Robots and respect: Assessing the case against Autonomous Weapon Systems

Professor Robert Sparrow, Department of Philosophy, Monash University.

A version of this manuscript appeared as

Sparrow, R. 2016. Robots and respect: Assessing the case against Autonomous Weapon Systems. *Ethics and International Affairs* 30(1): 93-116. doi:10.1017/S0892679415000647.

Please cite that version.

Abstract:

There is increasing speculation within military and policy circles that the future of armed conflict is likely to include extensive deployment of robots designed to identify targets and destroy them without the direct oversight of a human operator. My aim in this paper is twofold. First, I will argue that the ethical case for allowing autonomous targeting, at least in specific restricted domains, is stronger than critics have acknowledged. Second, I will attempt to uncover, explicate, and defend the intuition that even in this context there would be something ethically problematic about such targeting. I argue that an account of the non-consequentialist foundations of the principle of distinction suggests that the use of autonomous weapon systems is unethical by virtue of failing to show appropriate respect for the humanity of our enemies. However, the success of the strongest form of this argument depends upon understanding the robot itself as doing the killing. To the extent that we believe that, on the contrary, AWS are only the means whereby those who order them into action kill, the idea that the use of AWS fails to respect the humanity of our enemy will turn upon an account of what is required by respect that is essentially conventional. Thus, while the theoretical foundations of the idea that AWS are weapons that are "evil in themselves" are weaker than critics have sometimes maintained, they are nonetheless sufficient to the task of demanding a prohibition of the development and deployment of such weapons.

Keywords: robots; ethics; autonomous weapon systems; LARs; Just War Theory; UAVs.

Robots and respect: Assessing the case against Autonomous Weapon Systems

Introduction

The prospect of “killer robots” may sound like science fiction. However, the large amount of attention given to the operations of remotely piloted “drones” in recent years has also spotlighted the amount of research going on in military laboratories and universities into the technologies required to allow weaponized robots to reliably select targets and destroy them without the direct oversight of a human operator. Where the spokespeople of the armed services of major industrialized powers were once quick to distance themselves from the idea that they might deploy autonomous weapons, there is increasing speculation within military and policy circles — and within the US military in particular—that the future of armed conflict is likely to include extensive deployment of Autonomous Weapon Systems[1] (AWS).[2] The recent publication of a critical report, entitled “Losing Humanity”, by Human Rights Watch[3] and the launch of an international NGO-led campaign for an arms control treaty prohibiting autonomous weapons[4] has highlighted the extent to which this prospect is perceived by many to be a threatening one and intensified the ongoing ethical debate about it.[5]

My aim in this paper is twofold. First, I will argue that the ethical case for allowing autonomous targeting, at least in specific restricted domains, is stronger than critics have acknowledged.[6] A proper understanding of the nature and force of the argument for autonomous targeting is essential to any adequate response to it. Second, I will attempt to uncover, explicate, and defend the intuition that even in this context there would be something ethically problematic about such targeting.

Given the extent of my ambitions, the dialectic that follows is somewhat complicated and for this reason it will be useful to briefly sketch an outline the argument of the paper here.

In Section 1 of the paper, I introduce a working definition of “autonomous” weapons (§ 1.1) and describe the military dynamics driving the development of these systems (§ 1.2).

Section 2 of the paper surveys and evaluates the existing literature on the ethics of AWS. The bulk of this discussion is framed as an account of two “rounds” of debate between influential advocate for AWS, Ron Arkin, and his critics. In the first round (§ 2.1) I offer an initial treatment of the debate about whether AWS will be capable of discriminating between legitimate and illegitimate targets in accordance with the Just War doctrine of *jus in bello* (§ 2.1.1), discussing both the challenges of machine perception (§ 2.1.1.1) and the role of context and reasoning in applying the principles of *jus in bello* (§ 2.1.1.2), before considering and criticizing (§ 2.1.2) the idea that embracing the design of “ethical robots” (§ 2.1.2.1) or resorting to human oversight of AWS (§ 2.1.2.2) would avoid the problems associated with the limited capacity of robots to discriminate appropriately. Section 2.2 returns to Arkin’s (and others’) ethical case for the development and use of AWS, for a second round of examination. I outline and endorse the suggestion that confining the operations of AWS to narrowly restricted domains may greatly reduce the difficulties involved in distinguishing between legitimate and illegitimate targets (§ 2.2.1). I then turn to discuss Arkin’s suggestion that as long as AWS would do better than human beings typically do at this task of discrimination their use would be ethical and argue that this relies upon a controversial, essentially consequentialist, moral framework (§ 2.2.2); I also discuss and criticize what purports to be a “consent based” variation of this argument and suggest that it too ultimately relies on consequentialist intuitions (§ 2.2.3). My provisional conclusion (§ 2.3) at the end of this discussion, however, is that while the use of AWS against insurgents or infantry in urban environments is unlikely to be ethical for the foreseeable future, none of the criticisms of AWS I have considered up to this point succeed in demonstrating that they are inherently incapable of discrimination or that their use in certain constrained roles against particular sorts of targets could not be ethical.

In the third and final section of the paper I therefore turn to a deeper investigation of the philosophical foundations of the Just War doctrine of *jus in bello* in order to develop a new account of the origins and force of the intuition that the use of “killer

robots" would necessarily be morally problematic. Drawing on an influential paper by Thomas Nagel, I argue (§ 3.1) that a proper understanding of the non-consequentialist foundations of the principle of distinction suggests that the use of AWS is unethical by virtue of failing to show appropriate respect for the humanity of our enemies. However, I suggest (§ 3.2) that the success of the *strongest* form of this argument depends upon our thinking of the robot itself as doing the killing. To the extent that we believe that, on the contrary, AWS are only the means whereby those who order them into action kill others, the idea that the use of AWS fails to respect the humanity of our enemy will turn upon an account of what is required by respect, which is essentially conventional. However, I defend the idea that what we express through our treatment of our enemies, while partially determined by convention, may be crucial to the ethics of that treatment (§ 3.3). Thus, finally, I conclude that although the theoretical foundations of the idea that AWS are weapons that are "evil in themselves" are weaker than critics have sometimes maintained, they are nonetheless strong enough to support the demand for a prohibition of the development and deployment of such weapons.

§ 1. The military case for autonomy

§ 1.1. Defining autonomous weapon systems.

Any sensible discussion of autonomous weapons must begin with clarifying what the author understands by "autonomy". The difficulties involved in providing a definition of autonomy, which is broad enough to capture what people take to be (alternatively) exciting and/or problematic about these systems without begging central questions in the debate about the ethics of their use, goes a long way towards explaining why the literature on the ethics of AWS is so vexed.[7] A minimal definition of autonomy is that a weapon or weapon system must be capable of some significant operation without direct human oversight. Perhaps the strongest definition of autonomy requires that a system be *morally* autonomous, that is to say be a moral agent with free will who is responsible for their own actions.[8]

Thus, as a number of authors have suggested, it is helpful to think about lethal autonomous operations as situated in a spectrum between these two sorts of cases,

with, for instance, antipersonnel mines — which “decide” when to explode on the basis of input from a pressure sensor — at one end and human beings or (theoretical) strong artificial intelligences at the other.[9] Of course, if one models the operations of autonomous weapons on the assumption that they are merely sophisticated landmines, then it may be hard to see what all the fuss is about. Alternatively, if a weapon must have the capacities of a human being to be autonomous then it may appear that we have nothing to worry about — as we are a long way from knowing how to design such computers.

Where the debate about lethal autonomous operations gets interesting is somewhere in the middle, wherein the operations of the weapon system possess a complexity that problematizes understanding them as merely a complex sort of landmine — without necessarily being so sophisticated as to require strong AI. As I (and others) have argued elsewhere, at the upper end of this range questions begin to arise about the appropriate way to allocate responsibility for the consequences of the operations of such systems.[10] However, as I will argue here, even where these questions do not arise, many people have the intuition that there is something morally problematic about robots killing people.

