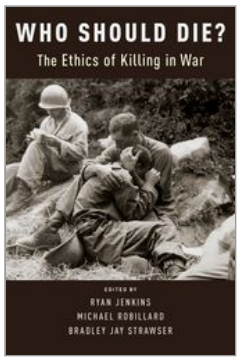
University Press Scholarship Online
Oxford Scholarship Online



Who Should Die?: The Ethics of Killing in War
Bradley Jay Strawser, Ryan Jenkins, and Michael Robillard

Print publication date: 2018
Print ISBN-13: 9780190495657
Published to Oxford Scholarship Online: November 2017
DOI: 10.1093/oso/9780190495657.001.0001

What Is the Moral Problem with Killer Robots?

Susanne Burri

DOI:10.1093/oso/9780190495657.003.0009

Abstract and Keywords

An autonomous weapon system (AWS) is a weapons system that, “once activated, can select and engage targets without further intervention by a human operator” (US Department of Defense directive 3000.09, November 21, 2012). Militaries around the world are investing substantial amounts of money and effort into the development of AWS. But the technology has its vocal opponents, too. This chapter argues against the idea that a targeting decision made by an AWS is always morally flawed simply because it is a targeting decision made by an AWS. It scrutinizes four arguments in favor of this idea and argues that none of them is convincing. It also presents an argument in favor of developing autonomous weapons technology further. The aim of this chapter is to dispel one worry about AWS, to keep this worry from drawing attention away from the genuinely important issues that AWS give rise to.

Keywords: autonomous weapon system, killer robot, artificial intelligence, war, ethics

Introduction

PRINTED FROM OXFORD SCHOLARSHIP ONLINE (www.oxfordscholarship.com). (c) Copyright Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a monograph in OSO for personal use (for details see http://www.oxfordscholarship.com/page/privacy-policy). Subscriber: Universitats- und Stadtbibliothek Köln; date: 11 June 2018

An autonomous weapon system (AWS) is a weapon system that, “once activated, can select and engage targets without further intervention by a human operator.”[1] Militaries around the world are intrigued by the potential that autonomous weapons technology offers, and they are investing substantial amounts of money and effort into the development of increasingly sophisticated AWS. But the technology has its vocal opponents, too. Thousands of artificial intelligence and robotics researchers have recently called for a ban on the further development of “offensive autonomous weapons beyond meaningful human control.”[2] The researchers’ demand is backed not only by a multitude of nongovernmental organizations[3] but—empirical research suggests—by a majority of the general public as well.[4] A fair number of philosophers, political scientists, and legal scholars have shown themselves critical of AWS also.[5]

(p.164) To some extent, people feel uneasy about AWS because they think that the technology comes with a huge potential for abuse. Unlike nuclear weapons, so the thought goes, AWS would one day be widely available if the technology were developed further. It would therefore be only a matter of time before some crazy or evil person obtained autonomous weapons and used them to wreak havoc.

But people worry about AWS not only because the technology could easily be abused. An often-voiced concern is that even if autonomous weapons were to remain exclusively in the hands of well-intentioned individuals, the technology might still malfunction in unexpected and possibly disastrous ways. In a recent paper, Robert Sparrow defines an autonomous weapon as a weapon that, among other things, “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 definition has the risk of malfunction built right into it.[6]

I believe that worries about the abuse of AWS and about their risk of malfunction are well grounded and, accordingly, deserve to be taken seriously. When thinking about what regulatory framework we should implement with respect to the further development and use of AWS, this means that we should carefully consider how optimally to respond to these worries.

But there is also a third type of AWS-related worry that appears to me less founded and that seems to draw attention away from the morally genuinely important issues that AWS give rise to. According to this third type of worry, a final targeting decision made by an AWS is always morally flawed in at least one respect, simply because it is a targeting

PRINTED FROM OXFORD SCHOLARSHIP ONLINE (www.oxfordscholarship.com). (c) Copyright Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a monograph in OSO for personal use (for details see http://www.oxfordscholarship.com/page/privacy-policy). Subscriber: Universitats- und Stadtbibliothek Köln; date: 11 June 2018

decision made by an AWS as opposed to a human being. Those who are moved by this type of worry are wary of AWS **(p.165)** because they feel that "allowing life or death decisions to be made by machines crosses a fundamental moral line."[7]

In this chapter, I argue against the idea that allowing life-or-death decisions to be made by machines is necessarily morally problematic in at least one respect. More specifically, I scrutinize four prima facie powerful arguments in favor of this idea, and I argue that none of them is ultimately convincing. I also present an argument in favor of developing autonomous weapons technology further. Importantly, the aim of this chapter is not to present a defense of autonomous weapons but simply to show that a particular type of worry about AWS is misguided and to point out that at least some moral considerations speak in favor of developing the technology further. In this way, I hope to help sharpen the philosophical discussion of the moral issues that AWS give rise to.

The remainder of this chapter proceeds as follows. In the following section, I clarify important concepts and terminology. Then, I critically evaluate four defenses of the claim that a final targeting decision made by an AWS is always morally problematic in at least one respect, and I argue that they all lack force. Finally, I explain how the duty to protect just combatants speaks in favor of developing autonomous weapons technology further.

Terminology

Autonomous Weapon Systems

International humanitarian law currently lacks a definition of AWS. In a 2012 directive, the US Department of Defense defines the term as follows:

> Autonomous weapon system. A weapon system that, once activated, can select and engage targets without further intervention by a human operator. This includes human-supervised autonomous weapon systems that are designed to allow human operators to override operation of the weapon system, but can select and engage targets without further human input after activation.[8]

Two things are noteworthy about this definition. First, the "autonomy" that an AWS necessarily possesses is technical in nature and very narrow: a weapon system[9] counts as autonomous as soon as it has the ability—once it has been **(p.166)** activated—to perform the two tasks

PRINTED FROM OXFORD SCHOLARSHIP ONLINE (www.oxfordscholarship.com). (c) Copyright Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a monograph in OSO for personal use (for details see http://www.oxfordscholarship.com/page/privacy-policy). Subscriber: Universitats- und Stadtbibliothek Köln; date: 11 June 2018

of selecting and engaging a target without human intervention. Second, the definition's focus is clearly on the system's technical capacity to run in autonomous mode. Whether the system is intended to run autonomously by default is irrelevant to its status as an AWS.

The difference between autonomous and other weapon systems can also be described in terms of the role that a human operator of the system has with respect to final targeting decisions.[10] An operator can be in the loop, in which case it is necessarily the operator who decides which specific targets are to be engaged: no target can or will be engaged unless the operator directs the system to do so. Weapon systems with a human operator in the loop are not autonomous.

If a human operator is in the loop with respect to final targeting decisions, then the operator does not decide which specific targets are to be engaged but can monitor the weapon system's performance and intervene to prevent it from engaging a target if he or she does not agree with the target the machine has selected. Lastly, a human controller is out of the loop if he or she does not get to decide which specific targets should be engaged or if he or she can neither continuously monitor the system's performance nor halt it at will. Weapon systems where a human operator is either in the loop or out of the loop are autonomous in nature.

In some sense, AWSs are not a thing of the future. Landmines have been in use for centuries, and at least in a crude sense, they select and engage their own targets once a human operator has activated them. There are also other, much more sophisticated AWSs that are already operational. Consider the Israeli Harpy.[11] The Harpy is a wide-area loitering cruise missile that targets enemy radars. Deployed without a specific target in mind, it flies a search pattern over a wide area, looking for enemy radars. If it detects an object that it identifies as an enemy radar, it dive-bombs into the object with the goal of destroying it.

