated, is intended to only engage individual targets or specific target groups that have been selected by a human operator' (DOD Directive 2012: 13–14). *Losing Humanity* defines a 'fully autonomous weapon' as either (a) a weapon system in which human operators are 'out-of-the-loop,' meaning that the machine is 'capable of selecting targets and delivering force without any human input or interaction'; or (b) a weapon system in which human operators are 'on-the-loop,' meaning that, in principle, a human operator is able to override the machine's target selection and engagement, but in practical fact, the human operators are 'out-of-the-loop' because mechanisms of supervision are so limited (Losing Humanity 2012: 2). These definitions of AWS differ in certain important ways, but they share a common view of what makes a weapon system 'autonomous': it is a matter of whether a human operator realistically is able to override an activated machine in the core targeting functions of target selection and engagement.

In a highly abstract sense, any weapon that does not require a human operator could be regarded as an AWS. Antipersonnel landmines would be a simple example of a weapon that is triggered without a human operator in-the-loop or on-the-loop, but instead is triggered by pressure or movement. Conceptually, at least, such mines might fit the definition of autonomy. This is so, however, only if 'select' is construed to mean merely 'triggered,' rather than 'selection among' targets. 'Selection among' emphasizes that there is a machine-generated targeting decision made; some form of computational cognition, meaning some form of AI or logical reasoning, is inherently part of AWS in the contemporary debate. The debates over what constitutes an AWS leaves aside weapons, such as landmines, that are conceptually 'autonomous' merely because they are so technologically unsophisticated that that they cannot be aimed, and we leave those aside as well. AWS in today's debates refer to technologically sophisticated systems in which capabilities for 'selection among' is a specific

design aim for the weapon, and in which the machine possesses some decisional capability to 'select' and 'engage.'

A feature of the above definitions of AWS, however, is that they are essentially categorical: a weapon is or is not autonomous. If so, this would certainly make regulation of AWS easier. But the practical reality is that the line between 'highly automated' and 'autonomous' is not clear-cut. Rather, 'automation' describes a continuum, and there are various ways to define places along it. Terms like 'semi-autonomous,' 'human-in-the-loop' and 'human-on-the-loop' are used to convey different levels and configurations of machine-human interaction and degrees of independent machine decision-making. Autonomy is not just about machine capabilities, but instead about the capabilities and limitations of both machines and human operators, interacting together. Rather than debate categorical definitions, a better starting point is that new autonomous systems will develop incrementally as more functions (not just of the weapon but also of the platform, e.g. the vehicle or aircraft) are automated. Incremental increases in automation will alter the human-machine interaction, and 'functional' autonomy (whether believed to be good or bad) will have to be assessed on a detailed examination of each system, case-by-case, assessing machine functions, human operator functions, and how they interact.

This continuum offers many possible gradations of automation, autonomy, and human operator control. For example, 'intermediate' automation of weapon systems might pre-program the machine to look for certain enemy weapon signatures and to alert a human operator of the threat, who then decides whether or not to pull the trigger. At a further level of automation, the system might be set so that a human operator does not have to give an affirmative command, but instead merely decides whether to override and veto a machine-initiated attack. Perhaps next in the gradation of automation, the system would be designed with the capability to select a target and engage autonomously—but also programmed to wait and call for human authorization if it identifies the presence of civilians or alternatively, more sophisticated yet (perhaps into the level of science fiction, perhaps not) programmed to assess possible collateral damage and not engage if it is estimated to be above a certain level.

In some cases, a human operator might control only a single or very few sets of sensor and weapon units. In others, he or she might control or oversee an integrated network of many sensor and weapon units, which might operate largely autonomously, though with the supervisor able to intervene with respect to any of the weapon units. In still other cases, the move to automate the weapon system (or even give it autonomy) might be driven by automation of all the *other non-weapon* systems of the platform with which the weapon has to be coordinated (including the ability to operate at the same speed at which the rest of the platform operates). Eventually, these systems may reach the point of full autonomy for which, once activated, the human role is vanishingly small (*functionally* out-of-the-loop, even if technically on-the-loop), and it may depend heavily on the operators' training

and orders from higher commanders. The tipping point from a highly automated system to an 'autonomous' one is thus very thin, a continuum rather than distinct categories, a function of both machine and human parameters together and, in practice, an unstable dividing line as technology moves forward.

It is important to be clear as to what kinds of highly automated or even autonomous weapons exist today. Weapon systems that would be able to assess civilian status or estimate harm as part of their own independent targeting decisions do not exist today and research toward such capabilities currently remains in the realm of theory (see Arkin 2009). That said, several modern highly automated—and some would call them autonomous—weapon systems already exist. These are generally for use in battlefield environments such as naval encounters at sea where risks to civilians are small, and are generally limited to defensive contexts against other machines in which human operators activate and monitor the system and can override its operation. The US Patriot and Aegis anti-missile systems and Israel's Iron Dome anti-missile system are both leading examples, but they will not remain the only ones (See Schmitt and Thurnher 2013 explaining existing types of sophisticated highly automated or autonomous weapon systems). New autonomous weapon systems are gradually becoming incorporated into warfare as technology advances and capabilities increase, one small, automated step at a time.

