American University Washington College of Law
Research Paper No. 2013-11

Columbia Public Law Research Paper

LAW AND ETHICS FOR AUTONOMOUS WEAPON SYSTEMS: WHY A BAN WON'T WORK AND HOW THE LAWS OF WAR CAN

Kenneth Anderson
Matthew C. Waxman

Electronic copy available at: http://ssrn.com/abstract=2250126

...

Legal and Ethical Requirements of Weapons

Arguments over the legal and ethical legitimacy of particular weapons (or their legitimate use)—poison as a weapon in war, for example, or the crossbow—go back very far in the history of warfare. Debates over autonomous robotic weapons (and also over UAVs) sometimes sound similar to those that arose with respect to technologies that emerged with the industrial era, such as the heated arguments of a century ago over submarines and military aviation. A core objection, then as now, was that they disrupted the prevailing norms of warfare by radically and illegitimately reducing combat risk to the party using them—an objection to "remoteness," joined to a claim (sometimes ethical, sometimes legal, and sometimes almost aesthetic) that it is unfair, dishonorable, cowardly, or not sporting to attack from a safe distance, whether with aircraft, submarines, or, today, a cruise missile, drone, or conceivably an autonomous weapon operating on its own. The law, to be sure, makes no requirement that sides limit themselves to the weapons available to the other side; weapons superiority is perfectly lawful and indeed assumed as part of military necessity.

Emergence of a new weapon often sparks an insistence in some quarters that the weapon is ethically and legally abhorrent and should be prohibited by law. Yet historical reality is that if a new weapon system greatly advantages a side, the tendency is for it gradually to be adopted by others that perceive they can benefit from it as well. In some cases, legal prohibitions on the weapon system as such erode, as happened with military submarines and aircraft; what survives is a set of legal rules for the use of the new weapon, with greater or lesser specificity. In other cases, legal prohibitions gain hold. The ban on poison gas, for example, has survived in one form or another with very considerable effectiveness over the history of the 20th century. The most recent, and many would say quite successful, ban on a weapon—the Ottawa Convention banning antipersonnel landmines—is very much the exception rather than the rule.

Where in this long history of new weapons and attempts to regulate them ethically and legally will autonomous weapons fit? What are the features of autonomous robotic weapons that raise ethical and legal concerns? How should they be addressed, as a matter of law and process—by treaty, for example—or by some other means? And what difference does the incremental shift from increasing automation to autonomy mean, if anything, to the legal and ethical concerns?

One answer to these questions is to wait and see: it is too early to know where the technology will go, so the debate over ethical and legal principles for robotic autonomous weapons should be deferred until a system is at hand. Otherwise it is just an exercise in science fiction. One does not have to embrace a ban on autonomous systems and their development to say that the wait-and-see view is shortsighted and faulty, however. Not all of the important innovations in autonomous weapons are far off on the horizon; some are possible now or will be in the near-term. Some of these innovations also raise serious questions of law and ethics even at their current research and development stage.

This is the time—before technologies and weapons development have become "hardened" in a particular path and before their design architecture is entrenched and difficult to change—to take account of the law and ethics that ought to inform and govern autonomous weapons systems, as technology and innovation let slip the robots of war. This is also the time—before ethical and legal understandings of autonomous weapon systems become hardened in the eyes of key constituents of the international system—to propose and defend a framework for evaluating them that advances simultaneously strategic and moral interests.

A recent and widely circulated report from the British Ministry of Defense on the future of unmanned systems made this point forcefully. It noted that as "technology matures and new capabilities appear, policy-makers will need to be aware of the potential legal issues and take advice at a very early stage of any new system's procurement cycle."[20] This is so whether the system is intended in the first place to be highly automated but not fully autonomous; is intended from the beginning to be autonomous in either target selection or engagement with a selected target, or both; or turns out upon review to have unanticipated or unintended autonomous functions (perhaps in how it interoperates with other systems).[21]

If early and continuing consideration of fundamental normative principles is a correct ethical and policy approach to the development of autonomous weapon technologies over time, what are the legal requirements that a weapon system must meet? There are three substantive rules: two drawn from the law of weapons, addressing the lawfulness of the weapon as such, and a third from the law of targeting, addressing the lawful uses of the weapon (and any limitations) in the conduct of hostilities. These three rules fit under a review process framework that is also a requirement of law in order to field a new weapon.