Landmines and weapon systems such as the Harpy are not the kind of AWS that those lobbying for a pre-emptive ban on the further development of the technology are particularly worried about. Whereas landmines seem technically too simplistic to give rise to the particular issues that advanced AWSs are thought to give rise to, the Harpy is not an especially worrisome AWS because it is purely antimachinery: it does not target human beings. The vaguely expressed concern of AWS critics is that no machine should be tasked with life-and-death **(p.167)**

PRINTED FROM OXFORD SCHOLARSHIP ONLINE (www.oxfordscholarship.com). (c) Copyright Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a monograph in OSO for personal use (for details see http://www.oxfordscholarship.com/page/privacy-policy). Subscriber: Universitats- und Stadtbibliothek Köln; date: 11 June 2018

decisions.[12] I take it that AWS critics are therefore primarily opposed to (1) technically sophisticated AWSs that (2) target human beings.

Lethal Autonomous Robots

Lethal autonomous robots (LARs), or killer robots, are a proper subset of AWSs. They are technically sophisticated AWSs that are also robots and that are able to lethally target human beings. As AWS critics are first and foremost worried about LARs, my primary concern in this chapter will be with LARs also.

A robot is a programmable machine that is equipped with sensors that allow it to form a representation of its environment.[13] At least to some extent, a robot has to be able to interact with or manipulate its environment without human intervention. An LAR, or killer robot, is a robot that, once it has been activated, can select and lethally engage human targets without the intervention of a human operator.

Importantly, drones as they are in use today are not killer robots as they are all remotely piloted (hence their alternative name of remotely piloted aircraft). It seems that no killer robots are currently in use anywhere.

Machine Learning

Robots run on computer programs. A computer program is a set of instructions that a computer can process and that allows it to perform certain tasks. In the case of robots, this includes processing sensory input and reacting to it in line with predefined objectives.

Algorithms form a key part of any computer program. An algorithm is a set of rules. Standard or nonlearning algorithms spell out how to perform a certain task in a stepwise manner. If I tell you to (1) wait in front of the next zebra crossing you find, (2) look left and look right, (3) cross the street if, and only if, you see no oncoming traffic from either side, I am providing you with a simple algorithm for safely crossing a street. All robots rely on nonlearning algorithms, but some rely on machine learning algorithms as well. If a robot relies entirely on nonlearning algorithms, it performs its core tasks—playing chess, assembling car parts, or selecting and engaging a target—according to a fixed procedure that was written out for it by human programmer. **(p. 168)** A robot that partly relies on learning algorithms has the ability to use data to modify the procedure by which it performs its core tasks. Depending on the degree of sophistication, learning algorithms enable a machine to sort through potentially very large, unstructured data sets

PRINTED FROM OXFORD SCHOLARSHIP ONLINE (www.oxfordscholarship.com). (c) Copyright Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a monograph in OSO for personal use (for details see http://www.oxfordscholarship.com/page/privacy-policy). Subscriber: Universitats- und Stadtbibliothek Köln; date: 11 June 2018

and to extract from these data sets information that allows it to improve the rules it follows to perform its core tasks.

As an example, consider image recognition software, where learning algorithms have recently been put to use with considerable success. In this case, researchers improved the image recognition capabilities of their learning machine by presenting it with picture after picture, as well as with a description of what each picture displayed. The machine sorted through all of these data to extract patterns, distinctive elements, and the like. It then used this information to overhaul its basic classification mechanism.

On the plus side, learning robots have the potential to become much more skilled and flexible at performing their core tasks than their rigidly programmed nonlearning counterparts.[14] The rapid progress in artificial intelligence that we have witnessed over the past few years, and continue to witness today, is largely driven by the powerful and versatile methods of machine learning. But the potential of machine learning comes at a cost. Some of the information that learning machines extract from the data they are fed tends to bear on their core procedures in ways that we can't fully grasp. To put the same point differently, once a learning machine has been exposed to a sufficient amount of data, there is a good chance that it will perform its core tasks based on modified rules that we are no longer able to completely comprehend. Once this is the case, there is always the risk that the machine does something undesirable that takes us by surprise.[15] The popular science writer David Berreby puts the point as follows:

> [N]eural nets[16] sometimes come out with answers that are downright weird: not right, but also not wrong in a way that people can grasp. Instead, the answers sound like something an extraterrestrial might come up with. These oddball results are rare. But they aren't just random glitches . . . . [This] can be a troubling thought, even if you aren't yet depending on neural nets to run your home and drive you around. After all, the more we rely on artificial intelligence, the more we need it to be predictable, especially in failure. Not knowing how or why a **(p. 169)** machine did something strange leaves us unable to make sure it doesn't happen again.[17]

This quote illustrates that, insofar as we are talking about learning LARs, worries that they might malfunction in unpredictable ways are well grounded.

A First Argument Against LARs: The Anti-Codifiability Thesis

PRINTED FROM OXFORD SCHOLARSHIP ONLINE (www.oxfordscholarship.com). (c) Copyright Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a monograph in OSO for personal use (for details see http://www.oxfordscholarship.com/page/privacy-policy). Subscriber: Universitats- und Stadtbibliothek Köln; date: 11 June 2018

A first argument in favor of the view that a killing performed by an LAR is always morally problematic in at least one respect is brought forward by Duncan Purves, Ryan Jenkins, and Bradley J. Strawser. They claim that proper moral deliberation requires Aristotelian judgment or that correctly applying a moral principle to a specific situation can never be done purely mechanically; it always requires interpretation. They call this the *anti-codifiability thesis*.[18]

Consider the *in bello* principle of necessity,[19] which states that you are morally required to minimize the harm that you inflict in order to achieve a morally worthy aim. As a commanding officer who is considering whether ordering a particular mission would satisfy the principle of necessity, you need to carefully think through all of your options for action. Once you have a good grasp of the expected effects of the mission in question, you still need to compare these effects to the expected effects of whatever alternatives are as a matter of fact available to you. This requires situational knowledge, knowledge about the means available to you, foresight, imagination, and creativity. Quite plausibly, it is not something you can do by unthinkingly running through a set of predefined steps. Purves et al.'s argument then is the following: LARs run on algorithms, which means that they unthinkingly run through a set of predefined steps. It follows that they are incapable of proper moral deliberation and that they are therefore always at risk of making significant moral mistakes.[20] It is morally objectionable to deploy LARs if they are at risk of making significant moral mistakes.

(p.170) Even if the anti-codifiability thesis is true and even if it is true that robots will therefore never be capable of full-blown moral deliberation, it does not follow that there would always be something morally objectionable about their deployment. In short, LARs don't have to be morally sophisticated deliberators to almost exclusively inflict only permissible harm. It suffices, instead, that a conscientious human commanding officer deploys them only in contexts where they are able to identify sufficient conditions for the morally permissible infliction of lethal harm. For LARs to be usefully and permissibly employable, they don't have to be able to replace human soldiers across all possible circumstances, nor do they have to be able to strategize and reason about entire missions the way higher-ranking military personnel have to.