Increasing automation in weapons technology results from advances in sensor and analytical capabilities and their integration into—and especially in response to the increasing tempo of—military operations. Some of this technology is highly particular to military battlefield requirements, but much of it is simply a military application of a new technology that finds wide uses in general society. For example, as private automobiles gradually incorporate new automation technologies—perhaps even a genuinely self-driving car—it would be inconceivable that military technologies would not incorporate them as well. This is no less true in the case of the targeting functions of weapons as for other weapon system functions, such as navigation or flying. Put another way, the ability to apply robotic systems to military functions depends upon advances and innovations in all the areas necessary to robotics—sensors, computational cognition and decision-making analytics, and physical movement and action mechanisms that make the machine robotic rather than a mere computer.

Increasing automation has other drivers, specific to the military, such as the desire among political leaders to protect not just one's own personnel on the battlefield but also civilian persons and property. Nonetheless, although automation will be a general feature across battlefield environments and weapon systems, genuine, full *autonomy* in weapons will likely remain rare for the foreseeable future, save in situations where special need justifies the expense and burden of weapons development. What are some of these special battlefield needs? A central and unsurprising one is the increasing tempo of military operations in which, other things being equal, the faster system wins the engagement (Marra and McNeil 2012). Automation permits

military systems of all kinds, not just weapons, to act more quickly than people might be able to do, in order to assess, calculate, and respond to a threat.

Moreover, speed, whether achieved through increased automation or genuine autonomy, might sometimes serve to make the deployment of force in battle more precise. By shortening the time, for example, between the positive identification of a target and its attack, there is less likelihood that the situation might have changed, that the target may have moved, or that civilians might have come into proximity. In the Libya hostilities in 2011, NATO-manned attack aircraft were reportedly too slow and had too little loiter time to permit accurate targeting of highly mobile vehicles on the ground in an urban battlefield with many civilians. In response, an appeal was made to the United States to initially supply surveillance drones, and then armed drones that could speed up the targeting process.[1] Some version of this will drive demand for automation, especially in competition with a sophisticated enemy's technology.

3. AWS Under the Existing Law of Armed Conflict

A peculiarity of the existing debates over AWS since 2012 is that some participants and certainly many ordinary observers appear to believe that AWS are not currently governed by existing international law, or at least not by a sufficiently robust body of international law. This misimpression lends greater weight and urgency to the call for some new law to address them, whether in the form of a ban treaty or a new protocol to the CCW. This is not the case, however; AWS of any kind—indeed, *all* weapons—are subject to LOAC. A requirement of LOAC is that states conduct legal reviews of weapons to determine if they are lawful weapons based on certain longstanding baseline requirements; if there are any legal restrictions on the battlefield environments for which they are lawful; or if there any legal limitations on how they can be used (see Thurnher 2013 for a non-technical exposition of these requirements). This matters because, despite the attention garnered by both the NGO campaign for a ban and demands for a new CCW protocol on AWS, there is already a robust legal process for the legal review of weapons.

Additionally, all the law of targeting and other fundamental rules of LOAC already apply to AWS, any form of automated weapon or any other form of weapon. Indeed, there are very few types of weapons, such as chemical weapons, that are governed by their own special set of international treaty rules. That sort of specialized regulation is the exception, not the rule. The vast majority of weapon systems—and the use of

those systems—are regulated by a well-established body of law that applies broadly, including to any new weapons that are invented.

There is a belief among some LOAC experts, perhaps particularly among LOAC lawyers in DOD and some other ministries of defence, that the whole debate over AWS has somehow got off on the wrong foot since 2012, with an assumption that this is legally ungoverned or only lightly governed space and therefore something must put in place. These LOAC lawyers might prefer to begin by asking what is wrong with the status quo of LOAC and its requirements, as they apply to AWS, now and in the future? And in what way has the existing process of legal weapons review been shown to be so inadequate that it needs to be replaced or supplemented by additional legal requirements—particularly given that for the most part, these remain *future* weapons with many unknown issues of design and performance? While it is certainly true, and recognized by LOAC lawyers, that legal weapon review of highly automated systems will require earlier review and legal guidance at the design stage, and quite possibly new forms of testing and verification of systems at a very granular level of a weapon system's engineering and software, in what way has the current system of legal review and regulation failed?

According to HRW, a weapon system that meets the definition of 'full autonomy' is inherently or inevitably illegal under LOAC. *Losing Humanity* states

> initial evaluation of fully autonomous weapons shows . . . such robots would appear to be incapable of abiding by key principles of international humanitarian law. They would be unable to follow the rules of distinction, proportionality, and military necessity . . . Full autonomy would strip civilians of protections from the effects of war that are guaranteed under the law (2012: 1–2).

Many LOAC experts—ourselves included—disagree that this is so as a matter of existing legal principle; the question, rather, is to examine any particular system and assess whether, and to what extent, it is, in fact, able to satisfy the requirements of LOAC in a given battlefield environment.[2] LOAC experts such as ourselves see arguments for a pre-emptive ban (or even greatly strengthened restrictions in a CCW protocol), moreover, as making of new law, not merely interpreting existing law, and doing so on the basis of certain factual predictions about the future of technology and how far it might advance in sophistication over the long-run. To understand this difference in perspectives, it is necessary to understand the basics of the existing LOAC framework (see Anderson, Reisner, and Waxman 2014, for a details discussion of these legal requirements as applied to AWS).