Article 36 of the 1977 Additional Protocol to the Geneva Conventions provides the framework for the legal review of new weapons. (The United States, while not party to Protocol I, very likely accepts the provisions under discussion here as customary international law binding on all parties.) In the "study, development, acquisition or adoption of a new weapon, means or method of warfare," says Article 36, a party is "under an obligation to determine whether its employment would, in some or all circumstances, be prohibited," either by Protocol I or by "any other rule of international law applicable" to such party. The United States, in its actual practice, has long undertaken extensive legal review of new weapon systems, for which the provisions of Protocol I are merely the starting point of a highly detailed legal and administrative review process.[22] In the past two decades, U.S. Defense Department lawyers have rejected proposed new weapons, including blinding laser weapons in the 1990s, and in recent years, reportedly, various cutting-edge cyber-technologies for use in cyber-conflict.[23]

The two substantive rules drawn from weapons law that must be part of any Article 36 legal review are, first, the rule against inherently indiscriminate weapons (Article 54(b)(4) of Protocol I) and, second, the rule against weapons that cause unnecessary suffering or superfluous injury (Article 35(2) of Protocol I). These two rules describe the lawfulness of the weapon itself. A weapon is deemed indiscriminate by its very nature if it cannot be aimed at a specific target and is as likely to hit civilians as combatants.[24] Any autonomous weapon system must comply with this rule; but the mere feature of autonomy as such does not per se rule compliance. In other words, the fact that an autonomous weapon system rather than a human being might make the final targeting decision would not in itself render the system indiscriminate by nature, so long as it is possible to supply the autonomous system with sufficiently reliable targeting information to ensure it can be aimed at a lawful target.[25] The second rule on the law of weapons prohibits a weapon as such if its nature is to cause unnecessary suffering or superfluous injury to combatants—weapons such as warheads filled with glass that could not be detected with X-rays and so, for example, would unnecessarily complicate treatment of the wounded. Again, the fact that an autonomous weapon system selects the target or undertakes the attack does not violate the rule.[26]

In sum, although specific circumstances might arise in which an autonomous weapon system would constitute an indiscriminate weapon by its nature, the fact of autonomy itself—the fact of machine selection of target and engagement with it—does not violate the law of armed conflict. Indeed, as the following sections discuss, it might turn out over time that for some purposes and forms of attack or defense, autonomous weapons may be able to be more discriminating and precise than human beings.

A Legal and Ethical Framework for Autonomous Weapon Systems

Even if an autonomous weapon is not illegal on account of its autonomy, targeting law still governs any particular use of that system. The baseline legal principles respecting the use of any weapon in hostilities are distinction and proportionality.

Distinction requires that a combatant, using reasonable judgment in the circumstances, distinguish between combatants and civilians, as well as military and civilian objects. The most significant effect of this targeting rule is that although use of autonomous weapon systems is not illegal per se, their lawful use—the ability to distinguish lawful from unlawful targets—might vary enormously from one system's technology to another. Some algorithms, sensors, or analytic capabilities might perform well; others badly. If one is a lawyer in a ministry of defense somewhere in the world, whose job is to evaluate the lawfulness of such weapon systems, including where and under what operational conditions they can lawfully be used, it will be indispensable to be able to test each system to know what it can and cannot do and under what circumstances.

The conditions in which the autonomous system will be used—the battlefield environment and operational settings—will be an important consideration not just in determining whether the system is lawful generally, but also in identifying where and under what legal limitations its use would be lawful. An autonomous system might be deemed inadequate and unlawful in its ability to distinguish civilians from combatants in operational conditions of infantry urban warfare, for example, but lawful in battlefield environments with few if any civilians present.

