A NATIONAL SECURITY AND LAW ESSAY

Law and Ethics for Autonomous Weapon Systems

Why a Ban Won't Work and How the Laws of War Can

by Kenneth Anderson and Matthew Waxman

Jean Perkins Task Force on National Security and Law
www.hoover.org/taskforces/national-security

task force on national security and law

Introduction

Public debate is heating up over the future development of autonomous weapon systems.[1] Some concerned critics portray that future, often invoking science-fiction imagery, as a plain choice between a world in which those systems are banned outright and a world of legal void and ethical collapse on the battlefield.[2] Yet an outright ban on autonomous weapon systems, even if it could be made effective, trades whatever risks autonomous weapon systems might pose in war for the real, if less visible, risk of failing to develop forms of automation that might make the use of force more precise and less harmful for civilians caught near it. Grounded in a more realistic assessment of technology—acknowledging what is known and what is yet unknown—as well as the interests of the many international and domestic actors involved, this paper outlines a practical alternative: the gradual evolution of codes of conduct based on traditional legal and ethical principles governing weapons and warfare.

A November 2012 U.S. Department of Defense policy directive on the topic defines an "autonomous weapon system" as one "that, once activated, can select and engage targets without further intervention by a human operator."[3] Some such systems already exist, in limited defensive contexts and for which human operators activate the system and can override its operation, such as the U.S. Patriot and Phalanx anti-missile systems and Israel's Iron Dome anti-missile system.[4] Others are reportedly close at hand, such as a lethal sentry robot designed in South Korea that might be used against hostile intruders near its border.[5] And many more lie ahead in a future that is less and less distant.[6]

Autonomous weapon systems are entering the battlefields of the future, but they are doing so one small automated step at a time. The steady march of automation (in different operational functions on battlefields that themselves vary greatly) is frankly inevitable, in part because it is not merely a feature of weapons technology, but of technology generally—anything from self-driving cars to high frequency trading programs dealing in the financial markets in nanosecond intervals too swift for human intervention. Automation in weapons technology is also inevitable as a response to the increasing tempo of military operations and political pressures to protect not just one's own personnel but also civilian persons and property.

Just as increased automation in many fields is inevitable, automation in weapons will occur, and is occurring, incrementally. Autonomy in weapon systems might positively promote the aims of the laws of war in some technological configurations and operational circumstances—but not in others. While autonomy for weapon systems for its own sake is not a strategic goal of the U.S. military in weapons design, many factors will push automation in some circumstances into genuine weapon autonomy. In some operational circumstances, as both the decision-making power of machines and the tempo of operations potentially increase, that human role will be likely to slowly diminish.[7] Though automation will be a general feature across battlefield environments and weapon systems, genuine *autonomy* in weapons will probably remain rare for the foreseeable future and driven by special factors such as reaction speeds and the tempo of particular kinds of operations.

The combination of *inevitable* and *incremental* development of automated systems to the point, in some cases, of genuine weapon autonomy raises not only complex strategic and operational questions but also profound legal and ethical ones. Advances in automation toward autonomy raise possibilities of tradeoffs between substantial gains in the ability to make war less destructive and harmful, primarily through the benefits of greater automated weapon precision, on the one hand, and significant dangers that military force will be used more destructively and with less ethical virtue on the other. Advancing automation raises cautious hopes among many, but also profound fears among some, about the future of war.

Highlighting that the incremental automation of some weapon systems in some forms is inevitable is not a veiled threat in the guise of a prediction. Nor is it meant to suggest that the path of these technologies is beyond rationally ethical human control. On the contrary, we believe that sensible regulation of these systems as they emerge is both possible and desirable. But regulation has to emerge along with the technologies themselves, and against the backdrop of a world that will likely come to adopt, over coming decades, technologies of autonomy for self-driving vehicles, or advanced nursing or elder-care robots,

or any number of other technologies that evolve from being increasingly automated to performing some functions with genuine machine autonomy. With many of these technologies, however, the machine will take actions with potentially lethal consequences—and it will happen largely because people conclude over successive decades that machines can sometimes do certain tasks better than humans can.

Recognizing the incremental evolution of these technologies is key to addressing the legal and ethical dilemmas associated with their inevitability. This is particularly so for the United States, because it is a leader both in developing and deploying new weapon technologies in ways visible to the world, and in formulating and conducting the legal, policy, and ethical processes of formal weapons reviews, as required by the laws of armed conflict. The certain yet gradual development and deployment of these systems, as well as the humanitarian advantages that may be created by the precision of some systems, make some proposed responses unworkable as well as normatively dubious, indeed wrong.[8] A sweeping international ban treaty proposed by some advocacy groups falls into that category.

The United States and its partners have grave interests—legal, moral, and strategic—in developing, simultaneously with new automated and autonomous weapons, a broadly shared normative framework and expectations for how these systems must perform to be lawful. They have an interest in discussions and exchanges of views and practices with others in the world who are developing and deploying these weapons. Those interests make it imperative, though, that the United States and its partners understand that shared norms only come about with some shared information about each party's own view of principles, policies, and practices regarding these weapons. The United States particularly must therefore resist its own natural impulses toward secrecy and reticence with respect to military technologies, at least where feasible. U.S. interests in technological and military secrecy must be balanced here against interests in shaping the normative terrain—the contours of how international law should be understood, interpreted, and applied to these new weapons, as well as informal international expectations about appropriate technological design—on which it and others will operate militarily as automation evolves.

Just as development of autonomous weapon systems will be incremental, so too will development of norms about acceptable systems and uses. The United States and its partners must act, however, before international expectations about these technologies harden around either of two extreme alternatives: imposing unrealistic, ineffective or dangerous bans based on sci-fi scenarios of killer robots rather than realistic understandings of the new technologies and their uses, or proceeding with few or no constraints at all, which might well result in

the development, deployment, and sale of "autonomous," but also patently illegal, weapons.[9]

For its part, the United States should assert that existing international law of war requirements and so-called Article 36 weapons reviews (based on Article 36 of the first additional protocol to the Geneva Conventions), must be applied by all parties in the development and deployment of automated weapon systems, with special scrutiny to systems that are autonomous with respect to target selection and engagement. At the same time, it should assert equally that these systems raise novel issues of legal review, and that while the United States has internal processes and standards seeking to give content to such reviews, it also understands these as ongoing attempts to develop shared frameworks of best practices and norms. It should propose and welcome discussion, comment, and the shared experience of other states, particularly those that are also actively developing new weapon systems.