For the purpose of this paper, then, and in order to avoid prejudicing my discussion of the larger literature by insisting on a more precise — and, therefore, inevitably more controversial — definition, I will understand an “autonomous” weapon as one that is capable of being tasked with identifying possible targets and choosing which to attack, without human oversight, and that is sufficiently complex such that, even when it is functioning perfectly, there remains some uncertainty about which objects and/or persons it will attack and why. This admittedly rough-and-ready definition represents my attempt to pick out an interesting category of systems whilst avoiding entering into the extended and difficult argument about the precise nature of machine autonomy.[11] In particular, while the first clause of this definition accords with the influential US Department of Defense definition of autonomy in weapon systems,[12] the second clause is intended to help distinguish between automatic systems such as the Phalanx Close-in Weapon System and potentially more problematic autonomous systems the eventual operations of which we will struggle to account for without adopting “an intentional stance” and hypothesizing “its”

"reasons" for action.[13] Perhaps the paradigmatic case of the latter would be an autonomous weapon wherein genetic algorithms or machine learning played a central role in determining its behavior.[14] However, I also intend this definition to capture robots of sufficient complexity which do not rely on these mechanisms.

§ 1.2. An arms race to autonomous operations?

As mentioned above, many authorities now speculate that the perceived success of remotely piloted "drones" and other "Unmanned Systems" (UMS) in recent military conflicts means that development of AWS is more or less inevitable.[15] There are at least three military and/or technological logics that drive powerfully in the direction of the development of weapons systems that can operate – and kill people – autonomously.

First, the communications infrastructure that enable the operation of existing remotely piloted weapons, such as the United States' Predator and Reaper drones, place significant limitations on the operations of these systems even in their existing roles and, perhaps more importantly, on their survivability in future inter-state conflicts. Operations of long-range Unmanned Aerial Vehicles (UAVs) require the transmission of extremely large amounts of data by military satellite and radio systems. This places an upper limit on the number of such systems that can be fielded at any point in any given theatre of operations. It also restricts the capacity to field long-range UAVs to those few nations that have the ability to launch and operate communication satellites (or who are able to access bandwidth provided by their allies). The need to be able to sustain regular communication with human operators also effectively rules out a major role for remotely-operated submersibles in future naval combat, given that, in order to avoid detection and destruction by enemy forces, submarines must operate in a communications blackout when in — and prior to — combat. The communication systems necessary to allow remote piloting of UMS are also vulnerable to electronic countermeasures and/or kinetic attacks on the physical infrastructure that sustains them. In particular, one might expect that military communication satellites would be one of the first targets of attack in any future large-scale conflict involving major industrialized powers.

Developing and deploying AWS would therefore allow more weapons to be fielded and for the systems to be more survivable.[16]

Second, a number of technological factors have combined to greatly increase the tempo of battle over recent decades, especially in air-to-air combat. In conflicts involving modern high-technology military systems, victory may depend upon decisions that require integrating information from multiple sources and that must be made in a matter of seconds. The demands of modern combat already push the limits of what the human nervous system is capable of: in the future it may become the case that only AWS are capable of reacting within the time necessary to facilitate survival in a hostile environment.[17]

Third, beyond the need to make complex decisions extremely quickly, a number of other features of the operations of UAVs and other unmanned weapons systems suggest that it would be preferable to remove human beings from their operations. The routine operations of UAVs such as Predator and Reaper are, on all accounts, extremely boring for the vast majority of the time they are in theatre and consequently pilot fatigue and error remain a significant cause of accidents involving these systems. Autonomous systems might be less prone to mishaps at launch and recovery, and while travelling to the battlespace, than those controlled by human operators. Moreover, training costs, salaries, and medical benefits for the operators of remote systems are significant expenses in the budget lines of those armed forces that operate them. Just as unmanned systems have been touted as cheaper than the manned systems they replace, eventually autonomous systems may become less expensive to operate than remotely piloted systems.[18]

§ 2. The ethical case for autonomy

Independently of the military/technological logics driving towards the development of autonomous weapons, the development of these weapons might be desirable on other grounds. Alternatively, even if the development of these systems *is* more or less inevitable, it may still be the case that we should resist it on ethical grounds. Indeed, given the role played by competition between states for military advantage in driving the development of these systems, a global arms control treaty prohibiting autonomous weapons may represent the only way to prevent their being developed

and fielded.[19] It is therefore necessary to consider the ethical case for (and, later, against) the development and deployment of autonomous weapons.

§ 2.1. Arkin and his critics: Round I

Ronald Arkin is perhaps the most vocal and enthusiastic advocate for developing AWS writing about the topic today and is also actively involved in the project of developing them.[20] In his influential "The case for ethical autonomy in unmanned systems", Arkin adduces a number of arguments in favor of autonomous operations.[21] He argues that, in the future, AWS may be better able to meet the requirements of ethical conduct of war than human beings because robots: can be designed to accept higher risks in the pursuit of confidence in targeting decisions; will have better sensors; will not be swayed by emotions, such as fear or anger, which often prompt humans to act unethically; need not suffer from cognitive biases that afflict human decision-making; and, will be better able to quickly integrate information from a wide variety of sources.[22] As I will discuss further below, his identification of the relevant standard against which the ethical use of AWS should be measured as that achieved by human warfighters is also a crucial intellectual move in the debate about the ethics of autonomous weapons.

§ 2.1.1 Difficulties with discrimination

Critics of Arkin's proposal have been quick to point out just how far existing robots are from being able to outperform human beings when it comes to adherence to the requirements of *jus in bello*.[23] In particular, Arkin systematically underestimates the extent of the challenges involved in designing robots that can reliably distinguish legitimate from illegitimate targets in war.

§ 2.1.1.1. The challenge of machine perception

Despite many decades of research — and much progress in recent years — perception remains one of the "hard problems" of engineering. It is notoriously difficult for a computer to reliably identify objects of interest within a given environment and to distinguish different classes of objects. This is doubly the case in

crowded and complex unstructured environments and when the environment and the sensor are in motion relative to each other.

In order for an AWS to be able to play a similar role to Reaper or Predator drones in Afghanistan, for instance, and identify, track, and target armed men, it would need to be able to distinguish a person carrying and assault rifle from a person carrying a metal tube or a folded umbrella. Moreover, in order to be able to assess the likelihood of collateral damage and thus the extent to which a particular attack would satisfy the *jus in bello* requirement of proportionality (of which, more below) autonomous weapons will need to be able to reliably identify and enumerate civilian targets as well as potential military targets. Thus, it won't be sufficient for an AWS to be able to identify and track armed persons (for instance by being able to recognize the LIDAR[24] signature of an AK-47) — it must also be able to identify and track unarmed persons, including children, in order to be able to refrain from attacks on military targets that would involve an unacceptably high number of civilian casualties. Weapons intended to destroy armored vehicles must be capable of distinguishing them from all the different cars and trucks manufactured all around the world; autonomous submarines must be able to distinguish warships from merchant vessels, et cetera. AWS must be capable of achieving these tasks while the sensor is in motion and from a wide range of viewing angles in visually cluttered environments and in a wide range of lighting conditions.

As I will discuss further below, these problems may be more tractable in some domains than others. However they remain a formidable challenge to the development of AWS.

§ 2.1.1.2. Context, discrimination, and reasoning

In fact, the problem of discriminating between legitimate and illegitimate targets is much more difficult than my discussion thus far suggests. To begin with, as the current conflict in Afghanistan illustrates clearly, not every person carrying a weapon is a combatant — that is, someone directly engaged in the armed conflict. In many parts of the world where conflicts are currently taking place, carrying a weapon is a matter of male honor. Similarly, if less commonly, (decommissioned) tanks may be parked in playgrounds for children to climb on; foreign warships may be passing

through the territorial waters of the enemy power; and neutral troops or peacekeeping forces may be present in areas in which other legitimate targets are located. Thus, in order to discriminate between combatants and non-combatants, it will not be sufficient to be able to detect whether someone (or something) is carrying a weapon. Whether someone (or something) is a combatant or not – and therefore a potential military target — is a matter of context… and often of political context. It will be extremely difficult to program robots to be able to make this kind of discrimination.[25]

Even if a weapon system could reliably distinguish combatants from non-combatants — that is to say enemy forces who are directly involved in the prosecution of an armed conflict from everyone (and everything) else — this would still fall well short of being able to distinguish legitimate from illegitimate targets under *jus in bello*. Attacks on combatants may be illegitimate under *jus in bello* in at least three sorts of circumstances: first, where such attacks may be expected to cause a disproportionate number of civilian casualties (*Additional Protocol I to the Geneva Conventions*, Article 57);[26] second, where they would constitute an unnecessarily destructive and excessive use of force ;[27] third, where the target has indicated a desire to surrender or is otherwise *hors de combat* (*Additional Protocol I to the Geneva Conventions*, Article 41*)*.[28]

Before it would be ethical to deploy an AWS, then, the system will need to be capable of making *these* sorts of discriminations, all of which involve reasoning at a high level of abstraction.