The proportionality rule requires that even if a weapon meets the test of distinction, any use of a weapon must also involve evaluation that sets the anticipated military advantage to be gained against the anticipated civilian harm (to civilian persons or objects). The harm to civilians must not be excessive relative to the expected military gain.[27] This calculus for taking into account civilian collateral damage is difficult for many reasons. While everyone agrees that civilian harm should not be excessive in relation to military advantages gained, the comparison is apples and oranges. Although there is a general sense that such excess can be determined in truly gross cases, there is no accepted formula that gives determinate outcomes in specific cases. Some military lawyers proceed largely casuistically, building on what was done in prior situations and examining similarities and differences. Difficult or not, proportionality is a fundamental requirement of the law and any completely autonomous weapon system would have to be able to address proportionality as well as distinction—though, as with distinction, reasonable judgments of proportionality would be highly dependent on the operational environment and battlefield in which the machine was deployed. Again, assessing proportionality is one thing in close-in infantry urban warfare, but altogether different in undersea, machine-on-machine war where few if any civilians are present.

These (and others could be added, such as precautions in attack) are daunting legal and ethical hurdles if the aim is to create a fully autonomous weapon, capable of matching or surpassing the standards we would expect of a human soldier performing the same function, in all battlefield circumstances and operational environments. Important work has been done in the past several years on whether and how these problems could be resolved as a matter of machine programming—algorithms, for example, that might capture these two fundamental principles of distinction and proportionality.[28] These research questions, unsurprisingly, are sharply debated, even as to whether machine programming could ever fully or adequately reproduce the results of human judgment in these fundamental law of war matters.[29]

In order to program distinction, for example, one could theoretically start with categories and samples of lawful targets—programmed targets could include persons or weapons that are firing at the robot—and gradually build upwards toward inductive reasoning about characteristics of lawful targets not already on the list. Or one could envision systems that integrate sensors and recognition processes to identify specific, known enemy combatants, perhaps also carrying weapons. Designers might use case-based reasoning and faster-than-real-time simulations to improve a machine's inductive learning. Perhaps these and other tools for distinguishing lawful from unlawful targets might gradually become good enough to be reasonable substitutes or even better than humans in the future—though perhaps not, or only for very limited operational environments and circumstances. Perhaps they are only appropriate not in a fully autonomous mode, but as a means of recommending and cuing up proposed targets for the final judgment of a human operator.

Proportionality, for its part, is a relative judgment that is easy to state as an abstract rule but very challenging to program in a machine: measure anticipated civilian harm and measure military advantage; subtract and measure the balance against some determined standard of "excessive"; if excessive, do not attack an otherwise lawful target. From a programming standpoint, this requires attaching values to various targets, objects, and categories of human beings, and calculating probabilistic assessments based on many complex contextual factors. It might also include inductive machine learning from human examples of judgments about proportionality, seeking to extract practical heuristics from them. Moreover, a machine's distinction and proportionality judgments will be probabilistic (as they are for humans, too), and an important legal, ethical, and policy question for any such system will be where to set the required confidence thresholds (again, this is so for humans, too). The appropriate threshold might—almost certainly will—also vary depending on specific operational context and mission (for example, permitting a system to fire only when anticipated collateral damage is close to zero and anticipated military gain is high). Although engineers and programmers might one day be able to do this well, today they are a long way off, even in basic conceptualizing, from creating systems sufficiently sophisticated to perform this function in situations densely populated with civilians and civilian property.

Yet difficult as these judgments seem to any experienced law-of-war lawyer, they (and others) are the fundamental conditions that the ethical and lawful autonomous weapon would have to satisfy and therefore what a programming development effort must take into account (along with the adequacy of sensor systems and weaponry). The ethical and legal engineering matter every bit as much as the mechanical or software engineering do. Legal and ethical assessments of autonomous systems will not be simply binary—that is, a system is either acceptable or unacceptable. Some systems might be capable of sufficient distinction and proportionality to be used only in environments likely to contain few or no civilians, or only for certain functions likely to pose little risk of damage to civilian property, or they would be intended for machine-on-machine operations, so that humans would not be an object of attack in any case. Autonomous weapons, like other sophisticated weapon systems, would be designed for specific purposes and operational environments.