National-level processes like these should be combined with international dialogue aimed at developing common ethical standards and legal interpretations. Such efforts will help develop a gradually coalescing set of shared practices and expectations that can be shaped over time as the technologies emerge and evolve.

Drones: Past Lessons and Future Automation

The incremental march toward automated lethal technologies of the future, and the legal and ethical challenges that accompany it, can be illustrated by looking at today's unmanned aerial vehicles (UAVs).[10] Unmanned aircraft piloted from afar are already a significant component of the U.S. arsenal. At this writing, close to one in three U.S. Air Force aircraft is remotely piloted (though this number also includes many tiny tactical surveillance drones) and the unmanned aircraft proportion will only grow.[11] Many other states are developing or importing such technology.[12] Current unmanned military aircraft are not autonomous in the firing of weapons—the weapon must be fired in real-time by a human controller—and so far there are no known plans or, apparently in the view of U.S. military planners, reasons today to take the human out of the weapon firing loop.[13]

Nor are today's UAVs truly autonomous as aircraft—they require human pilots in real-time to fly them, even when they are located far away. They *are*, however, increasingly automated in their flight functions—self-landing capabilities, for example, and particularly automation to the point that a single pilot can operate several unmanned aircraft at once, increasing efficiency considerably. The automation of flight is gradually increasing as sensors and aircraft control through computer programming improves. Looking into the future, some observers believe that one of the next generations of jet fighter aircraft will

no longer be manned or, at least, that manned fighter aircraft will be joined by unmanned aircraft.[14] Given that speed in every sense—including turning and twisting in flight, and reaction and decision times—is an advantage, design will emphasize automating as many of these functions as possible, in competition with the enemy's systems.[15]

Just as the aircraft might have to be maneuvered far too quickly for detailed human control of its movements, so too the weapons—against other aircraft, UAVs, anti-aircraft systems—might have to be utilized at the same speeds in order to match the beyond-human speed of the aircraft's own systems (as well as the enemy aircraft's similarly automated counter-systems).[16] Likewise, as alluded to above, defense systems on modern U.S. naval vessels have long been able to target incoming missiles automatically, with humans monitoring the systems' operation, because human decision-making processes are too slow to deal with multiple, inbound, high-speed missiles.[17] Some military operators regard many emerging automated weapons as a more sophisticated form of "fire and forget" self-guided missiles. And because contemporary fighter aircraft are designed not only for air-to-air combat but for ground attack missions as well, the design changes that reduce the role of the human controller of the aircraft platform may shade into automation of the weapons directed at ground targets, too.

Although current remotely-piloted UAVs, on the one hand, and future autonomous weapons, on the other, are based on different technologies and operational imperatives, they generate some similar concerns about their ethical legitimacy and lawfulness. Today's arguments over the legality of remotely-piloted, unmanned aircraft in their various missions (especially regarding targeted killing operations, and concerns that the United States is using technology to shift risk off its own personnel and onto remote-area civilian populations) presage the arguments that already loom over weapons systems that exhibit emerging features of autonomy—arguments, for example, about supposed limits of targeting precision, erosion of professional military ethics, and reduced thresholds for employing military force. Those arguments also offer lessons to guide short- and long-term policy toward autonomous weapons generally, including systems that are otherwise quite different.

Incremental Automation

These issues are perhaps easiest to imagine in the airpower context. But in other battlefield contexts as well, the United States and other sophisticated military powers will find increasingly automated lethal systems ever more attractive. So too, eventually, will unsophisticated powers and possibly non-state actors, as such technologies become commodified and offered for licit or illicit sale.

Moreover, as artificial intelligence improves, weapon systems will evolve from robotic "automation"—the execution of precisely pre-programmed actions

or sequences in a well-defined and controlled environment—toward genuine "autonomy," meaning the robot is capable of generating actions to adapt to changing and unpredictable environments.[18] This evolution will be propelled by a combination of several factors, including that automated machines may perform some functions better from a military perspective (in some cases faster, stealthier, or more precisely) than human equivalents, that they reduce risks to human soldiers, and that industry will be producing technologies for a range of applications that can adapted to military functions.

The fact that these technologies of weapon automation will spread over time, and that they will have good and bad uses, preoccupies many commentators, some imagining that had the United States not introduced such technologies, no one else would have either. It is widely believed that the United States will come to regret having triggered an "arms race" in UAV weapon systems or autonomous weapons.[19] The recent experience of U.S. lethal targeting policy and UAVs informs the recommendations we offer below. However, the United States has never been alone in pursuing these capabilities, the technologies of which are a function of advancing automation in many aspects of life, and it is a dangerous mistake to believe that it could have kept such capabilities in the box by foregoing them.

Initially, many autonomous weapon systems will be designed for use in operational environments in which two conditions make their use less legally or ethically problematic than usually imagined for these systems. First, some systems will be designed to be used only for "machine-on-machine" encounters, such as missile defense. This does not mean that there are no issues of targeting or collateral damage, of course, but in practical terms the concerns are not the same as, for example, close-in infantry urban warfare. Second, some systems will be designed only for use in operational environments in which there are few if any civilians present—an attack against an undersea submarine, for example. None of these is free from the possibility of target misidentification or selection errors, or the danger of civilian harm, but certainly not all autonomous weapons are designed for use in the most difficult environments for machine automation. It is a mistake, when imagining legal or ethical issues of autonomous weapons, to start with the most difficult operational environment, for which a lawful fully autonomous weapon would be the hardest (if even possible) to design.

The naval ship-borne, automated anti-missile systems described above are usually used in environments with few civilians present. Land-based, automated anti-missile systems are also machines designed to destroy other machines. But to take more difficult engineering and regulatory situations, consider efforts to protect peacekeepers facing the threat of snipers or ambush in an urban environment, or infantry teams working to secure a town. Small mobile robots with weapons could act as roving scouts for the human soldiers, with "intermediate" automation—the robot might be pre-programmed to look for

certain enemy weapon signatures and to bring the threat to the attention of a human operator, who then decides whether or not to pull the trigger. Automation might be used here not only to help keep soldiers safe but because it might help them distinguish hostile threats from innocent civilians, especially in situations where the former deliberately hides among the latter.