Thus, for instance, how many non-combatant deaths it would be permissible to knowingly cause in the course of an attack on a legitimate military target depends on: the military advantage the destruction of the target is intended to serve; the availability of alternative means of attacking the target; the consequences of not attacking the target at that time (which in turn is partially a function of the likelihood that an opportunity to attack the target will arise again); the availability of alternative means of achieving the desired military objective; and, the weaponry available to conduct the attack. Similarly, whether an attack would constitute an unnecessarily destructive use of force (which it may be even where there is no risk of killing non-combatants) is a function of: the nature of the military objective being targeted; the

extent of the military advantage the attack is intended to secure; and, the availability of alternative, less destructive, means of achieving it. Assessing these matters requires extensive knowledge and understanding of the world, including the capacity to interpret and predict the actions of human beings. In particular, assessing the extent to which an attack will achieve a definite military advantage requires an understanding of the balance and disposition of forces in the battlespace, the capacity to anticipate the probable responses of the enemy to various threats and circumstances, and an awareness of wider strategic and political considerations.[29] It is difficult to imagine how any computer could make these sorts of judgements short of the development of a human level general intelligence — that is, "strong" AI.[30]

Identifying when enemy forces have surrendered or are otherwise "*hors de combat*" is also a profound challenge for any autonomous system.[31] Perhaps it will be possible to program AWS to recognize the white flag of surrender or to promulgate a convention that all combatants will carry a "surrender beacon" that indicates when they are no longer participating in hostilities.[32] Yet these measures would not resolve the problem of identifying those who are *hors de combat*. An unconscious and gravely wounded soldier separated from his or her comrades is not a legitimate target even if he or she has not indicated the desire to surrender (indeed, he or she may have had no opportunity to do so) but may be extremely hard for a computer distinguish from a soldier lying in ambush. Similarly, a ship that has had its guns destroyed or that has been holed below the water so that all hands are required just to remain afloat — and is therefore no military threat — will not always have a different radar or infrared profile to a functioning warship. Human beings can often — if not always — recognize such situations with reference to context and expectations about how people will behave in various circumstances. Again, short of possessing a human level general intelligence, it is difficult to imagine how a computer could make these discriminations.

§ 2.1.2. Possible solutions? "Ethical" robots and human oversight

Arkin has offered two responses to these sorts of criticisms which are, I believe, inadequate.

§ 2.1.2.1. Ethical robots?

First, Arkin has suggested that it should be possible to build the capacity to comply with the relevant ethical imperatives "into" the weapon system in the form of what he calls an "ethical governor".[33] This will not, of course, address the problem of identifying and classifying objects in complex environments — although it is possible that improvements in computer vision technology will reduce these to a manageable level. More fundamentally, it presumes an impoverished account of ethics as a system of clearly defined rules with a clearly defined hierarchy for resolving clashes between them. The problem of identifying and adjudicating between the ethical imperatives relevant to the operation of an AWS is more profound than Arkin's project allows.

The sketches of deontological or utilitarian systems of ethics that philosophers have developed are just that – sketches. The task of ethical theory is to try to explain and systematize the ethical intuitions that properly situated and adequately informed persons evince when confronted with various ethical dilemmas. These intuitions are extremely complex and context dependent, which is why philosophers are still arguing about whether they are primarily deontological or consequentialist or perhaps virtue-theoretical. It is these — still poorly understood and often highly contested — intuitions that a machine would need to be capable of replicating in order for it to "do" ethics. Moreover, even the schematized accounts of some subset of these intuitions that philosophers have developed require agents to reason at a high level of abstraction and to be able to make complex contextual judgements for their application. For instance, consequentialists must be capable of predicting the consequences of our actions in the real world, making a judgement about when this attempt to track consequences — which, after all, are essentially infinite — may reasonably be curtailed, and assessing the relative value of different states of the world. It is unclear whether human beings can even do this reliably, which itself is a reason to be cautious about embracing consequentialism, but it seems highly unlikely that machines will ever be able to do so short of achieving human-level general intelligence. Similarly, Kantian ethics requires agents to identify the moral principles relevant to their circumstances and resolve any clashes between them — again a task that requires a high degree of critical intelligence.[34]

However, the most fundamental barrier to building an “ethical robot” is that ethics is a realm of meanings. That is to say, understanding the nature of our actions — what they mean — is fundamental to ethical reasoning and behavior.[35] Thus, for instance, intentionally killing a human being is murder — unless it happens in the course of a declared armed conflict and the killer and the victim are both combatants, it is done in self-defense, or it has been mandated by the state after a fair criminal trial. In order to be able to judge whether a particular killing is murder or not, then, one must be able to reliably track the application of concepts like “intention”, rights, legitimacy, and justice — a task which seems likely to remain well beyond the capacities of any computer for the foreseeable future. Perhaps more importantly, the *meaning* of murder — why it is a great evil — is not captured by any set of rules that distinguish murder from other forms of killing but only by its place within a wider network of moral and emotional responses. The idea that a properly programmed machine could behave ethically, short of becoming a full moral agent, only makes sense on the basis of deep-seated behaviorism of the sort that has haunted computer science and cognitive science for decades.

§ 2.1.2.2. Human oversight?

Arkin’s second suggestion is that weaponized robots could be designed to allow a human operator to monitor the ethical reasoning of the robot in order that they might intervene whenever they anticipate that the robot is about to do something unethical.[36] Other authors have suggested that AWS could be designed to contact and await instruction from a human operator whenever they encounter a situation their own programming is unable to resolve.[37]

This is problematic for two reasons. First, the need to “phone home” for ethical reassurance would mitigate two of the main military advantages of autonomous weapons, which are their capacity will to make decisions more rapidly than human beings can[38] and their capacity to operate in contexts where it is difficult to establish and maintain reliable communications to a human pilot.[39] If an “autonomous” weapon has to rely on human supervision to attack targets in complex environments, it would be at most *semi*-autonomous.[40] Second, it presumes that the problem of accurately representing the ethical questions at stake and/or determining when the ethics of an attack is uncertain is more tractable than the problem of resolving the

uncertainty. However, the appropriate description of the ethical issues at stake and/or the assessment that an ethical question is difficult or controversial itself requires ethical deliberation at the same level of complexity as the original ethical question. Thus, if we can't trust a machine to reliably make ethical judgements, we cannot trust it to identify when its judgements might be unreliable.

§ 2.2. Arkin and his critics: Round II

For these reasons, I believe that Arkin's critics are correct in arguing that the difficulties of reliably distinguishing between legitimate and illegitimate targets in crowded and complex environments probably rules out the ethical use of AWS in many roles for the foreseeable future.[41] However, Arkin does have two replies to these sorts of criticisms available to him at this point, which are more compelling. First, the problem of discriminating between legitimate and illegitimate targets is much more tractable in specific, restricted, domains than Arkin's critics — and the arguments above — suggest. Second, Arkin has argued that the relevant standard of reliability in discriminating between legitimate and illegitimate targets, which robots need to attain in order for their use to be ethical, is that achieved by human warfighters, which is much lower than might first appear. If AWS would kill fewer non-combatants than human troops, this establishes a strong consequentialist case for their deployment, regardless of other ethical concerns about them. As I will discuss below, a number of other advocates for autonomous weapons have also made a non-consequentialist version of an argument for the ethical use of autonomous weapons from their putative future reliability in distinction compared to human warfighters.