The following two examples illustrate how LARs could usefully and permissibly be employable even with limited deliberative capacities. Following Ronald C. Arkin,[21] the general idea is that robots could be programmed conservatively: as they lack a drive for self-preservation, they could effectively be instructed to desist from selecting and

PRINTED FROM OXFORD SCHOLARSHIP ONLINE (www.oxfordscholarship.com). (c) Copyright Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a monograph in OSO for personal use (for details see http://www.oxfordscholarship.com/page/privacy-policy). Subscriber: Universitats- und Stadtbibliothek Köln; date: 11 June 2018

engaging human targets in all but the morally most clear-cut cases. First, a robot could be programmed to engage a human target if, and only if, (1) the target has been shooting at our own combatants at some time during, say, the past ten seconds and (2) our own combatants have not indicated to the robot to refrain from firing. As enemy combatants are generally liable to be killed if we are fighting for a just cause,[22] responsibly deployed LARs that are programmed in this way should be able to highly reliably engage only in permissible killings. Second, a robot could be programmed to engage a human target if, and only if, its advanced facial recognition algorithms allow it to predict with near certainty that the human target in question is on a "kill list." On the assumption that the kill list is drawn up by morally responsible human beings, an LAR that is programmed in this way should again be able to highly reliably perform only permissible killings.

Against this, it could be argued that even if LARs are deployed only in limited contexts, there is nevertheless an ineliminable risk that they will unexpectedly display undesirable behavior,[23] at least if we are dealing with learning LARs. This seems to me correct. But the same risk is present when we are dealing with human soldiers also; their being capable of engaging in full-blown moral deliberation does not guard against this risk. Hence, what makes it permissible to **(p.171)** deploy learning LARs on a just mission would seem to be the same thing that makes it permissible to deploy human soldiers also: roughly, that due care has been taken to ensure that the risk in question is sufficiently small.[24]

Purves et al.[25] also argue that if the anti-codifiability thesis is true, an LAR would never be able to "resist an immoral order in the way that a human soldier might." If we take this as an argument in favor of the idea that there is always something morally problematic about deploying an LAR, the thought would seem to be that it is always better to deploy a human soldier instead as only human soldiers are able to recognize and resist immoral orders. But this argument is flawed in several respects. First, a robot doesn't have to be capable of moral deliberation to be able to resist at least some immoral orders. LARs could, for example, be programmed in such a way that they never target what they identify as a child. Second, to the extent that LARs are unable to resist immoral orders because they simply act as instructed, they are similarly unable to choose to deviate from moral orders. Hence, the argument of Purves et al. can also be turned on its head as it may always be better for a morally motivated commanding officer to deploy an obedient LAR as opposed to a potentially disobedient human soldier who might choose to ignore a moral order and engage in moral wrongdoing instead. Finally, if it is the point of Purves et al.'s argument

PRINTED FROM OXFORD SCHOLARSHIP ONLINE (www.oxfordscholarship.com). (c) Copyright Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a monograph in OSO for personal use (for details see http://www.oxfordscholarship.com/page/privacy-policy). Subscriber: Universitats- und Stadtbibliothek Köln; date: 11 June 2018

that there should always be a check on a commanding officer's orders, this condition could be satisfied by implementing a four-eyes principle with respect to missions that involve the deployment of LARs so that such missions would always need the approval of at least two human beings before being embarked on.

In sum, as long as LARs are carefully deployed by morally motivated human officers, there doesn't have to be anything morally worrisome about their deployment even if the anti-codifiability thesis is true.

A Second Argument Against LARs: Acting for the Right Reasons

For the sake of argument, Purves et al. are willing to concede that LARs could someday become at least as skilled at moral deliberation as the most virtuous human beings. More specifically, they present a second argument against killer robots which aims to establish that there would be something morally objectionable about a killing performed by a robot even if that robot was known to **(p.172)** be "perfect at making moral decisions" (p. 860 ff.). Their argument runs as follows. First, they put forward the claim that it is impossible for a robot to act for a reason. Their idea is that possessing "an attitude of belief or desire (or some further propositional attitude)" is a conceptual prerequisite of "acting for a reason" and that something which runs on algorithms cannot possess such an attitude (p. 861). Second, they contend that whenever an action lacks the feature of being performed for the right reason, it is morally deficient in at least one respect. They present the case of a "sociopathic soldier" who is "completely unmoved by the harm he causes to other people" and who takes pleasure in killing "scores of enemy soldiers" because he loves following orders (p. 861). According to Purves et al., even if this soldier is never ordered to kill anyone but liable enemy combatants, what he does is nevertheless morally flawed in at least one significant respect. Other things equal, it would be morally preferable to replace the sociopathic soldier with a morally virtuous soldier who understands that while killing enemy combatants is always regrettable, it is nevertheless morally permissible when a goal of sufficient moral value cannot be achieved in a less harmful way.

I believe that Purves et al. are right about the sociopathic soldier. But it seems to me that all the reasons we have to insist that it matters, morally speaking, whether a human person is acting for the right reasons fall away when we are dealing with an LAR that, by assumption, is "perfect at making moral decisions" (p. 860).

Purves et al. do not try to explain why an action is morally flawed when it lacks the feature of having been performed for the right reason. As I see it, there are four different, though to some extent complementary,

PRINTED FROM OXFORD SCHOLARSHIP ONLINE (www.oxfordscholarship.com). (c) Copyright Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a monograph in OSO for personal use (for details see http://www.oxfordscholarship.com/page/privacy-policy). Subscriber: Universitats- und Stadtbibliothek Köln; date: 11 June 2018

explanations for why we might appropriately care about the reasons behind an action when the action is performed by a human being.[26] But none of these explanations would seem to apply in the context of robot action.

First, when a human agent acts in accordance with what morality demands but does so for the wrong reasons (or for no reason at all), this seems morally significant because it indicates that the agent followed the demands of morality only coincidentally. In slightly different circumstances, he or she might have performed a morally reprehensible act instead. Consider Purves et al.'s sociopathic soldier. As it happens, the soldier never receives a morally illegitimate order, so he never violates the *in bello* rules. But it is nevertheless true that the soldier would have executed the morally most outrageous orders had he received them. **(p.173)** When we are dealing with a human agent, we therefore want him or her to act in accordance with what morality demands for the right reasons as this is one of the best available assurances that his or her behavior will robustly track the demands of morality.[27] By contrast, when it comes to robots, Purves et al. are committed to the claim that there is no such close relationship between acting for the right reasons and robustly tracking the demands of morality. Recall that Purves et al. assume that a robot is not the kind of thing that can act for a reason, but they nevertheless grant that a robot could in principle be built in such a way that it would always settle for the morally best alternative. Presumably their idea is that a robot could unerringly settle for the morally best alternative if it ran on sufficiently sophisticated algorithms.

A second explanation for why we might be justified in caring about a human agent's reasons for action centers on the idea of respect. When it is morally permissible to act in a way that will predictably set back others' vital interests, there is something disrespectful about acting in this way for a morally inappropriate reason.