"Programming the laws of war" at their conceptually most difficult (sophisticated proportionality, for example) is a vital research project over the long run, in order to find the greatest gains that can be had from machine decision-making within the law. Yet with respect to fielding autonomous weapons in the nearer term, some of the most difficult challenges to designing the "perfect" autonomous weapon (able to make judgments of distinction and proportionality better than expert humans) can be avoided for now. Instead of relying on complex balancing assessments of probabilistic valuations of advantages and harms, early generations of autonomous systems deployed by legally and ethically responsible states will likely be programmed with hard rules: say, that the weapon system may not fire (or must seek human operator approval) if it identifies any human within a specified radius of the target. The science-fiction problems do need to be addressed, but they do not need to be solved in order to field "autonomous" weapons that are clearly lawful because they are much more circumscribed in their operations than the full extent of the law would allow.

Four Major Arguments Against Autonomy in Weapons

If this is the cautiously optimistic vision of autonomous weapon systems, say, decades or even several generations from now, however, it is subject at the present time to four major objections. They are arguments against autonomy in weapon systems at all; for each of them, weapon autonomy as such is the problem and no mere regulation of autonomous weapons could ever be satisfactory. As "universal" objections to autonomy in weapons as such, unsurprisingly each of these figures prominently in calls for a sweeping preemptive ban on autonomous weapons or, as some advocates have said, even on the development of technologies or components of automation that could lead to fully autonomous lethal weapon systems.

The first is a broad claim that machine programming will never reach the point of satisfying the fundamental ethical and legal principles required to field a lawful autonomous lethal weapon.[30] Artificial intelligence has overpromised before, and once into the weeds of the judgments that these broad principles imply, the requisite intuition, cognition, affect, and judgment look ever more marvelously and uniquely human—especially amid the fog of war.[31] This is a core conviction held by many who favor 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 the legal or moral requirements of the laws of war. That is because, they believe, no machine system can, through its programming, replace the 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.

These assessments are mostly empirical. Although many who embrace them might also finally rest upon hidden moral premises denying in principle that a machine has the moral agency or moral psychology to make lethal decisions (a separate argument discussed next), 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, 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 behaviors.[32]

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; it is quite possible that no autonomous lethal weapon will pass the "ethical Turing Test." Yet the radical skepticism that underlies the argument 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 basis exists for such sweeping conclusions about the future of technology.

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). 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 addressing the weaknesses of human moral psychology directly.

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 vice as ends in themselves; human moral psychology is a means to those ends. 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, 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 second major argument against development of autonomous weapon systems is a moral one: it is simply wrong per se to take the human moral agent entirely out of the firing loop. A machine, no matter how good, cannot completely replace the presence of a true moral agent in the form of a human being possessed of a conscience and the faculty of moral judgment (even if flawed in human ways).[33] Perhaps we should make a societal choice, independent of consequences, and independent of how well machines might someday perform these tasks, to declare that the application of lethal violence should in no circumstances ever be delegated entirely to a machine.

This is a difficult argument to address, since it stops with a moral principle that one either accepts or does not accept. Whatever merit it has today, one must consider that 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 automatic 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 is likely to be one in which people expect those technologies to be applied to weapons and the battlefield, precisely because it regards them as better (and indeed might find morally objectionable the failure to use them). Moreover, this objection raises a further question as to what constitutes the tipping point into impermissible autonomy given that the automation of weapons' functions is likely to occur in incremental steps—there are many steps along the way to full autonomy at which the machine's contribution to a lethal decision would far exceed a human's.

The fundamental moral lesson that the current ban campaign seems to have drawn from the earlier campaign to ban landmines is that a weapon that is not aimed by a human being at the point of firing is inherently wrong—for the reason of not having a human fire it. The alternative and, in our view, correct deontological principle is that any weapon that undertakes target selection and firing at targets, irrespective of mechanism or agency, must be capable of meeting the fundamental requirements of the laws of war. We do not accept that a machine-made lethal decision is always and necessarily *mala in se*; and if that is ever accepted as a general moral principle, it promises to raise difficulties for machine systems far beyond weapons.[34] Machine-versus-human for these weapons-related activities might someday turn out to be morally incidental—a contingent, rather than morally inherent, feature of a weapon and its use. What matters morally is the ability consistently to behave in a certain way and to a specified level of performance. The "package" it comes in, machine or human, is not the deepest moral principle.