The legality of weapon systems turns on three fundamental rules. First, the weapon system cannot be indiscriminate by nature. This is not to ask whether there might be circumstances in which the weapon could not be aimed in a way that would comply with the legal requirement of 'distinction' between lawful military targets and civilians. That would be true of nearly any weapon, because any weapon could be deliberately misused. Rather, the rule runs to the nature of the

weapon *in the uses for which it was designed* or, as some authorities have put it, its 'normal' uses, i.e. the uses for which it was intended. This sets a very high bar for showing a weapon to be illegal as such; very few weapons are illegal per se, because they are indiscriminate by nature. The much more common problem arises when legal weapons are used in an indiscriminate manner—a serious violation of the law of armed conflict, certainly, but one that concerns the actual use of a weapon.

Second, a lawful weapon system cannot be 'of a nature' to cause 'unnecessary suffering or superfluous injury'. This provision aims to protect combatants from needless or inhumane suffering, such as shells filled with glass shards that would not be detectable by an x-ray of the wound. It is a rule that applies solely to combatants, not civilians (who are protected by other law of armed conflict provisions). Like the 'indiscriminate by nature' rule, it sets a high bar; this is unsurprising, given the many broad forms of violence that can lawfully be inflicted upon combatants in armed conflict.

Third, a weapon system can be deemed illegal per se if the harmful effects of the weapon are not capable of being 'controlled'. The rule against weapons with uncontrollable harmful effects is paradigmatically biological weapons, in which a virus or other biological agent cannot be controlled or contained; once released, it goes where it goes. Once again, even though many LOAC rules prevent the use of weapons in circumstances that might have uncontrolled effects, the bar to make the weapon itself illegal per se is high.

There is debate on this point, but many LOAC experts—including the authors of this chapter—believe that these rules do not render a weapon system illegal per se solely on account of it being autonomous (Schmitt and Thurnher 2013: 279, discussing that 'autonomous weapon systems are not unlawful per se'). Even if a weapon system is not per se illegal, however, it might still be prohibited in some—even most—battlefield environments, or in particular uses on a particular battlefield. But in other circumstances, the weapon might also be legal. With respect to new weapon technologies generally, the question is not whether the 'new technologies are good or bad in themselves, but instead what are the circumstances for their use' (ICRC 2011: 40).

Targeting law governs the circumstances of the use of lawful weapons and includes three fundamental rules: discrimination (or distinction), proportionality, and precautions in attack (see Boothby 2012 for a standard reference work with respect to targeting law). Distinction requires that a combatant, using reasonable judgment in the circumstances, distinguish between combatants and civilians, as well as between military and civilian objects. Although use of autonomous weapon systems is not illegal per se, a requirement for their lawful use—the ability to distinguish lawful from un-lawful targets—might vary enormously from one weapon system's technology to another. Some algorithms, sensors, or analytic capabilities might perform well, others poorly.

Such capabilities are measured with respect to particular uses in particular battlefield environments; the 'context and environment in which the weapon system operates play a significant role in this analysis' (Thurnher 2013). Air-to-air combat between military aircraft over the open ocean, for example, might one day take place between autonomous systems, as a result of the technological pressures for greater speed, ability to endure torque and inertial pressures, and so on. Distinction is highly unlikely to be an issue in that particular operational environment, however, because the combat environment would be lacking in civilians. Yet, there would be many operational environments in which meeting the requirements of distinction by a fully autonomous system would be very difficult—urban battlefield environments in which civilians and combatants are commingled, for example. This is not to say that autonomous systems are thereby totally illegal. Quite the opposite, in fact, as in some settings their use would be legal and in others illegal, depending on how technologies advance.

Proportionality requires that the reasonably anticipated military advantage of an operation be weighed against the reasonably anticipated civilian harms. As with the principle of distinction, there are operational settings—air-to-air combat over open water, tank warfare in remote uninhabited deserts, ship antimissile defence, undersea anti-submarine operations, for example—in which civilians are not likely to be present and which, in practical terms, do not require very much complex weighing of military advantage against civilian harms. Conversely, in settings such as urban warfare, proportionality is likely to pose very difficult conditions for machine programming, and it is widely recognized that whether and how such systems might one day be developed is simply an open question.

Precautions in attack require that an attacking party take feasible precautions in the circumstances to spare the civilian population. Precautions and feasibility, it bears stressing, however, are terms of art in the law of armed conflict that confer reasonable discretion on commanders undertaking attacks. The commander's obligation is grounded in reasonableness and good faith, and in 'planning, deciding upon or executing attacks, the decision taken by the person responsible has to be judged on the basis of all information available to him at the relevant time, and not on the basis of hindsight.'