In the next iteration, the system might be set with the human being not required to give an affirmative command, but instead merely deciding whether to override and veto a machine-initiated attack. Perhaps next the system will be designed to target and fire autonomously but to wait and call for higher-level authorization only when it assesses possible collateral damage above a certain level. In any of these permutations that human decision-maker also might not be a soldier on site, but an off-battlefield, remote robot-controller.

It is already clear that the communications link between human and weapon system could be jammed or hacked. One technological response might be to reduce the vulnerability of the communications link by severing it—making the robot dependent upon executing its own programming, or even rendering it genuinely autonomous. Moreover, as the speed and complexity of response in weapon platform maneuvers increases, the communications link might be simply too slow to control the craft, and greater automation, even autonomy, in at least some functions might prove an important step.

Covert or special operations will involve their own evolution toward incrementally autonomous systems (especially as enemy forces use civilians as cover). Consider the raid on the Osama bin Laden compound—tiny surveillance robots equipped with facial and voice recognition technology might have helped to affirmatively identify bin Laden earlier and to distinguish enemy fighters from innocent bystanders. It might not be a large step to weaponize such systems, and then perhaps go the next step to allow them to act autonomously—perhaps initially with a human remote-observer as a failsafe, but with very little time to override programmed commands.

It is important to note here that no one seriously expects remotely-controlled or autonomous systems to completely replace humans on the battlefield. Many military missions will always require humans on the ground, even if in some contexts they will operate alongside and in conjunction with increasingly automated, sometimes autonomous, systems.

The examples in the previous paragraphs have all been stylized to sound precise and cautiously controlled, carefully attentive to the legal requirements of weapons review. Consider also, however, that at some point in the not-distant future, someone—maybe in China, maybe in Russia, or someplace else—will likely design, build, deploy and sell for use in an urban, civilian-filled battlefield

an autonomous weapon system that is programmed to target only something—say a person or position—that is firing a weapon and is probably hostile rather than friendly. Such a weapon system might lack the ability altogether to take account of civilian presence and likely collateral damage.

Quite apart from the security and war-fighting implications, the U.S. and allied governments would have grave legal and humanitarian concerns about such a system offered for sale on the international arms markets, let alone deployed and used. They would then find themselves potentially facing a weapon system on the battlefield that conveys significant advantages to its user, but which they would not deploy themselves because (for reasons described below) they would not consider it a legal weapon. Such systems are much easier to create than lawful ones. The speed necessary to respond to such adversary systems in the field, though, might well create demand for defensive systems that feature greater autonomy in decision-making.

The implication is that evolution in battlefield robots will be more than simply a race for ever more sophisticated automated weapon systems. It will also feature the imperative to counter automated systems that are relatively easy to design and build, but not actually lawful. Defenses against the latter will come partly through technical means, but partly, as elaborated in the sections that follow, through international norms and diplomacy.

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.

International Treaties and the Challenge of Incremental Evolution

These four objections run to the whole enterprise of building autonomous weapon systems, and important debates could be held around each of them. Each has serious weaknesses, but whatever their merits in theory, they all face a practical difficulty in the incremental way autonomous weapon systems will develop out of gradually increasing automation of multiple and often discrete subsystems in a weapon system. The four grand objections are often voiced, after all, as though there was likely to be some determinable break-point between the fully human-controlled system and the fully machine-controlled one. It is unlikely to happen that way. Even if, in principle, a "fully autonomous weapon

system" is one in which the weapon system selects the target and engages it, with no human being "in" or even "on the loop" (to effect a human override), in practical terms, it is far more likely that the evolution of weapons technology will be gradual, slowly and indistinctly eroding the role of the human in both target identification and firing at the target. It will not be so clear when automation of the system has advanced so far that, for the purposes of the objections voiced by critics, the machine operates according to its own programming.

"Incrementality"—moving gradually from automation to genuine autonomy—does not by itself render any of the four universal objections wrong per se. But it does mean that another kind of discussion should be had about regulation of weapons systems undergoing step-by-step change. It is much less conceptual and intellectual than the four arguments above—but more consequential. A fully autonomous weapon system is easy to define in the abstract: in the Department of Defense directive's terminology, it is a weapon system in which the machine both selects the target and engages it without human intervention. But applying this definition will probably encounter trouble in many particular cases, mostly in determining whether the human operator has a sufficiently robust role to say that the system is not autonomous. Such assessment addresses only the question of whether the system exhibits full autonomy, in any event, not the further question of whether its automated capabilities are legally sufficient for its operational battlefield environment.

Instead, what engineers and designers do on a daily basis in developing these systems gives rise to the regulatory conversation that most matters. What matters in practical terms is the granular discussion of specific and particular programming, incremental changes to any and all subsystems—interwoven at each step of design and development with the normative requirements of weapons law.

As mentioned earlier, the United States is sometimes portrayed as engaged in heedless pursuit of technological advantage that will inevitably be fleeting as other countries mimic, steal, or reverse-engineer its technologies.[41] According to this view, if the United States would quit pursuing these technologies, the genie might remain in the bottle or at least emerge much more slowly. This is exaggerated or wrong, though, in part because the automation technologies at issue are being developed in other states as well and are already spreading with respect to general use far outside of military applications.[42]

It is also true, in the area of weapons development, that it is easier, faster, and cheaper for states competitively engaged with the United States to deploy systems that would be, in the U.S. view or those of its partners and close allies, ethically and legally deficient. Autonomous and unlawful is easy; autonomous and lawful, for the most difficult operational environments, is hard.

For this reason among many others, the United States and its partners *do* have strong interests in seeing that development and deployment of both highly automated and autonomous battlefield robots be reviewed and regulated in some fashion to ensure that those developed and deployed are lawful. Moreover, it is quite true (as some critics of U.S. weapons-development policy have pointed out) that, even if U.S. abstention from developing weapons based in these new technologies of automation and autonomy alone would not prevent their proliferation, it would be reckless nonetheless for the United States along with its partners to pursue them without a well-conceived and shared policy strategy—including a role for normative constraints—for responding to other states' or actors' design, development, deployment, and use of them.