§ 2.2.1. A way forward? Autonomous operation in restricted domains

One possible solution to the problems involved in ensuring that AWS are capable of complying with the *jus in bello* principles of distinction and proportionality would be to narrowly constrain the domain of the operations of AWS and/or the sorts of systems that they are tasked with destroying.[42] How difficult it is to distinguish between a military and non-military object depends on the features of each as revealed by the sensors available to robots. How difficult it is to avoid "collateral damage" when

attacking military targets depends upon the relative number of legitimate and illegitimate targets within the area of operations. In anti-submarine warfare, for instance, there are few civilian targets.[43] Similarly, in air-to-air combat, counter-artillery operations, or suppression of enemy air defenses, it is relatively straightforward to distinguish military from non-military systems.[44] Tanks and mechanized artillery, and — to a lesser extent — naval assets also have, for the most part, distinctive visual silhouettes and radar and infrared signatures as compared to the non-military systems (cars, trucks, merchant ships, et cetera) amongst which they might be found. When potential targets are mechanized and combat is confined to a distinct theatre of operations, it is much more plausible to hold that autonomous weapons will be capable of reliably identifying potential military targets and distinguishing them from non-combatants. Indeed, existing target identification systems are already capable of reliably distinguishing between military and civilian systems in these domains.[45] The claim that autonomous weapons will never be capable of reliably distinguishing between military and non-military targets in these domains therefore simply looks false.

Admittedly, as I observed above, not every military target is a legitimate target according to the principles of *jus in bello*. Even in these restricted domains, then, the challenge of discriminating between legitimate and illegitimate targets is harder than first appears. Systems must be capable not only of identifying potential military targets but the presence of civilian populations or installations that might also be damaged in the attack in order to be able to avoid causing disproportionate civilian casualties. They must also be able to determine when attacks on military targets are justified by the principle of necessity and secure a definite military advantage.[46] Yet when combat is occurring in a discrete geographical area, especially in the air or in space or underwater or (more controversially) when civilians have been given sufficient warning to vacate an area in which hostilities are about to commence, *and* when victory in this context would advance an important military objective, then it might prove possible to guarantee that the destruction of any of the military objects present would be justified. It is less clear, however, that the problem of identifying forces that have surrendered or otherwise *hors de combat* is any more tractable simply by virtue of being confined to a restricted geographical area. The idea of "surrender beacons" is, perhaps, more practicable when the forces engaged are

military assets rather than personnel. Yet the problem of identifying when enemy forces are illegitimate targets by virtue of being so incapacitated by wounds or damage that it is no longer reasonable to consider them to constitute a military threat remains profound. Nevertheless, it seems likely that by confining the operations of AWS to a carefully delineated “kill box” it might be possible to greatly reduce the risk of attacks on illegitimate targets.

§ 2.2.2. The consequentialist case for autonomy

Arkin has a further string to his bow at this point, moreover. He can simply concede that the task of designing an AWS capable of distinguishing between legitimate and illegitimate targets is a difficult one but claim that the problem of designing an AWS that does better than human beings in this task *is* tractable.[47] Indeed, by highlighting the real world attitudes and behaviors of US soldiers deployed in Iraq, Arkin has argued persuasively that human warfighters are actually quite bad at behaving ethically during wartime.[48]

However, this is arguably entirely the wrong standard to set when considering whether the use of a weapon system would be ethical.[49] Because what is at stake is the value of an innocent human life, when it comes to protecting non-combatants from deliberate (or negligent) attack it might be argued that the relevant ethical standard is perfection. It would not justify a human being deliberately or negligently killing two civilians for every ten combatants, for instance, to point out that other warfighters typically kill three civilians for every ten combatants in similar circumstances. There is a clear sense in which human warfighters *could* never deliberately target non-combatants or use disproportionate force, *et cetera*. Thus, it is reasonable to expect perfect ethical compliance from human beings even if in fact they seldom achieve this.[50] Putting a machine into combat when it would be *un*reasonable to expect that it *won’t* violate the requirements of distinction and proportionality could only be defensible if one believes that the only relevant consideration was the eventual number of civilian casualties. Arkin’s argument about the benefits of autonomous targeting therefore depends on adopting a consequentialist ethical framework that is concerned only with the reduction of civilian casualties, which is controversial, especially in the context of the ethics of warfare.

§ 2.2.3. Consent from civilians as grounds for autonomous operations?

There is, however, another version of this argument, which buttresses the claim about the relative effectiveness of weaponized robots with an appeal to the agency of those their operations might threaten. Thus for instance Brian Williams has argued that the civilian population in the autonomous tribal areas of Pakistan actually prefer operations against Al Qaeda militants to be conducted via drone attacks because the alternative — anti-terrorist operations by the Pakistani armed forces — is so much more destructive.[51] By appealing to the consent of the civilian population this version of the argument mobilizes powerful deontological intuitions.[52]

Yet, on closer inspection, the argument from consent is a red herring. Consider the nature of the circumstances in which civilians in the Sudan, for instance, might say that they would prefer the operations of an AWS to the deployment of human warfighters in the areas where they live. We must imagine that the area in which they live has become the theatre of operations for the military forces of another nation and their lives are at risk because of this. Importantly, they themselves may bear no responsibility for the events which led to this circumstance: their nation may be the victim of an unjust attack; or, they may have voted and demonstrated against their own government, which has in fact provided another state with a just cause to initiate hostilities.[53] In this context, they face the choice of being threatened with death as a result of the operations of poorly trained and/or racist heavily-armed 19-year-olds or autonomous weapons. Say that they prefer that the attacking state deploy AWS to the alternative in which they are at a higher risk of death. This is a less-than-ideal, to say the least, circumstance in which to be trying to secure a meaningful consent from someone in relation to how they are treated. Indeed, in many ways one would have to say that this “consent” is coerced. It is as though we had said to the civilian population in the theatre of operations “let us risk your life with AWS otherwise we will threaten you with our (human) armed forces”. While it may be rational for them to prefer the first option, the fact that they do so hardly justifies our proceeding with it.

Critics of existing drones and future AWS typically claim to be motivated, at least in part, out of a concern for the rights of those who live in the territories in which these

weapons have and will be deployed,[54] so the suggestion that critics are willing to ignore the expressed desires of those populations is therefore unsettling. However, while reference to the rights and agency of civilians is undoubtedly a powerful rhetorical strategy in defense of the use of robotic weapons, whatever force it has is not a function of consent but rather reflects the consequentialist considerations adduced above.

§ 2.3. The prospects for ethical autonomous targeting thus far

The prospects for ethical autonomous targeting are, therefore, according to my investigation here, mixed. Critics of AWS are correct in holding that the difficulties involved in operating in accordance with the principles of *jus in bello* — and, in particular, the requirements of distinction, proportionality, and military necessity — are profound and unlikely to be resolvable by computer in the context of combat in urban environments in operations against insurgents or infantry for the foreseeable future. On the other hand, in specific limited domains — and, in particular, in operations against naval assets, tanks and self-propelled artillery, and/or aircraft in a given geographical area — it may be possible for robots to distinguish between legitimate and illegitimate targets with a high degree of reliability. Indeed, in this context AWS might even prove *more* reliable than human beings at distinguishing between legitimate and illegitimate targets, as Arkin has argued. At the very least, the possibility of deploying AWS in this fashion establishes that they are not, as some have suggested, “inherently indiscriminate” weapons.

At this stage, then, it would be premature to conclude that any of the ethical arguments I have surveyed thus far stand as an insurmountable barrier to the ethical operations of AWS. If we are to explain the widespread ethical intuition that there is something profoundly disturbing about the prospect of “killer robots” we must delve deeper into the philosophical foundations of just war theory.