In applying these rules to AWS, it is essential to understand that before an AWS—like any weapon system, including highly-automated or autonomous ones—is used in a military operation, human commanders and operators employing it generally will continue to be expected to exercise caution and judgment about such things as the likely presence of civilians and the possibility that they may be inadvertently injured or killed; expected military advantage; particular environmental conditions or features; the weapon's capabilities, limitations, and safety features; as well as many other factors. The many complex legal issues involved in such scenarios make it hard to draw general conclusions in the abstract. In many cases, however, although a weapon system may be autonomous, much of the requisite legal analysis

would still be conducted by human decision makers who must choose whether or not to use it in a specific situation. Whether LOAC legal requirements are satisfied in a given situation will therefore depend not simply on the machine's own programming and technical capabilities, but also on human judgments.

In the end, at least in the view of some LOAC experts, there is no reason *in principle* why a highly automated or autonomous system could not satisfy the requirements of targeting law (Schmitt and Thurnher 2013: 279). How likely it is that it will do so *in fact* is an open question—indeed, as leading AI robotics researcher Ronald Arkin says, it should be treated as a hypothesis to be proved or disproved by attempts to build machines able to do so (Arkin 2014[3]). In practical terms, however, weapon systems capable of full or semi-autonomy, and yet lacking the capacity to follow all the LOAC rules, could still find an important future role, insofar as they are set with a highly restrictive set of parameters on both target selection and engagement. For example, an AWS could be set with parameters far more restrictive than those required by law; instead of proportionality, it could be set not to fire if it detects *any* civilian presence. Being an AWS does not mean, in other words, that it cannot be used unless it is capable of following the LOAC rules entirely on its own. As participants in the AWS are gradually coming to recognize, the real topic of debate is not AWS set loose on battlefield somewhere, but instead the regulation of machine-human interactions.

4. Substantive Arguments for a Pre-emptive Ban on AWS

Although the existing legal framework that governs AWS and any other weapon system AWS is primarily LOAC and its weapons review process (and some other bodies of law, such as human rights law, might apply in some specific contexts), advocates of a complete ban generally advance several arguments in favour of a complete, pre-emptive ban. Three of the most prominent are taken up in this section: (a) AWS should be banned on the pure moral principle that machines should not make decisions to kill; this morally belongs to people, not robotic machines; (b) machine programming and AI will never reach a point of being capable of satisfying the requirements of LOAC, law, and ethics, and because they will not be able to do so even in the future, they should be pre-emptively banned today; and (c) AWS should be banned because machine decision-making undermines, or even removes, the possibility of holding anyone accountable in the way and to the

extent that, for example, an individual human soldier might be held accountable for unlawful or even criminal actions.

AWS should be banned on the moral principle that only human beings ought to make decisions deliberately to kill or not kill in war. This argument, which has been developed in its fullest and most sophisticated form by ethicist Wendell Wallach, is drawn from a view of human moral agency (see Wallach 2015). That is, a machine, no matter how sophisticated in its programming, cannot replace the presence of a true moral agent—a human being possessed of a conscience and the faculty of moral judgment. Only a human being possessing those qualities should make, or is fully morally capable of making, decisions and carrying them out in war as to when, where, and who to target with lethal force. A machine making and executing lethal targeting decisions on its own programming would be, Wallach says, inherently wrong (Wallach 2013).

This is a difficult argument to address because, as a deontological argument, it stops with a moral principle that one either accepts or does not accept. One does not have to be a full-blown consequentialist to believe that practical consequences matter in this as in other domains of human life. If it were shown to be true that machines of the future simply did a vastly better job of targeting, with large improvements in minimizing civilian harms or overall destruction on the battlefield, for example, surely there are other fundamental principles at work here.

One might acknowledge, in other words, that there is something of genuine moral concern about the intentional decision to take a life and kill in war that diminishes the dignity of that life, if simply determined by machine and then carried out by machine. But at some point, many of us would say that the moral value of dignity, even in being targeted, has to give way if the machine, when it kills or unleashes violent force, clearly uses less violence, kills fewer people, causes less collateral damage, and so on.

In the foreseeable future, we will be turning over more and more functions with life or death implications to machines—such as driverless cars or automated robot surgery technologies—not simply because they are more convenient but because they prove to be safer—and our basic notions about machine and human decision-making will evolve. A world that comes, if it does, to accept self-driving autonomous cars may also be one in which people expect those technologies to be applied to weapons and the battlefield as a matter of course, precisely because it regards them as better (and indeed might find the *failure* to use them morally objectionable).

The second argument is that AWS should be banned because machine learning and AI will never reach the point of being capable of satisfying the requirements of LOAC, law, and ethics. The underlying premise here is that machines will not be capable, now or in the future, of the requisite intuition, cognition, and judgment to comply with legal and ethical requirements—especially amid the fog of war. This is a core conviction held by many who favour a complete ban on autonomous lethal weapons. They generally deny that, even over time and, indeed, no matter how

much time or technological progress takes place, machine systems will ever manage to reach the point of satisfying legal and ethical codes and principles applicable in war. That is because, they believe, no machine system will ever be able to make appropriate judgments in the infinitely complex situations of warfare, or because no machine will ever have the capability, through its programming, to exhibit key elements of human emotion and affect that make human beings irreplaceable in making lethal decisions on the battlefield—compassion, empathy, and sympathy for other human beings (Losing Humanity 2012: 4).