These considerations—and alarm at the apparent development of an arms race around these emerging and future weapons—have led many to believe that an important part of the solution lies in some form of multilateral treaty.[43] Some have proposed that a treaty might be a regulatory one, restricting acceptable weapons systems or restricting acceptable use, following the pattern of regulatory weapon treaties of the past—the 1868 St. Petersburg Declaration (banning some munitions below 400 grams weight as causing unnecessary suffering), for example. An international regulatory protocol might provide guidance as to the level of automation presumptively permissible—or impermissible—for certain categories of operations, such as machine-on-machine in naval warfare (with few if any civilians present), for example. Regulatory treaties for weapons face difficult conceptual drafting questions such as whether there are sufficiently specified technological design and operational conditions (but also reasonably stable as technologies change over time) that a protocol can offer legal guidance that will not be quickly outdated and made obsolete. Those who believe that an international regulatory agreement is immediately the best approach have to consider carefully whether and to what extent a new weapons protocol and its process of diplomatic negotiation, in this new and rapidly shifting technological area, can or should take the lead in seeking to formulate new binding law. Too quick a rush to a binding protocol risks instability of norms and rapid obsolescence through technological change, or vacuity through the inability to formulate the kinds of specific and yet not quickly outdated provisions worth enshrining in a binding treaty. It would be more prudent, in seeking a stable basis for law over the long run, to let other, less formal processes take the lead to allow genuinely widely shared norms to coalesce in a very difficult area.

Be that as it may, the limelight of public and media attention has been taken today, not by any *regulatory* treaty proposal, but instead by calls for a *prohibitory* treaty. The model of this kind of prohibitory weapons treaty is the 1997 Ottawa Convention banning antipersonnel landmines.[44] A coalition of advocacy groups called the International Committee for Robot Arms Control has been working

over the last few years to promote an international convention to prohibit "[f]urther development, acquisition, deployment, and use of armed autonomous robot weapons."[45] The call for a prohibitory international ban was raised to far greater prominence recently when, in November 2012, Human Rights Watch issued a report calling for a sweeping multilateral treaty that would ban outright the development, production, sale, deployment, or use of "fully autonomous weapons" programmed to select and engage targets without human intervention.[46] The report, "Losing Humanity: The Case Against Killer Robots," was controversial in several of its key assertions. It did not limit itself, for example, to saying that a ban treaty was necessary to create new law outlawing a category of new weapons; it said, rather, that its "initial evaluation" of fully autonomous weapons showed that they "would appear to be incapable of abiding by key principles of international humanitarian law," thus suggesting illegality under existing law.[47] Moreover, it did not limit itself to calling for a ban on fully autonomous weapons. It went much further to call for a ban on "development" of any fully autonomous weapon system.

In any case, ambitions for a multilateral treaty regulating or prohibiting autonomous weapon systems are misguided for several reasons. For starters, limitations on autonomous military technologies, although quite likely to find wide superficial acceptance among some states and some non-governmental groups and actors, will have little traction among those most likely to develop and use them. Some states may want the United States to be more aggressive in adopting the latest technologies, given that possible adversaries are likely to have far fewer compunctions about their own autonomous weapon systems, and others are likely to favor any technological development that extends the reach and impact of U.S. and allied forces or enhances their own ability to counter adversaries' capabilities.

Even states and groups inclined to support treaty prohibitions or limitations will find it difficult to reach agreement on scope or definitions because lethal autonomy will be introduced incrementally—as battlefield machines become smarter and faster, and the real-time human role in controlling them gradually recedes, agreeing on what constitutes a prohibited autonomous weapon will be unattainable. Even assuming agreement could be reached, there are the general challenges of compliance: the collective action problems of failure and defection that afflict all such treaty regimes, especially when dealing with dual-use (civilian and military) underlying technologies.

Finally, there are serious humanitarian risks to prohibition, given the possibility that autonomous weapons systems could in the long run be more discriminating and ethically preferable to alternatives. If all such systems are prohibited, and particularly if even research and development of relevant technologies is also prohibited, one never gets the benefits that might come from new technologies—

and future generations will not even be aware of the potential benefits that were given up, because these prohibitions on development meant they were never even pursued. Prohibition precludes the possibility of such benefits, and proponents of it must acknowledge and bear responsibility for this risk.

Principles, Policies, and Processes for Regulating Automating Weapon Systems

The risks and dangers associated with advancing autonomous robotic weapons are very real. Of course the grave dangers include the possible abuse of this technology by parties that do not respect existing international legal requirements—especially those that flagrantly cast aside the principles of distinction and proportionality. They also include less nefarious but nevertheless significant risks, such as unintended and unpredicted interactions between automated or autonomous systems; mistaken belief that a human operator will be able to meaningfully remain "on the loop" in real-time (whereas in practice the human operator may not have sufficient time or alternative sources of information to make any meaningful counter-decision); or even a decision algorithm for which the programmers did not sufficiently understand, incorporate, or test the requirements of the law of armed conflict for a particular situation that arises. The authors of the Defense Department directive on autonomous weapons were well aware that whether a system is merely highly automated or genuinely autonomous might well depend less on the machine's design than on the anticipated role for the human operators. If they cannot reasonably perform that role (perhaps because it is effectively beyond human capabilities or because those doing the operating have not been sufficiently trained), a system believed to be merely automated to a limited point might turn out to be effectively autonomous.

Given all of these risks, and to promote many other interests as well, the United States has a serious interest in promoting legal and ethical norms to guide the application of fundamental principles of the laws of war to the development and evaluation of emerging automated and autonomous weapons, and in guiding development in this context of international norms. By "international norms" here, we do not mean new binding legal rules only—whether treaty rules or customary international law—but instead the gradual fostering of widely-held expectations about legally or ethically appropriate conduct, whether formally binding or not. Among other reasons, such norms are important to the United States for guiding and constraining its internal practices, such as research and development (R&D) and eventual deployment of autonomous lethal systems that it regards as legal; they help earn and sustain necessary buy-in from the officers and lawyers who would actually use or authorize such systems in the field.[48] They help establish common standards among the United States and its partners and allies to promote cooperation and joint operations. And they raise the

political and diplomatic costs to adversaries of developing, selling, or using autonomous lethal systems that run afoul of these standards.