§ 3. Robots and respect

There is, I think, as I have argued elsewhere a case to be made against developing and deploying robotic weapons in general — both tele-operated and autonomous weapon systems — on the basis of the doctrine of *jus ad bellum*.[55] The fact that robotic weapons hold out the prospect of the use of force without risk of one's troops coming home in body bags and the likelihood that such systems will be used in more aggressive postures during peace time again due to the lack of threat to the life of the "pilot" suggests that these systems will lower the threshold of conflict and make war more likely.[56] Furthermore, as Paul Kahn has argued, the pursuit of risk-free warfare problematizes the justification of wars of humanitarian intervention by juxtaposing the high value based on the lives of our own military personnel against the lower value placed on the lives of those in the theatre of conflict, whose rights and welfare are supposed to justify the intervention, but who are placed at higher risk of death as a result of the use of robotic weapons.[57] However, these sorts of concerns are not specific to AWS and have force against a wider range of means of long-distance war fighting.[58]

§ 3.1. AWS and *jus in bello*

If there is going to be anything uniquely morally problematic about AWS, then, the explanation will need to be located within the doctrine of *jus in bello*. In a famous and influential article on the moral foundations of the principles of *jus in bello*, Thomas Nagel argued that the force of these principles could only be explained by the idea that they are founded in absolutist moral reasoning.[59] Nagel develops an account of the key injunctions of *jus in bello* by way of a (essentially Kantian) principle of respect for the moral humanity of those involved in war. He argues that even during wartime it is essential that we acknowledge the personhood of those with whom we interact and that,

> "… whatever one does to another person intentionally must be aimed at him as a subject, with the intention that he receive it as a subject. It should manifest an attitude to him rather than just to the situation, and he should be able to recognize it and identify himself as its object."[60]

Another way of putting this is that we must maintain an “interpersonal” relationship with other human beings even during wartime. Obviously, if this principle is to serve as a guide to the ethics of war—rather than a prohibition against war —the decision to take another person’s life must be compatible with such a relationship.[61]

Thus, on Nagel’s account, applying the principles of distinction and proportionality involves establishing this interpersonal relationship with those who are the targets of a lethal attack and acknowledging the morally relevant features that render them combatants or otherwise liable for being subject to a risk of being killed. In particular, in granting the possibility that they might have a right not to be subject to direct attack by virtue of being a non-combatant one is acknowledging their humanity.[62] This relationship is fundamentally a relationship between agents — indeed, between members of the Kantian “kingdom of ends”.

Immediately, then, we can see why AWS might be thought to be morally problematic regardless of how reliable they might be at distinguishing between legitimate and illegitimate targets.[63] When an AWS decides to launch an attack the relevant interpersonal relationship is missing.[64] Indeed, in some fundamental sense there is no one who decides whether the target of the attack should live or die. The absence of a human intention when it comes to the killing of a human being in war appears profoundly disrespectful.

§ 3.2. “Killer robots” or “robots for killing”?

I believe this intuition is central to popular concerns about “killer robots.” However, importantly, this way of understanding the ethics of AWS treats the robot as though “it” were doing the killing. Short of the development of artificial intelligences that are actually moral agents, this seems problematic. We might equally well think of the robot as a tool by which one person attempts to kill another—albeit an indeterminate other.[65] The relevant interpersonal relationship would then be that between the officer who authorizes the release of the weapon and those they intend to kill. Neither the fact that the person who authorizes the launch doesn’t know precisely who they are killing when they send the AWS into action, nor the fact that the identity of those persons may be objectively indeterminate at the point of launch, seems to rule out the possibility of the appropriate sort of relationship of respect. When a

missile officer launches a cruise missile to strike a set of GPS coordinates 1000 km away it is highly unlikely that he (or she) knows the identity of those he (or she) intends to kill.[66] Mines and Improvised Explosive Devices (IEDs) kill anyone who happens to trigger them and thus attack persons whose identity is actually indeterminate and not merely contingently unknown. If an interpersonal relationship is possible while using these weapons, it is not clear why there could not be an interpersonal relationship between the commanding officer who launches an AWS and the people it kills. Thus, neither of these features of AWS would appear to function as an absolute barrier to the existence of the appropriate relationship of respect.

It is worth observing that this comparison is not *entirely* favorable to either AWS or these other sorts of weapons. People often do feel uneasy about the ethics of anonymous long-range killing and also — perhaps especially — about landmines and IEDs.[67] Highlighting the analogies with AWS might even render people more uncomfortable with these more familiar weapons. Nevertheless, in so far as contemporary thinking about *jus in bello* is yet to decisively reject other sorts of weapons which kill persons whose identity is unknown or actually indeterminate without risk to the user, it might appear unfair to reject AWS on these grounds.

It is also worth noting that the language of machine autonomy sits uneasily alongside the claim that autonomous systems are properly thought of merely as tools to realise the intentions of those who wield them.[68] The more advocates of robotic weapons laud their capacity to make complex decisions without input from a human operator, the more difficult it is to believe that AWS connect the killer and the killed directly enough to sustain the interpersonal relationship that Nagel argues is essential to the principle of distinction. That is to say, even if the machine is not a full moral agent, it is tempting to think that it might be an “artificial agent” with sufficient agency, or a simulacrum of such, to problematize the “transmission” of intention. This is why I have argued elsewhere that the use of such systems may render the attribution of responsibility for the actions of AWS to their operators problematic.[69] As Heather Roff[70] has put it, drawing on the work of Matthias,[71] the use of autonomous weapons seems to risk a “responsibility gap” — and where this gap exists, it will not be

plausible to hold that when a commander sends an AWS into action he or she is acknowledging the humanity of those the machine eventually kills.

However, this argument about responsibility has been controversial and ultimately, I suspect, turns upon an understanding of autonomy that is richer and more demanding than that I have assumed here.[72] At least some of the "autonomous" weapons currently on the drawing boards of military research laboratories seem likely to possess no agency whatsoever and thus arguably *should* be thought of as transmitting the intentions of those who command their use.

§ 3.3. What it says about our attitude towards our enemies when we send AWS to kill them

Yet this is not the end of an investigation into the implications of a concern for respect for the ethics of AWS. As Nagel acknowledges in his original paper, there is a conventional element to our understanding of the requirements of respect.[73] What counts as humane or inhumane treatment of a prisoner, for instance, or as desecration of a corpse, is partially a function of contemporary social understandings. Thus, certain restrictions on the treatment of enemy combatants during wartime have ethical force simply by virtue of being widely shared. Moreover, there is ample evidence that existing social understandings concerning the respectful treatment of human beings argue against the use of AWS being ethical. A recent public opinion survey found high levels of hostility to the prospect of robots being licensed to kill human beings.[74] The most plausible interpretation of this hostility is that most people already feel strongly that sending a robot to kill human beings would express a profound disrespect of the value of an individual human life. To render someone liable to be killed by robot is to treat them like vermin; this is not how we should treat our fellow human beings.[75]

The idea that what we "express" when we treat our enemies in a certain way is sometimes crucial to the morality of warfare is evidenced by the extent and force of the intuition that the mutilation and mistreatment of corpses is a war crime. Such desecration, occurring after death as it does, does not inflict "unnecessary suffering" on the enemy; rather, it is wrong precisely because and insofar as it expresses a profound disrespect for them and their humanity. Importantly, while the content of

what counts as a “mistreatment” or “mutilation” is conventional and may change over time, the intuition that we are obligated to treat even the corpses of our enemies with respect is deeper and much less susceptible to revision.

The ethical principles of *jus in bello* allow that we may permissibly attempt to kill our enemy, even using means that will inevitably leave them dying of burns or mangled and bleeding to death in the mud. Yet these principles also place restrictions on the means we may use and on our treatment of the enemy more generally. I have argued — following Nagel — that a fundamental moral requirement is that this treatment should be compatible with respect for the humanity of our enemy and that the content of this concept is partially determined by shared social understandings regarding what counts as respectful treatment. Furthermore, I have suggested that widespread public revulsion at the idea of autonomous weapons should be interpreted as conveying the belief that the use of AWS is incompatible with such respect. If I am correct in this, then, even if an interpersonal relationship may be held to exist between the commanding officer who orders the launch of an autonomous weapon system and the individuals killed by that system, the appropriate description is that it is one of disrespect rather than respect.