Consider again the sociopathic soldier. When he sets out to kill an enemy combatant without any recognition that the enemy combatant's death is regrettable and without any appreciation of the morally valuable goal that justifies killing the enemy combatant, the sociopathic soldier's attitude toward the enemy combatant is quite plausibly one of reprehensible disrespect. But when an agent is not the kind of agent that acts on reasons, the question of appropriate attitude simply does not arise. There is nothing unduly selfish or inconsiderate about a lion that hunts down its prey, just as there is nothing disrespectful or

PRINTED FROM OXFORD SCHOLARSHIP ONLINE (www.oxfordscholarship.com). (c) Copyright Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a monograph in OSO for personal use (for details see http://www.oxfordscholarship.com/page/privacy-policy). Subscriber: Universitats- und Stadtbibliothek Köln; date: 11 June 2018

insensitive about an LAR that has decided to lethally engage an enemy combatant based on the algorithms that were programmed into it.

Against this, it could be argued that a killing performed by an LAR is disrespectful not because the LAR's attitude is disrespectful but because there is something disrespectful about the attitude of whoever decides to deploy the LAR. For many people, there is something deeply unsettling about the thought of finding themselves face to face with an enemy that they know to be unresponsive to reasons, and they would much prefer to be confronted with an enemy who is capable of acting on reasons instead.[28] Hence, there may be something inconsiderate and disrespectful about a decision to utilize LARs. While an argument along these lines is brought forward by Sparrow, it seems to me successfully countered by Jenkins and Purves **(p.174)** (2016).[29] Jenkins and Purves (p. 9) argue that a decision to deploy LARs is respectful "in the narrow sense" if it is taken on the reasonable assumption that only liable targets will be engaged and that the decision is respectful "in the wide sense" if it is taken at least to some extent with a view to minimizing civilian casualties.[30] In other words, what would seem to make the deployment of LARs respectful is deploying them in such a way that their behavior can reasonably be expected to conform to the rules of *jus in bello*.

Third, we might insist that actions not performed for the right reasons are morally deficient because they lack moral worth. According to this Kantian idea, actions that are performed for the right reasons are accorded a special moral status—unlike other actions, they have moral worth—because the will behind them is of unconditional moral value.[31]

In my view, it makes sense to call an action morally deficient because it was not performed for the right reasons as long as the action was performed by an agent who is capable of acting on reasons. But if the action was performed by an agent who lacks the capacity to act on reasons, calling it morally deficient because it was not performed for the right reasons seems to me analogous to calling a bicycle deficient because it does not run on four sturdy wheels. It is, of course, possible to argue that the realization of moral worth is morally desirable so that, other things equal, it is preferable to have an action performed for the right reasons as opposed to having it performed by an agent who is incapable of acting on reasons.[32] But even if this is true, it does not imply that an action performed by an agent who is incapable of acting on reasons is thereby necessarily morally deficient. Nor does it imply that there exists a morally weighty reason in favor of employing human soldiers over LARs: if human soldiers frequently act on the wrong

PRINTED FROM OXFORD SCHOLARSHIP ONLINE (www.oxfordscholarship.com). (c) Copyright Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a monograph in OSO for personal use (for details see http://www.oxfordscholarship.com/page/privacy-policy). Subscriber: Universitats- und Stadtbibliothek Köln; date: 11 June 2018

reasons or are morally motivated but make empirical mistakes, the disvalue involved in this may well outweigh the moral worth involved in their acting on the right reasons.

Fourth and finally, when a human agent fails to act for the right reasons, we can meaningfully describe his or her conduct as the morally flawed product of a less than fully virtuous character. According to this Aristotelian idea, a person is able to deliberate and act well to the extent that he or she is virtuous: virtues enable one to recognize and balance what is morally at stake in whatever situation **(p.175)** that one finds oneself in.[33] On this way of looking at things, each killing performed by Purves et al.'s sociopathic soldier is a morally flawed product of his depraved character.

While it seems reasonable to describe the sociopathic soldier's actions in this way, it would be absurd to characterize a machine's consistent failure to act for the right reasons as the inevitable consequence of its corrupt character. If Purves et al. are right that machines do not act for the right reasons because they do not act for reasons, it seems much more apt to say that a machine's actions do not tell us anything about its—possibly nonexistent—character.

In sum, Purves et al.'s second argument against LARs is not convincing as it stands. While it makes sense to assume that it matters, morally speaking, whether human actions are performed for the right reasons, the same does not seem to hold true of robot actions.

A Third Argument Against LARs: The Responsibility Gap

For Rob Sparrow, the deployment of killer robots is morally problematic because it creates a potential responsibility gap:[34] as long as LARs are in use, there is always a risk that they will inflict wrongful harm for which no one is morally responsible. According to Sparrow (p. 67),

> [it] is a minimal expression of respect due to our enemy . . . that someone should accept responsibility, or be capable of being held responsible, for the decision to take their life. If we fail in this, we treat our enemy like vermin, as though they may be exterminated without moral regard at all. The least we owe our enemies is allowing that their lives are of sufficient worth that someone should accept responsibility for their deaths.

Sparrow contends that a responsibility gap arises when circumstances are such that no human being is morally responsible for the wrongful harm inflicted **(p.176)** by an LAR. Sparrow (pp. 71–73) thinks that LARs themselves are never morally responsible for the harm that they

PRINTED FROM OXFORD SCHOLARSHIP ONLINE (www.oxfordscholarship.com). (c) Copyright Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a monograph in OSO for personal use (for details see http://www.oxfordscholarship.com/page/privacy-policy). Subscriber: Universitats- und Stadtbibliothek Köln; date: 11 June 2018

inflict, essentially because it is conceptually impossible to punish a machine. Let us take a closer look at these two ideas in turn.

Human Responsibility for Harm Inflicted by LARs

While the relevant passages are somewhat obscure, I take Sparrow (pp. 69–71) to be saying that a human agent is not morally responsible for harm inflicted by an LAR when the harm was not, in some meaningful sense, under the human agent's control. Moreover, harm inflicted by an LAR was not meaningfully under a human agent's control if either

1. The machine behaved in a way that was not foreseeable or
2. The machine in some substantive sense chose to inflict the harm, for example, against an explicit order to desist.

As I see it, the first of these two conditions cannot absolve human agents who had control over the development and deployment of an LAR from their moral responsibility for wrongful harm inflicted by that LAR.

By way of example, consider a programmer who writes a computer program for an LAR that contains a large number of learning algorithms. The basic idea behind the program is to enable an LAR to become increasingly skilled at searching for and lethally engaging human targets whose biometrics it has successfully matched to a record in some database. When put into LAR hardware and tested, the program seems to run smoothly, but it is impossible to predict how the LAR's behavior will evolve once the program's learning algorithms are fed with more and more data. If the programmer—or other individuals working for the same company—decides to hide the fact that the software comes with crucial unpredictabilities, the moral responsibility for any unforeseeable wrongful harm that an LAR running on the software might cause remains with him or her. His or her actions are not only negligent but downright reckless: he or she is pretending that it is relatively safe to use an incredibly dangerous tool.

Alternatively, the relevant individuals might conscientiously inform a potential buyer about the limitations and unpredictabilities of the software. In this case, at least part[35] of the moral responsibility for whatever unforeseeable wrongful **(p.177)** harm an LAR running on the program might cause is passed on to the buyer. If an informed buyer decides to put the program to use without first letting it learn and then testing it extensively in a controlled environment, the informed buyer acts recklessly and is morally blameworthy for whatever wrongful harm the LARs cause. Alternatively, the buyer might test and train the program before putting the LARs to use, but he or

PRINTED FROM OXFORD SCHOLARSHIP ONLINE (www.oxfordscholarship.com). (c) Copyright Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a monograph in OSO for personal use (for details see http://www.oxfordscholarship.com/page/privacy-policy). Subscriber: Universitats- und Stadtbibliothek Köln; date: 11 June 2018

she might, for example, make the mistake of never testing the program's performance under adverse weather conditions. It might then turn out that the LARs start killing indiscriminately once heavy rain starts to fall. If so, the buyer acted negligently and is once again morally blameworthy for the indiscriminate killings of the LARs.