A better approach than treaties for addressing these systems is the gradual development of internal state norms and best practices that, worked out, debated, and applied to the United States' own weapons-development process, can be carried outwards to discussions with others around the world. National-level processes should be combined with international dialogue aimed at developing common standards and legal interpretations. This requires a long-term, sustained effort combining internal ethical and legal scrutiny and external diplomacy and collaboration.

For its part, the government of the United States must resist two extreme instincts—its own instincts among officials to hunker down behind secrecy and avoid discussing and defending even guiding principles, on the one hand, and the instincts of critics or skeptics of autonomous lethal systems favoring the idea of some grand international treaty to regulate or even prohibit them, on the other. Instead the U.S. government should carefully and continually develop internal principles, processes, and practices that it believes are correct for the design and implementation of such systems. It should also prepare to articulate clearly to the world the fundamental legal and moral standards by which all parties ought to judge autonomous weapons, whether those of the United States or those of others.

The core, baseline standards can and should be drawn and adapted from the customary law of armed conflict framework: the principles distinction and proportionality. Even if we are a long way from having, or possibly ever wanting to have for actual deployment, an autonomous weapon that would be able to make the required judgments of distinction and proportionality as probabilistic judgments, we should undertake the research into it, and ensure that those doing the engineering and design research understand just how difficult the concepts are for human beings. Proportionality, for example, is partly a technical issue of designing systems capable of measuring predicted civilian harm, but also partly an ethical issue of attaching weights to the variables at stake. It also is important to bear in mind the point made earlier that distinction and proportionality assessments are probabilistic, so those programming these systems or those deciding how to set and use them must decide on minimum confidence levels, which might depend on specific contexts and missions. Some of the most important discussions among engineers, lawyers, commanders, and policy-makers will therefore be about where to set those parameters and value settings (forms of which are already done routinely in formulating military rules of engagement and conducting collateral damage assessments for bombing or targeting operations).[49] These will be legally, ethically, and pragmatically difficult discussions, and they should be.

Such questions move from overarching legal, ethical, and policy principles to processes that make sure principles are concretely taken into account—not just down the road at the deployment stage but much earlier, during the R&D stage. It will not work to go forward with design and only afterwards, seeing the technology, decide what changes need to be made in order to make the system's decision-making conform to legal requirements. By then it may be too late. Engineering designs will have been set for both hardware and software and significant national investment into R&D already undertaken will be hard to write off on ethical or legal grounds. Legal, ethical, and policy review by that stage could become a matter of justification at the back end, rather than seeking best practices at the front end. None of this urges a ban, as some call for, but it does urge careful attention to normative as well as design concerns at each step along the way.

The United States must develop a set of principles to regulate and govern advanced autonomous weapons not just to guide its own systems, but also to effectively assess the systems of other states. This requires that the United States work to bring along its partners and allies—most notably NATO members and technologically advanced Asian allies—by developing common understandings of norms and best practices as the technology evolves. Just as development of autonomous weapon systems will be incremental, so too will development of norms about acceptable systems and uses.

Internal processes should therefore be combined with public articulation of overarching policies. Various mechanisms for declaring policy might be utilized over time—perhaps directives by the Secretary of Defense (like the November 2012 directive already promulgated by the Deputy Secretary)—followed by periodic statements explaining the legal rationale behind decisions about R&D and deployment of weapon technologies. The United States has taken a similar approach in the recent past to other controversial technologies, most notably cluster munitions and landmines, by declaring commitment to specific standards that balance operational necessities with humanitarian imperatives.[50] These policy pronouncements establish parameters internal to the U.S. government and they serve as vehicles for explaining reasoning to outside audiences at home and abroad; they can also be adapted by other states through consultative processes.

Such a national-level process should be combined with international dialogues aimed at fostering agreement on—or at least narrowing the differences with respect to—efforts to adapt and translate the law of armed conflict to this technological context. As a possible model, an international grouping of legal experts commissioned by the NATO Cooperative Cyber Defence Centre of Excellence has been working for the past few years in another technologically transformative area of conflict: cyber warfare. This process is meant to develop

and propose interpretive guidance (including the Tallinn Manual on the International Law Applicable to Cyber Warfare[51]) for states' and other actors' consideration. Although the cyber context is different, insofar as there may be greater disagreement as to the appropriate legal framework, similar international processes—whether involving state representatives, or independent experts, or both—can help foster broad consensus or surface disagreements that require resolution with respect to autonomous weapon systems.

To be sure, this proposal risks papering over enormous practical and policy difficulties. The natural instinct of the U.S. defense community—likewise that of other major state powers—will be to discuss little or nothing, for fear of revealing capabilities or programming details to adversaries, or enabling industrial espionage and reverse-engineering of systems. Policy declarations will necessarily be more general and less factually specific than critics would like. Furthermore, one might reasonably question whether broad principles like distinction and proportionality can meaningfully be applied and discussed publicly with respect to high-tech systems distinguishable only in terms of digital ones and zeroes buried deep in programmed computer code.

These concerns are real, and they demand two mitigating answers. First, the United States will need to resist its own impulses toward secrecy and reticence with respect to military technologies, recognizing that the interests those tendencies serve are counterbalanced here by interests in shaping the normative terrain on which it and others will operate militarily as technology quickly evolves. The legitimacy of such inevitably controversial systems in the public and international view matters greatly. It is better that the United States work to set the global standard by actively explaining its compliance with it than to let it be set by other states or groups—whether those who would impose ineffective or counterproductive prohibitions or those who would prefer few or no constraints at all.

Of course, there are limits to transparency here, on account of both secrecy concerns and the practical limits of persuading skeptical audiences about the internal and undisclosed decision-making capacities of rapidly evolving weapon systems. A second part of the answer is therefore to emphasize the internal processes by which the United States considers, develops, and tests its weapon systems. Legal review of any new weapon system is required as a matter of international law; as explained above, consistent with Article 36, the U.S. military would conduct it in any event to ensure the basic lawfulness of new weapon technologies. Even when the United States cannot disclose publicly the details of its automated systems and their internal programming, however, it should be as open as it can about its vetting procedures, both at the R&D stage and at the deployment stage, including the standards and metrics it uses in its evaluations. That is, the United States should take the lead in emphasizing publicly the legal

principles it applies and the policies and processes it establishes to ensure compliance, encouraging others to do likewise.