These assessments are mostly empirical. Although many who embrace them might also finally rest upon moral premises denying in principle that a machine has the moral agency or moral psychology to make lethal decisions, they are framed here as distinct factual claims about the future evolution of technology. The argument rests on assumptions about how machine technology will actually evolve over decades or longer or, more frankly, how it will *not* evolve, as well as beliefs about the special nature of human beings and their emotional and affective abilities on the battlefield that no machine could ever exhibit, even over the course of technological evolution. It is as if to say that no autonomous lethal weapon system could ever pass an 'ethical Turing Test' under which, hypothetically, were a human and a machine hidden behind a veil, an objective observer could not tell which was which on the basis of their behaviours.

It is of course quite possible that fully autonomous weapons will never achieve the ability to meet the required standards, even far into the future. Yet, the radical scepticism that underlies the argument that they never will is unjustified. Research into the possibilities of autonomous machine decision-making, not just in weapons but across many human activities, is only a couple of decades old. No solid basis exists for such sweeping conclusions about the future of technology.

Moreover, we should not rule out in advance possibilities of positive technological outcomes—including the development of technologies of war that might reduce risks to civilians by making targeting more precise and firing decisions more controlled (especially compared to human-soldier failings that are so often exacerbated by fear, panic, vengeance, or other emotions—not to mention the limits of human senses and cognition).[4] It may well be, for instance, that weapons systems with greater and greater levels of automation can—in some battlefield contexts, and perhaps more and more over time—reduce misidentification of military targets, better detect or calculate possible collateral damage, or allow for using smaller quanta of force compared to human decision-making. True, relying on the promise of computer analytics and artificial intelligence risks pushing us down a slippery slope, propelled by the future promise of technology to overcome human failings rather than directly addressing the weaknesses of human moral psychology that lead to human moral and legal failings on the battlefield.

But the protection of civilians in war and reduction of the harms of war are 'not finally about the promotion of human virtue and the suppression of human vice'

as ends in themselves; human moral psychology is simply a means to those ends, and so is technology. If technology can further those goals more reliably and lessen dependence upon human beings with their virtues but also their moral frailties—by increasing precision; taking humans off the battlefield and reducing the pressures of human soldiers' interests in self-preservation; removing from battle the human soldier's emotions of fear, anger, and desire for revenge; and substituting a more easily disposable machine—this is to the good. Articulation of the tests of lawfulness that any autonomous lethal weapon system must ultimately meet helps channel technological development toward those protective ends of the law of armed conflict.

The last argument is that AWS should be banned because machine decision-making undermines, or even removes, the possibility of holding anyone accountable in the way and to the extent that an individual human soldier might be held accountable for unlawful or criminal actions in war. This is an objection particularly salient to those who put significant faith in accountability in war through mechanisms of individual criminal liability, such as international tribunals or other judicial mechanisms. One cannot hold a computer criminally liable or punish it. But to say that the machine's programmers can be held criminally liable for the machine's errors is not satisfactory, either, because although in some cases negligence in design might properly be thought to be so gross and severe as to warrant criminal penalties, the basic concept of civil product liability and design defect does not correspond to the what the actions would be if done by a human soldier on the battlefield—war crimes. Therefore, the difficulty is, as many have pointed out, that somehow human responsibility and accountability for the actions taken by the machine evaporate and disappear. The soldier in the field cannot be expected to understand in any serious way the programming of the machine; the designers and programmers operate on a completely different legal standard; the operational planners could not know exactly how the machine would perform in the fog of war; and finally, there might be no human actors left standing to hold accountable.

Putting aside whether there is a role of individual accountability in the use of AWS, however, it is important to understand that criminal liability is just one of many mechanisms for promoting and enforcing compliance with the laws of war (see Anderson and Waxman 2013 for an expanded discussion). Effective adherence to the law of armed conflict traditionally has come about through mechanisms of state (or armed party) responsibility. Responsibility on the front end, by a party to a conflict, is reflected in how a party plans its operations, through its rules of engagement and the 'operational law of war.' Although administrative and judicial mechanisms aimed at individuals play some important enforcement role, LOAC has its greatest effect and offers the greatest protections in war when it applies to a side as a whole and when it is enforced by sanctions and pressures that impose costs on parties to a conflict that breach their legal responsibilities under LOAC.

Hence, treating criminal liability as the presumptive mechanism of accountability risks blocking the development of machine systems that might, if successful,

overall reduce actual harms on the battlefield. It would be unfortunate indeed to sacrifice real-world gains consisting of reduced battlefield harm through machine systems (assuming there are any such gains) simply in order to satisfy an a priori principle that there always be a human to hold accountable.

5. The Processes of International Discussions Over AWS

The Stop Killer Robots campaign, distinguished by its willingness to frame its call for a ban in ways that explicitly draws on pop culture and sci-fi (no one could miss the references to The Terminator and Skynet, least of all the journalists who found the sci-fi framing of Killer Robots irresistible) were able to line up a variety of sympathetic countries to press for discussion of 'Killer Robots' in UN and other international community meetings and forums. Countries had a variety of reasons for wanting to open up a discussion besides a sincere belief that this technology needed international regulation beyond existing LOAC—wanting to slow down the US lead in autonomous military technologies, for example. But the issue was finally referred over to its logical forum—the mechanisms for review, drafting, and negotiation provided by the CCW. Periodic review meetings are built into the treaty, and this would be the normal place where such a discussion would go.