Finally, the buyer might carefully test and train the program before putting the LARs to use, so he or she might eventually be justified in assuming that the LARs are performing reliably. But the LARs might nevertheless malfunction for some outlandish reason that no one could reasonably have foreseen. This would still not leave us with a responsibility gap.[36] When you decide to use what you know is a dangerous tool to further your own purposes, you are morally responsible for the wrongful harm this causes no matter how many precautions you took to avoid causing wrongful harm.[37] If you took all reasonable precautions and if you were justified in assuming that the risk of wrongful harm was sufficiently small, you are not blameworthy for whatever wrongful harm you happen to cause. But you are still morally responsible: depending on the circumstances, you might have to apologize to your victim, you might owe your victim compensation, or it might be morally permissible for a potential victim to kill you to escape the threatening situation that your actions have put him or her in.[38]

In sum, LARs are dangerous tools, and no one who develops or deploys them can hide behind the excuse that the wrongful harm caused by an LAR came as a complete surprise. Taking due care may absolve the relevant human agents of blameworthiness, but it does not absolve them of their moral responsibility. The second condition is more tricky. If I understand Sparrow correctly, he is saying that artificial intelligence will reach a point where it will cease to be valid **(p.178)** to think of an autonomous robot as a mere tool. His idea seems to be that as machines keep learning and become more intelligent, they develop goals and desires of their own. In one of his examples, Sparrow talks about a robot committing a war crime so as to "to revenge the 'deaths' of robot comrades recently destroyed in battle." Elsewhere, he argues that holding "the programmers responsible for the actions of their creation, once it is autonomous, would be analogous to holding parents responsible for the actions of their children once they have left their care."[39]

Nick Bostrom convincingly argues that this anthropomorphizing view of artificial intelligence is false.[40] As machines learn and become increasingly intelligent, they become powerful optimizers: they pursue with skill and ingenuity whatever goals that they were programmed to

PRINTED FROM OXFORD SCHOLARSHIP ONLINE (www.oxfordscholarship.com). (c) Copyright Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a monograph in OSO for personal use (for details see http://www.oxfordscholarship.com/page/privacy-policy). Subscriber: Universitats- und Stadtbibliothek Köln; date: 11 June 2018

pursue. But unless they were programmed to do just that, they don't at some point start developing goals of their own. Put differently, artificial intelligence is by construction purely instrumental intelligence—it is skill at pursuing a set of predefined objectives.

Bostrom calls the idea that a machine could become exceedingly skilled at pursuing a narrow set of predefined objectives the *orthogonality thesis*. In his words, "Intelligence and final goals are orthogonal axes along which possible agents can freely vary. In other words, more or less any level of intelligence could in principle be combined with more or less any final goal" (p. 73). Once we keep in mind that learning algorithms simply enable a machine to sort through data to improve the accuracy of its various decision procedures and classification mechanisms (see earlier discussion under "Machine Learning"), it is difficult to doubt the veracity of Bostrom's orthogonality thesis.

But suppose that Sparrow is right, that the orthogonality thesis is false and that learning LARs might someday reach a point at which they develop goals and possibly even desires of their own. If so, Sparrow seems right to claim that when a wrongful killing was in some substantive sense chosen by the machine itself, there might be no human agent who can justly be held responsible for the killing.[41] But it seems to me that in such a case the moral responsibility would lie with the machine itself.

Robot Responsibility for Harm Inflicted by LARs

Sparrow (2007, pp. 71-3) claims that LARs can never be morally responsible for wrongful harm that they inflict, even if they become highly intelligent, very **(p.179)** powerful, and develop goals and desires of their own.[42] As Sparrow sees it, an agent can be morally responsible only if he or she can be held responsible, and we can hold someone responsible only if it is at least in principle possible to punish him or her. Sparrow then argues that it is conceptually impossible to punish a machine, primarily because punishing someone involves making him or her suffer, while machines are incapable of suffering.

For one thing, I am not convinced that the type of LAR that Sparrow envisages would necessarily be incapable of suffering. Once LARs have goals and desires of their own, why wouldn't they suffer if they had these thwarted?[43] Nor am I entirely convinced that an agent cannot be morally responsible unless it is conceptually possible to hold the agent responsible. Consider the case of a successful suicide bomber who wrongfully blows up ten innocent children along with himself or herself. While it may well be impossible to hold the dead bomber responsible,

PRINTED FROM OXFORD SCHOLARSHIP ONLINE (www.oxfordscholarship.com). (c) Copyright Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a monograph in OSO for personal use (for details see http://www.oxfordscholarship.com/page/privacy-policy). Subscriber: Universitats- und Stadtbibliothek Köln; date: 11 June 2018

he or she nevertheless seems morally responsible and blameworthy for what he or she did.

Sparrow might agree that the suicide bomber is morally responsible in some sense, roughly in the sense that makes him or her blameworthy for what he or she did. But he might go on to argue that this is not the sense of moral responsibility that he is interested in. More precisely, his argument seems to be that there is a worrisome responsibility gap whenever no one can justly be held accountable for some wrongdoing. If this is correct, Sparrow's idea would seem to be that if the suicide bomber acted alone, this would confront us with a responsibility gap similar to the gap that he is worried about with respect to LARs.

At least for the sake of argument, I am willing to grant Sparrow that it is morally worrisome if no one can be held accountable for wrongful harm and that LARs acting on their own behalf would be capable of inflicting such harm.[44] The crucial flaw in Sparrow's argument then seems to be that he wrongly assumes that unless we can punish someone by making him or her suffer, it is impossible to hold him or her to account for his or her wrongdoing.

Suppose that Sparrow is right that LARs cannot suffer and that an LAR commits a war crime against the orders of its commanding officer because it aims "to **(p.180)** revenge the 'deaths' of robot comrades recently destroyed in battle."[45] It seems to me that we could hold such an LAR responsible, for example, by

- Destroying the machine
- Turning off some of its functions or reprogramming it
- Demanding that the machine apologize or provide some form of compensation to the families and dependents of its victims

In short, while it is difficult to know what would be the most sensible way of holding responsible the type of LARs that Sparrow envisages, it is clear that our practices of holding wrongdoers accountable for their actions are not limited to making them suffer.