Although the United States cannot be very open publicly with the results of its tests, for fear of disclosing details of its capabilities to adversaries, it should at least be prepared to share them with its military allies as part of an effort to establish common standards. It should take the lead in inviting both high-level state-to-state discussions of these informal best practices, as well as discussions among weapons designers, engineers, lawyers, and others.

NGOs have an important role to play here in promoting transparency and best practices. They should press states to ensure that the standards and processes of review take all relevant legal, ethical, strategic, and engineering factors into account, and they should also press states to be open about the specific content of those standards and processes, including how they conduct legal reviews in fielding new weapons. The International Committee of the Red Cross's traditional role in fostering dialogue about proper standards and processes is important in this area, as it is in others.

Looking more speculatively ahead, the standards the United States applies internally in developing its systems might eventually form the basis of export control standards, in sharing technologies with allies and other states. As other countries develop their own autonomous weapons systems, and refine their policies and processes for regulating them, the United States can lead in forging a common export control regime and standards of acceptable weapons available on international markets.

In the end, one might still raise an objection from an entirely different direction to these proposals: that the United States (or any rational state in a similar position) should not unnecessarily constrain itself in advance under a set of normative commitments, given vast uncertainties about future technology, threats, and the broader security environment. Better, the argument might go, that the United States cautiously wait and avoid binding itself to one or another legal interpretation or policy commitment until it needs to do so. This objection fails to appreciate, however, that although significant deployment of highly autonomous systems may be far off, R&D decisions are already upon us. Development of highly automated systems today is already intertwined with development of future autonomous systems. Moreover, shaping international norms is a long-term process, and unless the United States and its allies accept some risk in starting it now, they may lose the opportunity to do so later.

Conclusion

The incremental development and deployment of autonomous weapon systems is inevitable, and any attempt at a global ban will be ineffective in stopping their use by the states whose acquisition of such weaponry would be most dangerous. Autonomous weapon systems are not inherently unlawful or unethical. Existing legal norms are sufficiently robust to enable us to address the new challenges raised by robotic systems. The best way to adapt existing norms to deal with these new technologies is a combined and international-national dialogue designed to foster common standards and spread best practices.

Taken as a whole, these policy proposals reflect a rather traditional approach—relying on the gradual evolution and adaptation of long-standing law of armed conflict principles—to regulate what seems to many like a revolutionary technological and ethical predicament. That is in part because the challenge of regulating apparently radical innovations in weaponry within a long-standing legal and ethical framework is hardly novel.

Some view the emergence of automated and autonomous weapon systems as a crisis for the law and ethics of war. To the contrary, provided we start now to incorporate legal and ethical norms adapted to weapons that incorporate emerging technologies of automation, the incremental movement from automation to machine autonomy can be both regulated and made to serve the ends of law on the battlefield.

Notes

1 This paper expands on ideas first published in Kenneth Anderson and Matthew Waxman, "Law and Ethics for Robot Soldiers," *Policy Review,* December 2012, http://www.hoover.org/publications/policy-review/article/135336.

2 See Bill Keller, "Smart Drones," *New York Times,* March 17, 2013.

3 Department of Defense Directive 3000.09, Autonomy in Weapon Systems 13 (November 21, 2012).

4 U.S. Navy, *MK-15 Phalanx Close-In Weapons System* (CIWS), http://www.navy.mil/navydata/fact_display.asp?cid=2100&tid=487&ct=2; Federation of American Scientists, *MK 15 Phalanx Close-In Weapons System (CIWS)*, January 9, 2003, http://www.fas.org/man/dod-101/sys/ship/weaps/mk-15.htm; Michael N. Schmitt and Jeffrey S. Thurnher, *Out of the Loop: Autonomous Weapon Systems and the Law of Armed Conflict,* Harvard National Security Journal (forthcoming 2013), working draft at SSRN, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2212188.

5 "South Korea deploys robot capable of killing intruders along border with North," *The Telegraph,* July 13, 2010, http://www.telegraph.co.uk/news/worldnews/asia/southkorea/7887217/South-Korea-deploys-robot-capable-of-killing-intruders-along-border-with-North.html.

6 The starting point for these and many other automated and autonomous battlefield weapons scenarios is Peter Singer, *Wired for War: The Robotics Revolution and Conflict in the 21st Century* (Penguin Press, 2009).

For a dissenting view, that removing humans from military targeting is unlikely for the foreseeable future, see Werner J.A. Dahm, "Killer Drones Are Science Fiction," *Wall Street Journal,* February 15, 2012.

7 This is approximately the view of a leading U.S. Air Force strategy document on the future evolution and use of drones—"remotely-piloted vehicles" or "unmanned aerial vehicles" (UAVs) in more technical language. See "United States Air Force Unmanned Aircraft Systems Flight Plan, 2009–2047" (May 18, 2009), at 16, http://www.fas.org/irp/program/collect/uas_2009.pdf.

8 For a survey of possible modes of international governance of lethal autonomous robotics, see Gary E. Marchant, et al., "International Governance of Autonomous Military Robotics," *Columbia Science and Technology Law Review* 12 (2011), at 272, http://www.stlr.org/cite.cgi?volume=12&article=7.

9 Markus Wagner, "Taking Humans Out of the Loop: Implications for International Humanitarian Law," *Journal of Law, Information and Science* 21 (2011).

10 See, generally, Shane Harris, "'Out of the Loop: The Human-free Future of Unmanned Aerial Vehicles," in *Emerging Threats in National Security and Law* essay series (Hoover Institution, 2012), http://media.hoover.org/sites/default/files/documents/EmergingThreats_Harris.pdf.

11 Spencer Ackerman and Noah Shachtman, "Almost 1 in 3 U.S. Warplanes Is a Robot," *Wired* Danger Room, January 9, 2012, http://www.wired.com/dangerroom/2012/01/drone-report.

12 Micah Zenko, "Reforming U.S. Drone Strike Policies," Council on Foreign Relations (January 2013), at 18–20, http://www.cfr.org/wars-and-warfare/reforming-us-drone-strike-policies/p29736.

13 We are not including here, for example, weapons that can be pre-programmed toward specific geographic coordinates, such as a cruise missile or its updated UAV cousins.

14 See "Unmanned Aerial Warfare: Flight of the Drones: Why the future of air power belongs to unmanned systems," *The Economist,* October 8, 2011.