A third major argument holds that autonomous weapon systems that remove the human being from the firing loop are unacceptable because they undermine the possibility of holding anyone accountable for what, if done by a human soldier, might be a war crime.[35] If the decision to fire is taken by a machine, who should be held responsible—criminally or otherwise—for mistakes? The soldier who allowed the weapon system to be used where it made a bad decision?[36] The commander who chose to employ it on the battlefield? The engineer or designer who programmed it in the first place?[37]

This is an objection particularly salient to those who put significant faith in law of armed conflict accountability through mechanisms of individual criminal liability, especially international tribunals or other judicial mechanisms. In some instances, to be sure, there will still be human decision-makers who can be held individually accountable for grossly improper design or deployment decisions. Indeed, those involved in programming autonomous weapons systems or their settings for particular circumstances will confront directly very difficult value judgments that may even be subjected to new forms of close scrutiny because they are documented in computer code rather than individual minds. The recent Defense Department policy directive is innovative in its insistence upon training human soldiers in the proper operation of systems, including choosing whether an automated or autonomous system is appropriate to particular battlefield conditions. These provisions in the directive point to practical ways to strengthen human accountability as automated systems are brought online.

Narrow focus on post-hoc judicial accountability for individuals in war is a mistake in any case. It is just one of many mechanisms for promoting and enforcing compliance with the laws of war. Excessive devotion to individual criminal liability as the presumptive mechanism of accountability risks blocking development of machine systems that might, if successful, reduce actual harms to soldiers as well as to civilians on or near the battlefield. Effective adherence to the law of armed conflict traditionally has been 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, the law of armed conflict has its greatest effect and offers the greatest protections in war when it applies to a side as a whole.

It would be unfortunate 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. It would be better to adapt mechanisms of collective responsibility borne by a "side" in war, through its operational planning and law, including legal reviews of weapon systems and justification of their use in particular operational conditions.

Finally, the long-run development of autonomous weapon systems faces an objection that by removing human soldiers from risk and reducing harm to civilians through greater precision, the disincentive to resort to armed force is diminished.[38] The two features of precision and remoteness (especially in combination) that make war less damaging in its effects are the same two features that make it easier to undertake. Automation, and finally autonomy, might well carry these features to whole new levels. The result might be a greater propensity to wage war or to resort to military force.[39]

This argument is invoked frequently, though it is morally and practically misconceived. To start with, to the extent it entails deliberately foregoing available protections for civilians or soldiers in war, for fear that political leaders would resort to war more than they ought, morally amounts to holding those endangered humans as hostages, mere means to pressure political leaders.

Furthermore, this concern is not special to autonomous weapons. The same objection has already been made with respect to remotely-piloted UAVs and high-altitude bombing before that. Generally, it can be made with respect to any technological development that either reduces risk to one's own forces or reduces risk to civilians, or both.[40] Yet it is not generally accepted as a moral proposition in other contexts of war—indeed, quite the other way around. All things equal, as a moral matter (even where the law does not require it), sides should strive to use the most sparing methods and means of war; there is no good reason why this obvious moral notion should suddenly be turned on its head.

The argument rests on further questionable assumptions, not just about morality and using people as mere means, but about the "optimal" level of force and whether it is even a meaningful idea in a struggle between two sides with incompatible aims. Force might conceivably be used "too" often, but sometimes it is necessary to combat aggression, atrocities, or threats of the same. Technologies that reduce risks to human soldiers (or civilians) may also facilitate desirable—even morally imperative—military action. More broadly, trying to reduce warfare and the resort to force by seeking to control the availability of certain weapon systems—particularly those that might make war less risky or less damaging in its conduct—is the tail wagging the dog: how much war occurs and at what intensity and level of destructiveness depends on a slew of much more significant factors, ranging across law, politics, diplomacy, the effectiveness of international institutions, the nature of threats, and many other things.