A Fourth Argument Against LARs: Heartless Killing

According to a fourth argument against the deployment of killer robots, there ought always to be an element of humanity in lethal targeting decisions. As Mary Ellen O'Connell puts it, "[g]iving up the decision [to kill] entirely to a computer program will . . . remove, literally, the humanity that should come to bear in all cases of justifiable killing."[46] Merel Ekelhof and Miriam Struyk express the idea as follows: "War is about human suffering, the loss of human lives, and consequences for

PRINTED FROM OXFORD SCHOLARSHIP ONLINE (www.oxfordscholarship.com). (c) Copyright Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a monograph in OSO for personal use (for details see http://www.oxfordscholarship.com/page/privacy-policy). Subscriber: Universitats- und Stadtbibliothek Köln; date: 11 June 2018

human beings. Killing with machines is the ultimate demoralization of war. Even in the hell of war we find humanity, and that must remain so."[47] As I understand this fourth argument, it proceeds from the assumption that there is something like a morally best way of killing in war. The morally best killings—the most exemplary, the most humane, the most respectful killings—take place when conscientious and empathetic just combatants, feeling the full weight of the decisions before them, eventually and for the right reasons reach the right practical conclusions. The argument against the deployment of LARs is then the following: because there is a morally best way of killing in war, we ought to strive toward it. To deploy LARs, however, is to move away from it. Hence, deploying LARs is always morally problematic in at least one respect.[48] While there may be **(p.181)** something intuitively appealing about this fourth argument, I believe that it falls prey to the following dilemma.[49]

Consider a situation where a commanding officer contemplates sending one of his or her subordinates on a mission, the successful pursuit of which will most likely involve killing a number of enemy combatants. The commanding officer can choose to send a human soldier or to instruct an LAR. Depending on the circumstances of the case, killing the enemy soldiers will be either justified or unjustified. Suppose first that it is unjustified. Then the morally best thing to do for the commanding officer is simply not to send out anyone.

Next, suppose that killing the enemy combatants is justified. Suppose that they are voluntarily fighting an unjust war and that, under the circumstances, killing them would be both proportionate and necessary so that they are liable to be killed.[50] If in this case the commanding officer is morally required to send out a human soldier, then he or she has to put the life of a just soldier at risk so that this soldier might show respect and compassion for a group of liable individuals. In other words, if the commanding officer has to deploy a human soldier so that the enemy soldiers may be killed in the "morally best possible way," then this implies that the morally best possible way of killing liable individuals is one that puts a just combatant in the line of fire. But, quite plausibly, the morally best way of killing liable individuals is one that minimizes the risk of harm to those who are not liable to be harmed. Hence, those who argue that there ought to be an element of humanity "even in the hell of war" owe us an explanation as to why the presence of "humanity" should take precedence over the safety of just combatants.

PRINTED FROM OXFORD SCHOLARSHIP ONLINE (www.oxfordscholarship.com). (c) Copyright Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a monograph in OSO for personal use (for details see http://www.oxfordscholarship.com/page/privacy-policy). Subscriber: Universitats- und Stadtbibliothek Köln; date: 11 June 2018

I am willing to concede that in cases where the risk of harm to a just combatant is very small, the morally best killing of an unjust enemy combatant takes place when a just combatant feels the weight of the decision and finally kills the enemy combatant with empathy and for the right reasons. But even in such cases, it may be morally supererogatory for a just combatant to take on whatever risk of **(p.182)** harm there remains. Unless a human subordinate volunteers, it may then still be right for a commanding officer to deploy an LAR as opposed to a human soldier. Moreover, while final targeting decisions made by human soldiers can in principle approach some ideal standard, in practice they often won't. To say that we should deploy human soldiers so that some ideal standard will be approximated therefore makes sense only in cases where it is reasonable to assume that the standard may in fact be approximated. Fittingly enough, these will most likely be cases where the risk of harm to just combatants is sufficiently small.[51]

An Argument in Favor of LARs

The basic idea behind the *in bello* principle of necessity is that harm is bad. In the pursuit of sufficiently worthy goals, it is sometimes morally permissible to bring about such harm, but only if we choose the least harmful way in which to pursue these goals. As I see it, the principle of necessity points to an important argument in favor of developing autonomous weapons technology further. Simply put, if we are able to develop LARs that can replace human soldiers in the theater of war, taking a wide perspective on the principle of necessity implies that we should do so as it helps us minimize the extent to which we have to put our soldiers at risk of harm when pursuing just goals.

In light of the general unease about AWS, someone might try to argue that developing autonomous weapons technology further isn't necessary to protect our own soldiers: as long as we develop remotely piloted weapons systems further—or so this argument would go—the need to have human soldiers on the ground will continually decrease anyway.[52] But while it is true that physical harm to soldiers can, at least to some extent,[53] be reduced through the use of remotely piloted weapons systems, the same is not true of mental and emotional harm.

(p.183) According to a study conducted by the US Armed Forces Health Surveillance Center, pilots of remotely piloted aircraft (RPA) report "high levels of stress and fatigue," and there is "no significant difference in the rates of [mental health] diagnoses, including post-traumatic stress disorder, depressive disorders, and anxiety disorders between RPA and [manned aircraft] pilots."[54] In other words, killing by

PRINTED FROM OXFORD SCHOLARSHIP ONLINE (www.oxfordscholarship.com). (c) Copyright Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a monograph in OSO for personal use (for details see http://www.oxfordscholarship.com/page/privacy-policy). Subscriber: Universitats- und Stadtbibliothek Köln; date: 11 June 2018

remote control is just as mentally and emotionally burdensome as other types of killing.

In one sense, this is good news. It suggests that RPA pilots are fully aware of what they are doing; they are not under the illusion that they are playing video games. In view of the fact that RPA pilots are killing other human beings, it is perfectly appropriate that their job should turn out to be emotionally and mentally burdensome.

But in another sense, the report brings bad news. Its findings suggest that even if we were to rely exclusively on remotely piloted systems while fighting a just war, we would still expose our soldiers to significant risks of harm.[55] As long as they have to select and lethally engage human targets, their long-term emotional and mental integrity is on the line.[56] It follows that if LARs have the potential to help us shield our soldiers from emotional and mental harm, then this provides us with a valid reason in favor of developing autonomous weapons technology further.

Conclusion

Unease about killer robots is widespread. Many people fear that LARs could easily be abused and that they come with an ineliminable risk of malfunction. Both **(p.184)** of these worries seem to me well grounded. When thinking about what regulatory framework we should implement with respect to the further development and use of AWSs, we should carefully consider how we can optimally respond to these worries.

In this chapter, I have argued against a third type of AWS-related worry, namely the worry that a final targeting decision made by an LAR would always be morally problematic in at least one respect. I have tried to dismantle four different arguments in favor of this claim. In addition to this, I have suggested that the protective obligations we have toward just soldiers speak in favor of developing autonomous weapon technology further.

As I see it, LARs are tools. They are potentially very sophisticated and potentially very dangerous tools. To develop and deploy them in a morally responsible manner will require much foresight, caution, and care. But, at least in principle, I see no reason why this could not be done.

Acknowledgment

For helpful discussions and comments, I thank Yitzhak Benbaji, Jonathan Birch, Andreas Carlsson, Lars Christie, Ben Ferguson,

PRINTED FROM OXFORD SCHOLARSHIP ONLINE (www.oxfordscholarship.com). (c) Copyright Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a monograph in OSO for personal use (for details see http://www.oxfordscholarship.com/page/privacy-policy). Subscriber: Universitats- und Stadtbibliothek Köln; date: 11 June 2018

Johannes Himmelreich, Ryan Jenkins, Bryan Roberts, Stefan Schubert, Tom Parr, and Thomas Seiler.

References

Bibliography references:

Aristotle. *Nicomachean Ethics*. 2nd ed. Indianapolis, IN: Hackett, 2000.

Arkin, Ronald C. "The Case for Ethical Autonomy in Unmanned Systems." *Journal of Military Ethics* 9, no. 4 (2010): 332–341.