Interestingly, conceiving of AWS as the means whereby the person who authorizes the launch of the system attempts to kill their targets explains why this means of killing may be profoundly disrespectful even though those who are killed by robots may be killed using weapons that are – understood in a more narrow sense — identical to those that a human being might use.[76] Thus, for instance, in conventional military terminology, a Predator drone – and by extension, perhaps, a future AWS — would ordinarily be understood as a platform from which a weapon (a Hellfire missile) may be delivered. Correspondingly, defenders of AWS have suggested that robot weapons couldn’t be morally problematic “in themselves” because it could make no difference to the suffering or the nature of the death of those they kill whether a Hellfire missile was fired from an AWS, from a (remotely piloted) Predator drone, or from a (manned) Apache helicopter.[77] Yet if we are going to understand the AWS as the means whereby the person who launches it attacks targets when it comes to a concern for the appropriate relationship of respect for the humanity of our enemies, we must also understand it as the weapon he/she uses to kill. Indeed, it is

quite clear that the officer who launches an AWS is *not* launching a Hellfire missile. Consequently, there is nothing especially problematic with describing an AWS as an illegitimate means of killing.

§ 3.4. The case for banning AWS

I believe that the contemporary campaign to ban autonomous weapons should therefore be understood as an attempt to entrench a powerful intuitive objection to the prospect of a disturbing new class of weapons in international law: AWS should be acknowledged as "*mala in se*" by virtue of the extent to which they violate the requirement of respect for the humanity of our enemies which underlies the principles of *jus in bello*.[78] That the boundaries of such respect are sometimes — as in this case — determined by convention (in the sense of shared social understandings rather than formal rules) does not detract from the fact that it is fundamental to the ethics of war.

A number of critics of the campaign to ban AWS have objected that this proposal is premature and that until we have seen robot weapons in action, we cannot judge whether they would be any better or worse morally speaking than existing weapons systems.[79] Yet insofar as a ban on AWS is intended to acknowledge that *the use* (rather than the effects) of robotic weapons disrespects the humanity of their targets, this objection has little force.

There is, of course, something more-than-a-little intellectually unsettling about the attempt to place a class of weapons in the category of the *mala in se* through legislation or (legal) convention: this category is supposed to be characterized precisely by the fact that its members may be recognized independently of positive law. Yet if we are honest about the matter, we will admit that there has always been controversy about the extent of this class and that some types of weapons now held to be evil in themselves were once widely believed to be legitimate means of waging war; only after a period of contestation and moral argument were technologies such as chemical and nuclear weapons acknowledged as prohibited means of war.[80] The current situation regarding the campaign against AWS is therefore analogous to the campaign against the use of chemical weapons at the beginning of the 20th century or the use of cluster munitions in the 1990s.[81] Should this campaign ultimately prove

successful, we will understand it to have recognized truths about these weapons which existed independently of — and prior to — the resulting prohibition.[82] In the meantime, the strengths and popular currency of the intuition that the use of AWS would profoundly disrespect the humanity of those they are tasked to kill is, I would submit, sufficient justification to try to establish such a prohibition.

Conclusion

I have argued that the prospects of AWS being capable of meeting the *jus in bello* requirements of distinction, proportionality, and military necessity in the context of counter insurgency warfare and/or complex urban environments is remote. However, in particular limited domains, the difficulties of making the relevant discriminations are much reduced and the major barrier to AWS being able to reliably distinguish between legitimate and illegitimate targets would appear to be their capacity to detect when enemy forces have surrendered or are otherwise *hors de combat*. If these difficulties can be overcome or if the risk of errors along these lines is low in context then concerns about the capacity of AWS to identify and attack only the appropriate targets are unlikely to rule out their use being ethical.

The strength of the case for autonomous weapons will also depend on how we assess the relative weight of consequentialist and deontological considerations in the ethics of war. To the extent that we believe that our main consideration should be to reduce the number of non-combatant deaths that occur in war, it becomes easier to imagine AWS being ethical: all that would be required would be for them to be better than human beings at distinguishing between legitimate from illegitimate targets in some given domain.[83] However, if we are concerned with what we owe non-combatants and others who are not legitimately subject to lethal force, then the merely statistical form of discrimination achievable by robots may be insufficient.

The deeper issue regarding the ethics of AWS, though, concerns whether the use of these weapons is compatible with the requirement of respect for the humanity of our enemies, which, I have argued, underpins the principles of *jus in bello*. If we understand AWS as “artificial agents”, which choose which targets to attack and when, it is likely that the necessary relationship of respect is absent and, therefore, that their use would be unethical. Yet in many cases it may in fact be more plausible

to consider the AWS as the means whereby the person who is responsible for its launch kills those that it is tasked to attack. However, this means may itself be unethical by virtue of expressing a profound disrespect for the humanity of our enemies. In so far as it relies on an account of the content of what is required by respect that is essentially conventional the case against AWS is (much) weaker than critics of these systems might prefer. Nevertheless, it is, as I have argued here, equal to the task of grounding an international treaty prohibiting the development and deployment of AWS on the grounds that such weapons are "evil in themselves".

There are, of course, further questions about whether it is realistic to imagine such a prohibition coming into force — let alone being effective in preventing (or at least significantly delaying) the deployment of AWS.[84] States that have the capacity to develop or field AWS will also have to confront the question as to whether the ethical case for any such treaty is worth whatever sacrifice of military advantage might be involved with signing it.[85] These are matters for further discussion and argument — and where, moreover, the style of argument appropriate to a philosophy paper may have little to contribute.[86] What I have shown here is that there is an ethical case to be made for working towards such a treaty.[87]

---

[1] Elsewhere in the literature AWS are sometimes referred to as Lethal Autonomous Robots (LARs).
[2] Kenneth Anderson and Matthew C. Waxman, "Law and Ethics for Robot Soldiers," *Policy Review* 176 (2012); Ronald C. Arkin, *Governing Lethal Behavior in Autonomous Robots* (Boca Raton: CRC Press, 2009); Gary E. Marchant et al., "International Governance of Autonomous Military Robots," *Columbia University Review of Science & Technology Law* 12 (2011); Department of Defense, *Unmanned Systems Integrated Roadmap: FY2011-2036* (Washington, DC: Department of Defense, 2012); Michael N. Schmitt and Jeffrey S. Thurnher, "Out of the Loop: Autonomous Weapon Systems and the Law of Armed Conflict," *Harvard National Security Journal* 4, no. 2 (2013); Peter W. Singer, *Wired for War: The Robotics Revolution and Conflict in the 21st Century* (New York: The Penguin Press, 2009); Robert O. Work and Shawn Brimley, *20YY: Preparing for War in the Robotic Age* (Centre for a New American Security, 2014).
[3] Human Rights Watch, *Losing Humanity: The Case against Killer Robots* (2012), http://www.hrw.org/reports/2012/11/19/losing-humanity-0.
[4] Campaign to Stop Killer Robots, "Campaign to Stop Killer Robots: About Us," (2013), http://www.stopkillerrobots.org/about-us/.
[5] Charli Carpenter, "Beware the Killer Robots: Inside the Debate over Autonomous Weapons," *Foreign Affairs (Online)*, July 3, 2013, http://www.foreignaffairs.com/articles/139554/charli-carpenter/beware-the-killer-robots#.
[6] I will leave the task of determining the *legality* of AWS under international humanitarian law to those better qualified to address it. However, in so far as the Just War Theory doctrine of *jus in bello* is developed and expressed in both legal and philosophical texts, I will occasionally refer to the relevant legal standards in the course of my argument, which concerns the *ethics* of autonomous targeting.
[7] For the argument that this problem also impacts on the science and engineering of AWS, see Defense Science Board: US Department of Defense, *The Role of Autonomy in DoD Systems*, (Washington, DC: Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, 2012), 23-24.