The CCW process began with the convening of several 'expert meetings', in which recognized experts in the field were invited in their individual capacities to open discussion of the issues. One of these was convened in spring 2014 and a second in spring 2015. Parallel to this intergovernmental treaty process, interested international NGOs (particularly member organizations of the Stop Killer Robots campaign) sponsored their own meetings, in a process of government/NGO parallel meetings that has become familiar since the 1990s and the international campaign to ban landmines.

It is not clear that an actual protocol on AWS will emerge from the CCW discussions, open for signature and ratification by states. We do not want to predict those kinds of substantive outcomes. However, it is very likely that pushing formalized international law—a treaty, a protocol—too quickly out of the box will fail, even with respect to a broadly shared interest among responsible states to ensure that clearly illegal autonomous weapons do not enter the battlefield. As we previously wrote with Daniel Reisner, a better approach to the regulation of AWS than quick promulgation of a new treaty is to:

reach consensus on some core minimum standards, but at the same time to retain some flexibility for international standards and requirements to evolve as technology evolves. Such an instrument is not likely to have compliance traction with States over time unless it largely codifies standards, practices, protocols and interpretations that States have converged upon over a period of actual development of systems (Anderson, Reisner, and Waxman 2014: 407).

The goals of legitimate normative regulation of AWS might well require an eventual treaty regime, and most likely in the form of a new protocol to the CCW convention. But the best way to achieve international rules with real adherence is to allow an extended period of gestation at the national level, within and informally among states' military establishments. Formal mechanisms for negotiating treaties create their own international political and diplomatic pressures. As we also previously wrote with Daniel Reisner, the process of convergence among responsible states is likely to be most successful if 'it takes place gradually through informal discussions among States, informed by sufficiently transparent and open sharing of relevant information, rather than through formal treaty negotiations that if initiated too early tend to lock States into rigid political positions' (Anderson, Reisner, and Waxman 2014: 407).

In other words, the best path forward is for a group of responsible states at or near the cutting edge of the relevant technologies—such as the United States, its NATO and Asian allies—to promote informal discussion about the evolving nature of the technologies at issue in autonomy, to focus on gradual and granular consideration of the legal, design, engineering, and strategic issues involved in autonomous weapons, and to foment, through the shared communications and discussions of leading states a set of common understandings, common standards, and proposals for best practices for such questions. It is slow and it is unapologetically state-centric, rather than being focused on international institutions or international NGOs and advocacy groups, but such an approach would adapt better to the evolution of the technologies involved in automation, autonomy, AI, and robotics.

A gestational period of best practices and informal state exchanges of legal interpretations over specific technologies and their uses has other advantages with respect to using process to advance more durable international norms for AWS. Discussions that are informal and directly among states, yet not part of an international 'negotiation,' and initially making no claim to creating new law, allow states to more freely expound, explore, evolve, and converge with others in their legal views. Moreover, rapid codification of treaty language, in advance of having actual designs and technology to address, inevitably favours categorical pronouncements, sweeping generalities and abstractions. What is needed, however, is not generalities, but concrete and specific norms emerging from concrete technologies and designs; LOAC already supplies the necessary general and abstract principles.

Among the many complex, concrete, and deeply technical issues that a gradual coalescence of best practices and informal norms might address, for example, is

how legal standards 'translate into terms of reliability engineering that are "testable, quantifiable, measurable, and reasonable"' (Anderson, Reisner, and Waxman 2014: 409, quoting Backstrom and Henderson 2012: 507). Such concrete and often technical matters (both in law and engineering) are the real issues for elaborating norms to govern AWS, not sweeping statements of first principles with which LOAC is already properly equipped. That said, however, the ability gradually to evolve widely shared international norms—norms that are concrete and often technical in nature—for AWS will necessarily depend on leading players, such as the US and its allies, being willing to see they have strategic interests in greater levels of transparency than they might otherwise prefer. Shared norms require at least some shared information.

6. Conclusions and the 'Meaningful Human Control' Standard

This discussion of AWS concludes by leaving the political, diplomatic, and negotiating issues of international treaty processes and returning to issues of regulatory substance. Discussions in the CCW meetings as well as in academic and policy forums have recently taken up the idea of a legal requirement of 'meaningful human control' (MHC) with respect to highly automated or autonomous weapon systems (see Horowitz and Scharre 2015). The idea is undeniably attractive—who would not want to require that machine weapon systems have appropriate and proper levels of human control? It is a concept found, for example, in the DOD Directive, where it is offered as one of the purposes for the special requirements imposed on AWS (DOD Directive 2012: 2).