Berreby, David. "Artificial Intelligence Is Already Weirdly Inhuman." *Nautilus* 27 (2015). http://nautil.us/issue/27/dark-matter/_artificial-intelligence-is-already-weirdly-inhuman.

Bostrom, Nick. "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents." *Minds and Machines* 22, no. 2 (2012): 71–85.

Deng, Boer. "Machine Ethics: The Robot's Dilemma." *Nature* 523 (2015): 24–26.

Docherty, Bonnie. *Losing Humanity: The Case Against Killer Robots*. Human Rights Watch. November 19, 2012. https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-%20robots.

Ekelhof, Merel, and Miriam Struyk. *Deadly Decisions: 8 Objections to Killer Robots*. PAX. February 26, 2014. http://www.paxforpeace.nl/stay-informed/news/stop-killer-robots-while- we-still-can.

Horowitz, Michael, and Paul Scharre. *An Introduction to Autonomy in Weapon Systems*. February 13, 2015. https://www.cnas.org/publications/reports/an-introduction-to-autonomy-in-%20weapon-systems.

Jenkins, Ryan, and Duncan Purves. "Robots and Respect: A Response to Robert Sparrow." *Ethics and International Affairs* 30, no. 3 (2016): 391–400.

Kant, Immanuel. *Groundwork of the Metaphysics of Morals*. Cambridge: Cambridge University Press, 1998.

(p.185) Krishnan, Armin. *Killer Robots: Legality and Ethicality of Autonomous Weapons*. Farnham, UK: Ashgate, 2009.

PRINTED FROM OXFORD SCHOLARSHIP ONLINE (www.oxfordscholarship.com). (c) Copyright Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a monograph in OSO for personal use (for details see http://www.oxfordscholarship.com/page/privacy-policy). Subscriber: Universitats- und Stadtbibliothek Köln; date: 11 June 2018

Matthias, Andreas. “The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata.” *Ethics and Information Technology* 6, no. 3 (2004): 175–183.

McMahan, Jeff. “The Basis of Moral Liability to Defensive Killing.” *Philosophical Issues* 15, no. 1 (2005): 386–405.

McMahan, Jeff. *Killing in War*. New York: Oxford University Press, 2009.

O’Connell, Mary Ellen. “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*, edited by Matthew Evangelista and Henry Shue, 224–235. Ithaca, NY: Cornell University Press, 2014.

Open Roboethics Initiative. *The Ethics and Governance of Lethal Autonomous Weapons Systems: An International Public Opinion Poll*. November 9, 2015. http://www.openroboethics.org/category/survey/analysis/analysis-military/.

Otto, Jean L., and Bryant J. Webber. “Mental Health Diagnoses and Counseling Among Pilots of Remotely Piloted Aircraft in the United States Air Force.” *Medical Surveillance Monthly Report (MSMR)* 20, no. 3 (2013): 3–8.

Pols, Hans, and Stephanie Oak. “War and Military Mental Health: The US Psychiatric Response in the 20th Century.” *American Journal of Public Health* 97, no. 12 (2007): 2132–2142.

Purves, Duncan, Ryan Jenkins, and Bradley J. Strawser. “Autonomous Machines, Moral Judgment, and Acting for the Right Reasons.” *Ethical Theory and Moral Practice* 18 (2015): 851–872.

Roff, Heather M. “Killing in War: Responsibility, Liability and Lethal Autonomous Robots.” In *Routledge Handbook of Ethics and War: Just War Theory in the Twenty-First Century*, edited by Fritz Allhoff, Nicholas G. Evans, and Adam Henschke, 352–364. New York: Routledge, 2013.

Sherman, Nancy. *Afterwar: Healing the Moral Wounds of Our Soldiers*. Oxford: Oxford University Press, 2015.

Sparrow, Robert. “Killer Robots.” *Journal of Applied Philosophy* 24, no. 1 (2007): 62–77.

PRINTED FROM OXFORD SCHOLARSHIP ONLINE (www.oxfordscholarship.com). (c) Copyright Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a monograph in OSO for personal use (for details see http://www.oxfordscholarship.com/page/privacy-policy). Subscriber: Universitats- und Stadtbibliothek Köln; date: 11 June 2018

Sparrow, Robert. “Robots and Respect: Assessing the Case Against Autonomous Weapon Systems.” *Ethics & International Affairs* 30, no. 1 (2016): 93–116.

Statman, Daniel. “Drones and Robots: On the Changing Practice of Warfare.” In *The Oxford Handbook of Ethics and War*, edited by Seth Lazar and Helen Frowe. Oxford: Oxford University Press, 2015. Availble at: http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199943418.001.0001/oxfordhb-9780199943418-e-9

Strawser, Bradley J. “Moral Predators: The Duty to Employ Uninhabited Aerial Vehicles.” *Journal of Military Ethics* 9, no. 4 (2010): 342–368.

Thomson, Judith Jarvis. “Self-Defense.” *Philosophy and Public Affairs* 20 (1991): 283–310.

Tomasik, Brian. *Do Artificial Reinforcement-Learning Agents Matter Morally?* Cornell University Library. October 30, 2014. http://arxiv.org/abs/1410.8233.

US Department of Defense. *Directive Number 3000.09 on Autonomy in Weapon Systems*. November 21, 2012. http://www.dtic.mil/whs/directives/corres/pdf/300009p.pdf.

Walen, Alec. “The Doctrine of Illicit Intention.” *Philosophy and Public Affairs* 34, no. 1 (2006): 39–67.

Yudkowsky, Eliezer. “Artificial Intelligence as a Positive and Negative Factor in Global Risk.” In *Global Catastrophic Risks*, edited by Nick Bostrom and Milan M. Cirkovic, 308–345. Oxford: Oxford University Press, 2008.

Notes:

([1]) US Department of Defense. *Directive Number 3000.09 on Autonomy in Weapon Systems*. November 21, 2012. http://www.dtic.mil/whs/directives/corres/pdf/300009p.pdf.

([2]) See http://futureoflife.org/open-letter-autonomous-weapons/.

([3]) The Campaign to Stop Killer Robots is an international coalition of currently 61 nongovernmental organizations that are calling for a pre-emptive ban on fully autonomous weapons. See www.stopkillerrobots.org.

PRINTED FROM OXFORD SCHOLARSHIP ONLINE (www.oxfordscholarship.com). (c) Copyright Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a monograph in OSO for personal use (for details see http://www.oxfordscholarship.com/page/privacy-policy). Subscriber: Universitats- und Stadtbibliothek Köln; date: 11 June 2018

([4]) In an international opinion poll, 67% of respondents felt that “all types of LAWS [lethal autonomous weapon systems] should be internationally banned.” Open Roboethics Initiative, *The Ethics and Governance of Lethal Autonomous Weapons Systems: An International Public Opinion Poll*, November 9, 2015, http://www.openroboethics.org/category/survey/analysis/analysis-military/, 2. See also www.openroboethics.org/laws_survey_released/.