[8] Robert Sparrow, "Killer Robots," *Journal of Applied Philosophy* 24, no. 1 (2007).
[9] Human Rights Watch, *Losing Humanity*, 6-20; Schmitt and Thurnher, "Out of the Loop"; Armin Krishnan, *Killer Robots: Legality and Ethicality of Autonomous Weapons* (Farnham, England: Ashgate Publishing, 2009), 43-45.
[10] Heather M. Roff, "Killing in War: Responsibility, Liability, and Lethal Autonomous Robots," in *Routledge Handbook of Ethics and War: Just War Theory in the 21st Century*, eds. Fritz Allhoff, Nicholas G Evans, and Adam Henschke (Milton Park, Oxon: Routledge, 2013); Sparrow, "Killer Robots."
[11] I have discussed this question at more length in Sparrow, "Killer Robots."
[12] U.S. Department of Defense, *DoD Directive 3000.09: Autonomy in Weapon Systems*, (Washington, DC: Department of Defense, 2012), http://www.dtic.mil/whs/directives/corres/pdf/300009p.pdf
[13] Daniel C. Dennett, *The Intentional Stance* (Cambridge, Massachusetts: MIT press, 1987).
[14] Andreas Matthias, "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata," *Ethics and Information Technology* 6, no. 3 (2004).
[15] Anderson and Waxman, "Law and Ethics for Robot Soldiers"; Arkin, *Governing Lethal Behavior in Autonomous Robots*; Marchant et al., "International Governance of Autonomous Military Robots"; Work and Brimley, *20YY*. The United States' Department of Defense (2012) clarified its own policy in relation to the development and use of AWS in DoD *Directive 3000.09: Autonomy in Weapon Systems*, which some (see, for instance, Spencer Ackerman, "Pentagon: A Human Will Always Decide When a Robot Kills You," *Wired*, 26 November 2012) have read as prohibiting the use of AWS armed with lethal weapons against human targets. However, see Mark Gubrud, "Us Killer Robot Policy: Full Speed Ahead," *Bulletin of the Atomic Scientists* (2013).
[16] Kenneth Anderson and Matthew C. Waxman, "Law and Ethics for Autonomous Weapon Systems: Why a Ban Won't Work and How the Laws of War Can," *Jean Perkins Task Force on National Security and Law Essay Series* (2013): 7; Schmitt and Thurnher, "Out of the Loop," 238.
[17] Thomas K. Adams, "Future Warfare and the Decline of Human Decision Making," *Parameters* 31, no. 4 (2001).
[18] Schmitt and Thurnher, "Out of the Loop." The costs of developing and fielding AWS are another matter entirely. Complex information technology systems are notoriously prone to running over budget and delivering less than was promised. Developing the computer software required for autonomy and debugging it effectively may be very expensive indeed.
[19] Mark Gubrud, "Stopping Killer Robots," *Bulletin of the Atomic Scientists* 70, no. 1 (2014); Human Rights Watch, *Losing Humanity*; International Committee for Robot Arms Control, *Mission Statement*, (2009), http://icrac.net/statements/; Robert Sparrow, "Predators or Plowshares? Arms Control of Robotic Weapons," *IEEE Technology and Society* 28, no. 1 (2009).
[20] Ronald C. Arkin, "On the Ethical Quandaries of a Practicing Roboticist: A First-Hand Look," in *Current Issues in Computing and Philosophy*, eds. Adam Briggle, Katinka Waelbers, and Philip Brey (Amsterdam: IOS Press, 2008); Arkin, *Governing Lethal Behavior in Autonomous Robots*.
[21] Ronald C. Arkin, "The Case for Ethical Autonomy in Unmanned Systems," *Journal of Military Ethics* 9, no. 4 (2010). See also Marchant et al., " International Governance of Autonomous Military Robots," 279-81; Department of Defense, *Unmanned Systems Integrated Roadmap: FY2011-2036*, 43-51.
[22] For a critical evaluation of these claims, see Ryan Tonkens, "The Case against Robotic Warfare: A Response to Arkin," *Journal of Military Ethics* 11, no. 2 (2012).
[23] Human Rights Watch, *Losing Humanity*; Noel E. Sharkey, "Autonomous Robots and the Automation of Warfare," *International Humanitarian Law Magazine* 2 (2012); Noel E. Sharkey, "The Evitability of Autonomous Robot Warfare," *International Review of the Red Cross* 94, no. 886 (2012).
[24] LIght Detection And Ranging.
[25] For a useful discussion of the ways in which applying the principle of distinction requires assessment of intention and of just how hard this problem is likely to be for a machine, see Marcello Guarini and Paul Bello, "Robotic Warfare: Some Challenges in Moving from Noncivilian to Civilian Theaters," in *Robot Ethics: The Ethical and Social Implications of Robotics*, eds. Patrick Lin, Keith Abney, and George A. Bekey (Cambridge, MA; London: MIT Press, 2012). Again, as Guarini and Bello concede and I will discuss further below, in some — very restricted — circumstances it *may* be reasonable to treat every person carrying a weapon and every weaponised system, within a narrowly defined geographical area, as a combatant.
[26] *Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (Protocol I), 8 June 1977*, http://www.icrc.org/Ihl.nsf/INTRO/470?OpenDocument.
[27] Geoffrey S. Corn et al., *The Law of Armed Conflict: An Operational Approach* (New York: Wolters Kluwer Law & Business, 2012), 115-17.

[28] Corn et al, *Law of Armed Conflict*, 165-66; *Additional Protocol I.*
[29] In the course of writing this paper, I was fortunate enough to read a draft of a manuscript by Heather Roff addressing the prospects for autonomous weapons meeting the requirements of the principle of distinction. My discussion here has undoubtedly been influenced by her insightful treatment of the topic. Schmitt and Thurnher, "Out of the Loop," also contains a useful discussion of this question.
[30] Schmitt and Thurnher, "Out of the Loop," 265-66; Markus Wagner, "Taking Humans out of the Loop: Implications for International Humanitarian Law," *Journal of Law Information and Science* 21, no. 2 (2011).
[31] Robert Sparrow, "Twenty Seconds to Comply: Autonomous Weapon Systems and the Recognition of Surrender," *International Law Studies* 91 (2015).
[32] That importance of this requirement is noted in Marchant et al., "International Governance of Autonomous Military Robots," 282.
[33] Arkin, *Governing Lethal Behavior in Autonomous Robots*.
[34] For an account of what would be required to produce "ethical robots", which is more sympathetic to the idea than I am here, see Wendell Wallach and Colin Allen, *Moral Machines: Teaching Robots Right from Wrong* (Oxford: Oxford University Press, 2009).
[35] Raimond Gaita, *Good and Evil: An Absolute Conception*, 2nd ed. (Abingdon: Routledge, 2004), 264-82.
[36] Arkin, *Governing Lethal Behavior in Autonomous Robots*, 203-09.
[37] Donald P Brutzman et al., "Run-Time Ethics Checking for Autonomous Unmanned Vehicles: Developing a Practical Approach" (paper presented at the *18th International Symposium on Unmanned Untethered Submersible Technology (UUST)*, Portsmouth, New Hampshire, 2013); Alex Leveringhaus and Tjerk de Greef, "Keeping the Human 'in-the-Loop': A Qualified Defence of Autonomous Weapons," in *Precision Strike Technology and International Intervention: Strategic, Ethico-Legal and Decisional Implications*, eds. Mike Aaronson, et al. (Abingdon; New York: Routledge, 2015).
[38] Adams, "Future Warfare and the Decline of Human Decision Making."
[39] Anderson and Waxman, "Law and Ethics for Autonomous Weapon Systems," 7; Schmitt and Thurnher, "Out of the Loop," 238; Work and Brimley, *20YY*, 24.
[40] See also Brutzman et al., "Run-Time Ethics Checking for Autonomous Unmanned Vehicles." This is not to deny that there would be some military advantages associated with the development of such systems, as long as the communications infrastructure necessary to allow contact with a human operator as required was in place. For instance, by removing the need for direct human supervision it would multiply the number of systems that could operate in the context of a given amount of band width and also make it possible for one human operator to oversee the activities of a number of robots.
[41] Guarini and Bello, "Robotic Warfare"; Human Rights Watch, *Losing Humanity*; Sharkey, "The Evitability of Autonomous Robot Warfare." Indeed, I have argued this myself elsewhere. See Robert Sparrow, "Robotic Weapons and the Future of War," in *New Wars and New Soldiers: Military Ethics in the Contemporary World*, eds. Jessica Wolfendale and Paolo Tripodi (Surrey, UK; Burlington, VA: Ashgate, 2011).
[42] Michael N. Schmitt, "Autonomous Weapon Systems and International Humanitarian Law: A Reply to the Critics," *Harvard National Security Journal* (2013).
[43] Brutzman et al., "Run-Time Ethics Checking for Autonomous Unmanned Vehicles"
[44] Guarini and Bello, "Robotic Warfare."
[45] Leveringhaus and de Greef, "Keeping the Human 'in-the-Loop'."
[46] To the extent that this requirement is not recognised by LOAC, the *legal* barriers to the ethical use of AWS in sufficiently restricted domains will be correspondingly lower.
[47] Ronald C. Arkin, "Lethal Autonomous Systems and the Plight of the Non-Combatant," *AISB Quarterly* 137 (2013).
[48] Arkin, "The Case for Ethical Autonomy in Unmanned Systems."
[49] The argument in this paragraph owes much to remarks made by Daniel Brunstetter in a session on drone warfare at the International Studies Association conference in San Francisco in April 2013. See also Megan Braun and Daniel R. Brunstetter, "Rethinking the Criterion for Assessing CIA-Targeted Killings: Drones, Proportionality and Jus Ad Vim," *Journal of Military Ethics* 12, no. 4 (2013). George Lucas subsequently brought it to my attention that he had in fact rehearsed this argument in a paper in 2011. See George R. Lucas Jr, "Industrial Challenges of Military Robotics," *Journal of Military Ethics* 10, no. 4 (2011).
[50] Lucas Jr, "Industrial Challenges of Military Robotics."