There are, however, several reasons to be cautious about embracing MHC. The first is that the basis on which many parties seem to have embraced MHC as a way out of conceptual and political difficulties is because it offers strategic ambiguity. This principle can be read many different ways, and it begs questions of what is meant by 'meaningful' and what is meant by 'control'. Sometimes strategic ambiguity is a good idea in international politics, as a way of defusing tensions. But much of the time, strategic ambiguity ends in disappointment. It is not generally a good idea to embrace treaty phrasing about which the parties hold radically opposed or at least inconsistent ideas as to what it means. At some point, the contradictions can no longer be elided. This threatens to be the case with MHC—the US can make itself comfortable with the MHC standard because it says that, of course, its AWS have the proper amount of MHC; the Stop Killer Robots campaign and its

sympathetic governments will understand exactly the same language to mean that no truly autonomous system can ever have MHC; and a not-insignificant number of militarily advanced countries will urge everyone to embrace it (especially their rival, the United States) while secretly developing AWS with capabilities that will be known only when deployed.

Secondly, although some of its proponents view the MHC standard as flowing from LOAC, in some important respects it is quite at odds with the fundamental structure of LOAC, and its core principles of necessity, distinction, proportionality, and humanity. Each of these four principles is directed to, and evaluated by, its effects in armed conflict. Necessity authorizes violent hostilities, but also limits their effects. Distinction authorizes attacks on some persons, but also limits the effects of attacks, by limiting those who can be directly targeted. Proportionality authorizes attacks that might foreseeably lead to civilian harm or deaths, but it also limits the scope of permissible collateral harm. Humanity, in its LOAC meaning, seeks to relieve the burdens of those trapped in armed conflict, but it does so by reference to the effects that one action or another has on those people.

MHC is different. Insofar as its requirements are not already part of the others, it means obligations that are not finally measured by their effects, but instead by an insistence on a certain mode of weapons and hostilities. It is not a law of nature, however, that weapons that put a human being 'meaningfully' in control of it, in some fashion, necessarily do the best job at minimizing battlefield harms. It is not beyond possibility that at some point, in some circumstances, a machine might do it better, on its own.

It is not clear at this writing how or even whether the international debate over a new treaty will proceed; neither is it clear what arguments or concepts might come to dominate in that debate. Perhaps it will be MHC—or perhaps something else. As an alternative to MHC, however, we would suggest that debate over standards or rules for automated or autonomous systems should remain neutral as between human or machine, and should affirmatively reject any a priori preference for human over machine.

The principle of humanity is fundamental, but it refers, not to some idea that humans must operate weapons, but instead to the promotion of means or methods of warfare that best protect humanity within the lawful bounds of war, irrespective of whether the means to that end is human or machine or some combination of the two. Whether to favour an ethical insistence on an element of human control or to instead to favour strict neutrality as between 'who' or 'what', to be settled solely on the basis of effects and who or what performs better in minimizing battlefield harms: this is an essential debate today over the normative regulation of autonomous weapon systems, and surely not irrelevant to many other debates arising today over the law and ethics of automation and robotic technologies.[5]

Notes

1. See, e.g. Julian E. Barnes, 'US Launches Drone Strikes in Libya' (*Wall Street Journal*, 22 April 2011) A6 ('Drones have been used for reconnaissance missions from the start of the conflict, but in recent days, NATO commanders had asked the US to provide armed Predator strikes.')
2. For a general and legally thorough introduction to the legal requirements and processes of weapons review in international law, from a US perspective, see Hays Parks, 'Conventional Weapons and Weapons Reviews' (2005) 8 Yearbook of International Humanitarian Law 55.
3. For example, remarks by Ronald C Arkin in a public panel discussion on AWS, on (regarding the ability of machine systems gradually to advance in capabilities to make algorithmic determinations that would conform to LOAC requirements, not as a certainty or impossibility, but instead as a 'testable hypothesis'), University of Pennsylvania School of Law, Conference on Autonomous Weapon Systems, November 14, 2014.
4. As the ICRC put it in its 2011 'Contemporary Challenges of Armed Conflict' report, p. 40: 'After all, emotion, the loss of colleagues and personal self-interest is not an issue for a robot and the record of respect for [the law of armed conflict] by human soldiers is far from perfect, to say the least.' See also, 'Out of the Loop,' p. 249 ('Although emotions can restrain humans, it is equally true that they can unleash the basest instincts. From Rwanda and the Balkans to Darfur and Afghanistan, history is replete with tragic examples of unchecked emotions leading to horrendous suffering').
5. Readers interested in additional resources on AWS and their legal and ethical considerations are referred to the Center for a New American Security, 20YY Warfare Initiative, Ethical Autonomy Project, which since 2014 has maintained a running bibliography on AWS issues from the standpoints of technology, strategy, law, and ethics; website at http://cnas.org.