15 This amounts to the famous "OODA Loop" by any other name. OODA refers to the techno-strategic concept first developed by fighter pilot and strategist John Boyd—"observe, orient, decide, and act." See, generally, Frans P.B. Osinga, *Science, Strategy and War: The Strategic Theory of John Boyd* (Routledge, 2006).

16 This is, then, another expression of the OODA Loop, because, in air-to-air combat, "Boyd's insight was that . . . the advantage lay with the fighter pilot whose OODA Loop was faster and more accurate than his opponent's, and who was able to throw his opponent's loop out of sync." William C. Marra and Sonia K. McNeil, "Automation and Autonomy in Advanced Machines: Understanding and Regulating Complex Systems," Lawfare Research Paper Series 1-2012 (April 2012), at 9. The fastest OODA Loop of the future combat plane is likely to be an automated one—automated in both flight and weapons functions, and unmanned as well.

17 Singer, *Wired for War,* 124.

18 Marra and McNeil, "Automation and Autonomy in Advanced Machines," at 15–28.

19 Kristin Roberts, "When the Whole World Has Drones," *National Journal,* March 21, 2013, http://www.nationaljournal.com/magazine/when-the-whole-world-has-drones-20130321.

20 U.K. Ministry of Defense, "Joint Doctrine Note 2/11: The U.K. Approach to Unmanned Aircraft Systems, Developments, Concepts and Doctrine," March 30, 2011 (MOD 2011 Report), at para. 508.

21 See, e.g., Department of Defense Directive, Enclosure 2 at (a), providing that systems will go through rigorous testing including "analysis of unanticipated emergent behavior resulting from the effects of complex operational environments on autonomous or semi-autonomous systems."

22 For a recent example of the U.S. Air Force directive on procedures to be followed for review of the legality of weapon systems and cyber capabilities, see Order of the Secretary of the Air Force, "Legal Reviews of Weapons and Cyber Capabilities," Air Force Instructions 51-402, July 27, 2011, http://www.fas.org/irp/doddir/usaf/afi51-4w02.pdf.

23 Personal interviews with authors.

24 For an excellent summary of these issues, see Jeffrey S. Thurnher, "The Law That Applies to Autonomous Weapon Systems," ASIL Insights 17, no. 4, January 18, 2013. (Lt. Col. Thurnher is a U.S. Army JAG officer and professor in the international law department of the U.S. Naval War College.) See also Schmitt and Thurnher, "Out of the Loop"; Michael N. Schmitt, "Autonomous Weapon Systems and International Humanitarian Law: A Reply to the Critics," *Harvard National Security Journal* (2013); Alan Backstrom and Ian Henderson, "New Capabilities in Warfare: An Overview of Contemporary Technological Developments and the Associated Legal and Engineering Issues in Article 36 Weapons Reviews," *International Review of the Red Cross* (forthcoming 2013).

25 Thurnher, ASIL Insights.

26 Thurnher, ASIL Insights.

27 The customary formulation as found, for example, in 1977 Additional Protocol I, Art. 51(5)(b).

28 Perhaps the most ambitious of these efforts is by roboticist Ronald C. Arkin, who describes his work on both distinction and proportionality in his "Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture," Technical Report GIT-GVU-07-11, Georgia Tech Robotics Lab (2007), at 43–53, http://www.cc.gatech.edu/ai/robot-lab/online-publications/formalizationv35.pdf. For a general discussion, see also Ronald C. Arkin, *Governing Lethal Behavior in Autonomous Robots* (Chapman and Hall, 2009).

29 See, e.g., Wendell Wallach and Colin Allen, *Moral Machines: Teaching Robots Right from Wrong* (Oxford University Press, 2009); Armin Krishnan, *Killer Robots: Legality and Ethicality of Autonomous Weapons* (Ashgate, 2009); Markus Wagner, "Autonomy in the Battlespace: Independently Operating Weapon Systems and the Law of Armed Conflict," in *International Humanitarian Law and the Changing Technology of War* (Martinus Nijhoff, 2013), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2211036; Peter M. Asaro, "Modeling the Moral User," *IEEE Technology and Society Magazine* (Spring 2009), at 20–24 (Asaro directly addresses Arkin's modeling assumptions about moral machines).

30 See *Economist,* "Flight of the Drones," which notes that some "question whether artificial intelligence, which always seems just a few years away, will ever work well enough . . . [a]nd decisions about what is proportionate often require fine distinctions and sophisticated judgment . . . there will be an almost infinite combination of contingencies facing drones."

31 See, e.g., Noel Sharkey, "March of the Killer Robots," *London Telegraph,* June 15, 2009; Noel Sharkey, "The Automation and Proliferation of Military Drones and the Protection of Civilians," *Law, Innovation and Technology* 3 (December 2011), at 236–37; Robert Sparrow, "Building a Better WarBot: Ethical Issues in the Design of Unmanned Systems for Military Applications," *Science and Engineering Ethics* 15 (June 2009), at 178.

32 See, e.g., Mathias Gutmann, Benjamin Rathgeber, and Tareq Syed, "Action and Autonomy: A Hidden Dilemma in Artificial Autonomous Systems," in Michael Decker and Mathias Gutmann, eds., *Robo- and Informationethics: Some Fundamentals* (LIT Verlag Munster, 2012), at 233.

33 A direct moral case against autonomous weapons as such is made in Matthew Bolton, Thomas Nash, and Richard Moyes, "Ban autonomous armed robots," Article36.org, March 5, 2012, http://www.article36.org/statements/ban-autonomous-armed-robots.

34 See, e.g., Wendell Wallach, "Terminating the Terminator: What to do About Autonomous Weapons," *Science Progress,* January 29, 2013, http://scienceprogress.org/2013/01/terminating-the-terminator-what-to-do-about-autonomous-weapons.

35 See W.J. Hennigan, "New Drone Has No Pilot Anywhere, So Who's Accountable?" *Los Angeles Times,* January 26, 2012, http://articles.latimes.com/2012/jan/26/business/la-fi-auto-drone-20120126.

36 See Mary L. Cummings, "Automation and Accountability in Decision Support System Interface Design," *Journal of Technology Studies* 32 (Winter 2006), at 23–31; see also M.L. Cummings, "Creating Moral Buffers in Weapon Control Interface Design," *IEEE Technology and Society* (Fall 2004), at 28–33.