[51] Brian G. Williams, *Predators: The CIA's Drone War on Al Qaeda* (Washington, DC: Potomac Books, 2013).
[52] Williams' argument proceeds by providing evidence of actual consent but in most cases the argument will need to proceed by way of reference to "hypothetical consent" — that is, what civilians in the area of operations *would* prefer.
[53] Robert Sparrow, ""Hands up Who Wants to Die?": Primoratz on Responsibility and Civilian Immunity in Wartime," *Ethical Theory and Moral Practice* 8, no. 3 (2005).
[54] Human Rights Watch, *Losing Humanity*; International Human Rights and Conflict Resolution Clinic at Stanford Law School and Global Justice Clinic at NYU School of Law. *Living Under Drones: Death, Injury, and Trauma to Civilians From US Drone Practices in Pakistan*. 2012. http://www.livingunderdrones.org/download-report/.
[55] Sparrow, "Robotic Weapons and the Future of War."
[56] Singer, *Wired for War*, 319; Sparrow, "Predators or Plowshares?"
[57] Paul W. Kahn, "The Paradox of Riskless Warfare," *Philosophy & Public Policy Quarterly* 22, no. 3 (2002).
[58] Bradley J. Strawser, "Moral Predators: The Duty to Employ Uninhabited Aerial Vehicles," *Journal of Military Ethics* 9, no. 4 (2010).
[59] Thomas Nagel, "War and Massacre," *Philosophy and Public Affairs* 1 (1972).
[60] Nagel, "War and Massacre," 136.
[61] Nagel, "War and Massacre," 138.
[62] Peter Asaro, "On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-Making," *International Review of the Red Cross* 94, no. 886 (2012): 701.
[63] Asaro, "On Banning Autonomous Weapon Systems."
[64] Mary Ellen O'Connell, "Banning Autonomous Killing: The Legal and Ethical Requirement That Humans Make Near-Time Lethal Decisions," in *The American Way of Bombing: Changing Ethical and Legal Norms, from Flying Fortresses to Drones*, eds. Matthew Evangelista and Henry Shue (Ithaca, NY: Cornell University Press, 2014).
[65] Arkin, "Lethal Autonomous Systems and the Plight of the Non-Combatant"; V Kanwar, "Post-Human Humanitarian Law: The Law of War in the Age of Robotic Weapons," *Harvard National Security Journal* 2, no. 2 (2011): 619-20; Schmitt and Thurnher, "Out of the Loop," 268.
[66] They should, of course, be appropriately confident that the persons at the facility or location they are attacking are legitimate targets.
[67] Indeed, the stockpiling and use of anti-personnel mines at least is prohibited by the Ottawa treaty.
[68] Defense Science Board, "Role of Autonomy," 1-2.
[69] Sparrow, "Killer Robots."
[70] Roff, "Killing in War."
[71] Matthias, "The Responsibility Gap."
[72] Thomas Hellström, "On the Moral Responsibility of Military Robots," *Ethics and Information Technology* 15, no. 2 (2013); Leveringhaus and de Greef, "Keeping the Human 'in-the-Loop'"; Gert-Jan Lokhorst and Jeroen Van Den Hoven, "Responsibility for Military Robots," in *Robot Ethics: The Ethical and Social Implications of Robotics*, ed. Patrick Lin, Keith Abney, and George A. Bekey (Cambridge, MA
London: MIT Press, 2012); Sparrow, "Killer Robots."
[73] Nagel, "War and Massacre," 135, note 7.
[74] Charli Carpenter, "Us Public Opinion on Autonomous Weapons," (University of Massachusetts 2013), http://www.duckofminerva.com/wp-content/uploads/2013/06/UMass-Survey_Public-Opinion-on-Autonomous-Weapons.pdf
[75] Gubrud, "Stopping Killer Robots," 40; Aaron M. Johnson and Sidney Axinn, "The Morality of Autonomous Robots," *Journal of Military Ethics* 12, no. 2 (2013); Sparrow, "Robotic Weapons and the Future of War."
[76] Asaro, "On Banning Autonomous Weapon Systems."
[77] Schmitt and Thurnher, "Out of the Loop," 10.
[78] Gubrud, "Stopping Killer Robots; Human Rights Watch, *Losing Humanity*; Wendell Wallach, "Terminating the Terminator: What to Do About Autonomous Weapons," *Science Progress* (2013). The legal basis for doing so might be found in the "Marten's clause" in the Hague Convention and the prohibition on weapons which are "contrary to the dictates of the public conscience." Human Rights Watch, "Losing Humanity."

[79] Arkin, “Lethal Autonomous Systems and the Plight of the Non-Combatant”; Anderson and Waxman, “Law and Ethics for Robot Soldiers”; Anderson and Waxman, “Law and Ethics for Autonomous Weapon Systems”; Schmitt and Thurnher, “Out of the Loop.”
[80] The moral status of nuclear weapons remains controversial in some quarters. However, the last two decades of wars justified with reference to states’ possession of “weapons of mass destruction” suggests that there is an emerging international consensus that such weapons are *mala in se*.
[81] Johnson and Axinn, “The Morality of Autonomous Robots,” 137.
[82] Should this campaign fail, it is possible that public revulsion at sending robots to kill people will be eroded as AWS come into use and become a familiar feature of war — as has occurred with a number of weapons, including artillery and submarines, in the past. In that case, the argument that such killing disrespects the humanity of our enemies will eventually lapse as the social conventions around respect for the humanity of combatants are transformed. It might be argued that even if a prohibition on AWS is achieved, conventional understandings of the appropriate relations between humans and robots may shift in the future as people become more familiar with robots in civilian life. While this cannot be ruled out a priori, I suspect that it is more likely that outrage at robots being allowed to kill humans will only intensify as a result of the social and psychological incentives to maintain the distinction between “us” and “them”.
[83] Arkin, “Lethal Autonomous Systems and the Plight of the Non-Combatant.”
[84] For cynicism about the prospect of such, see: Anderson and Waxman, “Law and Ethics for Autonomous Weapon Systems”; Arkin, “Lethal Autonomous Systems and the Plight of the Non-Combatant”; Marchant et al., “ International Governance of Autonomous Military Robots.” For a countervailing perspective, see O'Connell, “Banning Autonomous Killing.”
[85] Wendell Wallach and Colin Allen, “Framing Robot Arms Control,” *Ethics and Information Technology* 15, no. 2 (2013); Schmitt and Thurnher, “Out of the Loop.”
[86] For an important contribution to this project, see Jürgen Altmann, “Arms Control for Armed Uninhabited Vehicles: An Ethical Issue,” *Ethics and Information Technology* 15, no. 2 (2013).
[87] Thanks are due to Ron Arkin, Daniel Brunstetter, Ryan Jenkins, Mark Gubrud, Duncan Purves, Heather Roff, BJ Strawser, Michael Schmitt, Ryan Tonkens, several anonymous referees, and the editors of *Ethics and International Affairs* for comments on drafts of this manuscript. Mark Howard ably assisted me with sources and with preparing the paper for publication.