Further Reading and Research Sources

Anderson K and Waxman M, 'Law and Ethics for Autonomous Weapon Systems: Why a Ban Won't Work and How the Laws of War Can, National Security and Law Task Force Essay' (The Hoover Institution, Stanford University, 2013) <http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2250126> accessed 17 November 2015

Anderson K, Reisner D, and Waxman M, 'Adapting the Law of Armed Conflict to Autonomous Weapon Systems' (2014) 90 International Legal Studies 398 ('Adapting the Law of Armed Conflict')

Arkin R, *Governing Lethal Behavior in Autonomous Robots* (Chapman and Hall 2009)

Article 36, 'Home Page' (2015) <http://article36.org> accessed 17 November 2015

Asaro P, 'On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decisionmaking' (2012) 94 International Review of the Red Cross 687 <www.icrc.org/eng/resources/documents/article/review-2012/irrc-886-asaro.htm> accessed 17 November 2015

Backstrom A and Henderson I, 'New Capabilities in Warfare: An Overview of Contemporary Technological Developments and the Associated Legal and Engineering Issues in Article 36 Weapons Reviews' (2012) 94 International Review of the Red Cross 483

Boothby W, *The Law of Targeting* (OUP 2012)

Calo R, 'Robotics and the Lessons of Cyberlaw' (2015) 103 California Law Review 513

Campaign to Stop Killer Robots, 'About Us' (*Stop Killer Robots*, 2013) <https://www.stopkillerrobots.org/about-us/> accessed 17 November 2015

Center for a New American Security (CNAS), '20YY Future of Warfare Initiative, Ethical Autonomy Project" (2015) <www.cnas.org/research/us-defense-policy-and-military-operations/20yy-warfare-initiative> accessed 17 November 2015

Department of Defense, 'Autonomy in Weapon Systems' (2012) Directive Number 3000.09 ('Directive' or 'DOD Directive')

Horowitz M and Scharre P, 'Meaningful Human Control in Weapon Systems: A Primer' (Center for a New American Security, 2015) <www.cnas.org/sites/default/files/publications-pdf/Ethical_Autonomy_Working_Paper_031315.pdf> accessed 17 November 2015

Human Rights Watch, 'Arms' (2015) <www.hrw.org/topic/arms> accessed 17 November 2015

Human Rights Watch, *Losing Humanity: The Case Against Killer Robots* (International Human Rights Clinic at Harvard Law School, 2012) ('Losing Humanity')

International Committee for Robot Arms Control (ICRAC), <http://icrac.net> accessed 17 November 2015

International Committee of the Red Cross (ICRC), 'International Humanitarian Law and the Challenges of Contemporary Armed Conflict: Report Prepared for the 31st International Conference of the Red Cross and Red Crescent 40' (2011) ("Challenges of Contemporary Armed Conflict")

International Committee of the Red Cross (ICRC), 'New Technologies and Warfare' (2012) 94 (886) International Review of the Red Cross <www.icrc.org/eng/resources/international-review/review-886-new-technologies-warfare/review-886-all.pdf> accessed 17 November 2015

International Committee of the Red Cross (ICRC), 'New Technologies and IHL' (2015) <www.icrc.org/en/war-and-law/weapons/ihl-and-new-technologies> accessed 17 November 2015

Marra W and McNeil S, 'Automation and Autonomy in Advanced Machines: Understanding and Regulating Complex Systems' (Lawfare Research Paper Series, 1-2012, April 2012) http://lawfareblog.com

Parks H, 'Conventional Weapons and Weapons Reviews' (2005) 8 Yearbook of International Humanitarian Law 55

Schmitt M, 'Autonomous Weapon Systems and International Humanitarian Law: A Reply to the Critics' (2013) 4 Harvard National Security Journal <http://harvardnsj.org/2013/02/autonomous-weapon-systems-and-international-humanitarian-law-a-reply-to-the-critics/> accessed 17 November 2015

Schmitt M and Thurnher J, '"Out of the Loop": Autonomous Weapon Systems and the Law of Armed Conflict' (2013) 4 Harvard National Security Journal 234 <http://harvardnsj.org/2013/05/out-of-the-loop-autonomous-weapon-systems-and-the-law-of-armed-conflict/> accessed 17 November 2015

Stockton Center for the Study of International Law, 'Autonomous Weapons Forum' (US Naval War College, 2014) 90 International Legal Studies <www.usnwc.edu/Research-Gaming/

International-Law/New-International-Law-Studies-(Blue-Book)-Series/International-Law-Blue-Book-Articles.aspx?Volume=90> accessed 17 November 2015

Sharkey N, 'The Evitability of Autonomous Robot Warfare' (2012) 94 International Review of the Red Cross 787 www.icrc.org/eng/resources/documents/article/review-2012/irrc-886-sharkey.htm> accessed 17 November 2015

Thurnher J, 'The Law That Applies to Autonomous Weapon Systems' (American Society of International Law 2013) 17 Insights <www.asil.org/insights/volume/17/issue/4/law-applies-autonomous-weapon-systems> accessed 17 November 2015

United Nations, '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 (and Protocols) (As Amended on 21 December 2001)' (1980) 1342 UNTS 137 ('CCW')

US Department of Defense, 'Law of War Manual' (2015) <www.dod.mil/dodgc/images/law_war_manual15.pdf> accessed 17 November 2015

US Department of Defense, 'Task Force Report: the Role of Autonomy in DoD Systems' (Defense Science Board, 2012) <www.acq.osd.mil/dsb/reports/AutonomyReport.pdf> accessed 17 November 2015

Wallach W, 'Terminating the Terminator: What to Do About Autonomous Weapons' (Science Progress 2013)

Wallach W, A *Dangerous Master: How to Keep Technology from Slipping Beyond Our Control* (Basic Books 2015)

Wallach W and Allen C, *Moral Machines: Teaching Robots Right from Wrong* (OUP 2008)