37 See Sparrow, "Building a Better WarBot," at 178–79.

38 *Economist,* "Flight of the Drones": "Looking farther ahead, there are fears that UAVs and other roboticised killing machines will so lower the political threshold for fighting that an essential element of restraint will be removed. Robert E. Lee said 'it is well that war is so terrible, otherwise we would grow too fond of it.' Drones might make leaders fonder of war." See also, MOD 2011 Report, at para. 517.

39 Singer, *Wired for War,* at 431–433; Paul W. Kahn, "The Paradox of Riskless Warfare," *Philosophy & Public Policy Quarterly* 22 (Summer 2002), at 2–8; Robert Sparrow, "Predators or Plowshares? Arms Control of Robotic Weapons," *IEEE Technology & Society Magazine* (Spring 2009), at 26. Some of these issues are considered in relation to Walzer's account of just war theory by Peter M. Asaro, "How Just Could a Robot War Be?" in Philip Brey, Adam Briggle and Katinka Waelbers, eds., *Current Issues in Computing and Philosophy* (IOS Press, 2010), http://www.cybersophe.org/writing/Asaro%20Just%20Robot%20War.pdf.

40 Kenneth Anderson, "Efficiency Jus ad bellum and in bello: Making the Use of Force Too Easy?" in Claire Finkelstein, Jens David Ohlin, Andrew Altman, eds., *Targeted Killings: Law and Morality in an Asymmetrical World* (Oxford University Press 2012), at 374–399.

41 See, e.g., Scott Shane, "Coming Soon: The New Drones Arms Race," *New York Times,* October 8, 2011.

42 Sparrow discusses this "dual-use" challenge for international regulation of military robotics. "Predators or Plowshares?" at 28.

43 See, for example, the work of the International Committee for Robot Arms Control, http//www.icrac.co.uk. Their campaign is described in Nic Fleming, "Campaign Asks for International Treaty to Limit War Robots," *New Scientist,* September 30, 2009, http://www.newscientist.com/article/dn17887-campaign-asks-for-international-treaty-to-limit-war-robots.html.

44 For example, the U.K.-based NGO "Article 36" (the name refers to the provision of 1977 Additional Protocol I requiring legal review of new weapons systems) has called for a ban on autonomous lethal weapons systems. See, e.g., Bolton, Nash, and Moyes, "Ban Autonomous Armed Robots." This article explicitly argues that autonomous lethal weapons are morally the equivalent of an indiscriminate antipersonnel landmine and should be banned on the same logic. See also Jürgen Altmann, "Preventive Arms Control for Uninhabited Military Vehicles," in R. Capurro & M. Nagenborg, eds., *Ethics and Robots* (AKA Verlag Heidelberg, 2009), at 69–82.

45 International Committee for Robot Arms Control, 2010 Berlin Statement, http://icrac.net/statements.

46 "Losing Humanity: The Case Against Killer Robots," Human Rights Watch and the International Human Rights Clinic at Harvard Law School (November 2012), http://www.hrw.org/reports/2012/11/19/losing-humanity-0.

47 Losing Humanity, at 30. It goes on to reinforce the inference that these weapons, as a category, are already illegal under existing law, adding that fully autonomous weapons "would be unable to follow the rules of distinction, proportionality, and military necessity and might contravene the Martens Clause." For a critique of this legal view, see Schmitt and Thurnher, "Out of the Loop," at 9.

48 For example, the success of the NGO campaign led by Human Rights Watch and the International Committee of the Red Cross to ban blinding laser weapons—which finally resulted in the addition to the Convention on Chemical Weapons of Protocol IV (entry into force 1998; U.S. ratification 2009)—depended in considerable part on the resistance of U.S. military officers to blinding lasers as a weapon of war. See Ann Peters, "Blinding Laser Weapons: New Limits on the Technology of Warfare," Loyola International and Comparative Law Review 18 (1996), at 733.

49 Gregory S. McNeal, "Are Targeted Killings Unlawful? A Case Study in Empirical Claims Without Empirical Evidence," in Claire Finkelstein, Jens David Ohlin and Andrew Altman, eds., *Targeted Killings: Law and Morality in an Asymmetrical World* (Oxford University Press, 2012), at 327–346.

50 See, for example, Department of Defense Press Release, Cluster Munitions Policy Released, July 9, 2008, http://www.defense.gov/releases/release.aspx?releaseid=12049; U.S. Department of State, Landmine Policy White Paper, February 27, 2004, http://www.fas.org/asmp/campaigns/landmines/FactSheet_LandminePolicyWhitePaper_2-27-04.htm.

51 Tallinn Manual on the International Law Applicable to Cyber Warfare, Michael N. Schmitt, ed. (Cambridge University Press, 2013).

Matthew C. Waxman photo by Eileen Barroso/Columbia University.

Kenneth Anderson
Kenneth Anderson is a professor of international law at Washington College of Law, American University, Washington, DC, and a member of the Hoover Institution's Task Force on National Security and Law. He is the author of Living with the UN: American Responsibilities and International Order *(2011) and specializes in international law, human rights, and international business law.*

Matthew C. Waxman
Matthew C. Waxman is a professor of law at Columbia Law School, an adjunct senior fellow at the Council on Foreign Relations, and a member of the Hoover Institution's Task Force on National Security and Law. He previously served in senior positions at the State Department, Defense Department, and National Security Council.

Jean Perkins Task Force on National Security and Law

The National Security and Law Task Force examines the rule of law, the laws of war, and American constitutional law with a view to making proposals that strike an optimal balance between individual freedom and the vigorous defense of the nation against terrorists both abroad and at home. The task force's focus is the rule of law and its role in Western civilization, as well as the roles of international law and organizations, the laws of war, and U.S. criminal law. Those goals will be accomplished by systematically studying the constellation of issues—social, economic, and political—on which striking a balance depends.

The core membership of this task force includes Kenneth Anderson, Peter Berkowitz (chair), Philip Bobbitt, Jack Goldsmith, Stephen D. Krasner, Shavit Matias, Jessica Stern, Matthew C. Waxman, Ruth Wedgwood, Benjamin Wittes, and Amy B. Zegart.

For more information about this Hoover Institution Task Force please visit us online at www.hoover.org/taskforces/national-security.

HOOVER INSTITUTION
STANFORD UNIVERSITY