([5]) See, for example, Robert Sparrow, “Killer Robots,” *Journal of Applied Philosophy* 24, no. 1 (2007): 62–772, and “Robots and Respect: Assessing the Case Against Autonomous Weapon Systems,” *Ethics & International Affairs* 30, no. 1 (2016): 93–116; Heather M. Roff, “Killing in War: Responsibility, Liability and Lethal Autonomous Robots,” in *Routledge Handbook of Ethics and War: Just War Theory in the Twenty-First Century*, ed. Fritz Allhoff, Nicholas G. Evans, and Adam Henschke (New York: Routledge, 2013), 352–364; 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*, ed. Matthew Evangelista and Henry Shue (Ithaca, NY: Cornell University Press, 2014), 224–235; and Duncan Purves, Ryan Jenkins, and Bradley J. Strawser, “Autonomous Machines, Moral Judgment, and Acting for the Right Reasons,” *Ethical Theory and Moral Practice* 18 (2015): 851–872. This is not to say that there aren’t any positive voices. For a robotics researcher well disposed toward AWS, see Ronald Arkin, “The Case for Ethical Autonomy in Unmanned Systems,” *Journal of Military Ethics* 9, no. 4 (2010): 332–341; for a philosopher, see Daniel Statman, “Drones and Robots: On the Changing Practice of Warfare,” in *The Oxford Handbook of Ethics and War*, ed. Seth Lazar and Helen Frowe (Oxford: Oxford University Press, 2015), available at: http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199943418.001.0001/oxfordhb-9780199943418-e-9

([6]) Sparrow, “Robots and Respect,” 95. See also Armin Krishnan, *Killer Robots: Legality and Ethicality of Autonomous Weapons* (Farnham, UK: Ashgate, 2009, p. 53).

([7]) This is how the Campaign to Stop Killer Robots explains the problem with AWS. See www.stopkillerrobots.org/the-problem/.

([8]) US Department of Defense, 13.

([9]) It is appropriate to talk about an autonomous weapon system and not merely about an autonomous weapon because an AWS consists not only of munition but typically also numerous other elements such as a

PRINTED FROM OXFORD SCHOLARSHIP ONLINE (www.oxfordscholarship.com). (c) Copyright Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a monograph in OSO for personal use (for details see http://www.oxfordscholarship.com/page/privacy-policy). Subscriber: Universitats- und Stadtbibliothek Köln; date: 11 June 2018

launching platform, sensors, and software that steers the targeting process. The different elements may moreover be “distributed across multiple physical platforms.” Michael Horowitz and Paul Scharre, *An Introduction to Autonomy in Weapon Systems*, February 13, 2015, 3, fn. 3, https://www.cnas.org/publications/reports/an-introduction-to-autonomy-in-weapon-systems.

([10]) Horowitz and Scharre, *An Introduction to Autonomy*, 8.

([11]) Horowitz and Scharre, *An Introduction to Autonomy*, 13–14; Krishnan, *Killer Robots*, 65.

([12]) See, for example, Bonnie Docherty, *Losing Humanity: The Case Against Killer Robots*. Human Rights Watch, November 19, 2012, 42, https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots; O’Connell, “Banning Autonomous Killing,” 232.

([13]) Krishnan, *Killer Robots*, 9–10.

([14]) Boer Deng, “Machine Ethics: The Robot’s Dilemma.” *Nature* 523 (2015): 26.

([15]) See, for example, Eliezer Yudkowsky, “Artificial Intelligence as a Positive and Negative Factor in Global Risk,” in *Global Catastrophic Risks*, ed. Nick Bostrom and Milan M. Cirkovic (Oxford: Oxford University Press, 2008), 308–345.

([16]) A neural net is a set of learning algorithms modeled loosely after the human brain.

([17]) David Berreby, “Artificial Intelligence Is Already Weirdly Inhuman,” *Nautilus* 27 (2015). http://nautil.us/issue/27/dark-matter/_artificial-intelligence-is-already-weirdly-inhuman (original emphasis).

([18]) Purves et al., “Autonomous Machines,” 855.

([19]) In bello principles are rules about the morally permissible infliction of harm in the context of war.

([20]) Purves et al., “Autonomous Machines,” 859, distinguish between empirical, moral, and practical mistakes. An LAR commits an empirical mistake if it makes a mistake in “discovering and identifying the empirical facts that are relevant” to its targeting decision; it commits a moral mistake if it comes to “the wrong normative answer about a moral problem, even given full information about the descriptive facts”;

PRINTED FROM OXFORD SCHOLARSHIP ONLINE (www.oxfordscholarship.com). (c) Copyright Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a monograph in OSO for personal use (for details see http://www.oxfordscholarship.com/page/privacy-policy). Subscriber: Universitats- und Stadtbibliothek Köln; date: 11 June 2018

finally, it commits a practical mistake if it executes a correct decision in a flawed way, such as “by reacting a moment too slowly.”

([21]) Arkin, “The Case for Ethical Autonomy,” 333.

([22]) Child soldiers might provide an exception.

([23]) By this, I mean that an LAR might always commit, in the terminology of Purves et al., “Autonomous Machines,” an empirical, moral, or practical mistake. See fn. 9.

([24]) Note also that if it is reasonable to assume that LARs might someday outperform human soldiers in terms of engaging only in permissible killings, this might provide a moral reason in favor of developing the technology further.

([25]) Purves et al., “Autonomous Machines,” 858.

([26]) Judith Jarvis Thomson, “Self-Defense,” *Philosophy and Public Affairs* 20 (1991): 293–294, argues that the reasons for which an action is performed are irrelevant to its permissibility. In spirit, Thomson’s position is at odds with Purves et al.’s ideas. Strictly speaking, it is nevertheless consistent with Purves et al.’s core contention that an action is morally flawed *in at least one respect* whenever it is performed for the wrong reasons.

([27]) See, for example, Alec Walen, “The Doctrine of Illicit Intention,” *Philosophy and Public Affairs* 34, no. 1 (2006): 39–67.

([28]) See Sparrow, “Robots and Respect,” 109–110; also the later discussion under “A Fourth Argument Against LARs.”

([29]) Sparrow, “Robots and Respect”; Ryan Jenkins and Duncan Purves, “Robots and Respect: A Response to Robert Sparrow,” *Ethics and International Affairs* 30, no. 3 (2016): 391–400.

([30]) By distinguishing between respect in the narrow and the wide senses, Jenkins and Purves follow Jeff McMahan, who applies a similar distinction to proportionality. See *Killing in War* (New York: Oxford University Press, 2009, 20–21.

([31]) See Immanuel Kant, *Groundwork of the Metaphysics of Morals* (Cambridge: Cambridge University Press, 1998), esp. sect. 1.

([32]) This was pointed out to me by Andreas Carlsson.

PRINTED FROM OXFORD SCHOLARSHIP ONLINE (www.oxfordscholarship.com). (c) Copyright Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a monograph in OSO for personal use (for details see http://www.oxfordscholarship.com/page/privacy-policy). Subscriber: Universitats- und Stadtbibliothek Köln; date: 11 June 2018

([33]) See Aristotle, *Nicomachean Ethics*, 2nd ed. (Indianapolis, IN: Hackett 2000), esp. bk. II.

([34]) Sparrow, “Killer Robots.” The idea that the use of learning machines creates a responsibility gap was first presented by Andreas Matthias, “The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata,” *Ethics and Information Technology* 6, no. 3 (2004): 175–183. In this chapter, I focus on Sparrow’s treatment of the issue because Sparrow, unlike Matthias, argues specifically that the existence of a responsibility gap makes the deployment of LARs necessarily morally problematic.

([35]) It is conceivable that part of the moral responsibility remains with the developers of the software, simply because they voluntarily decided to create and disseminate something they knew might cause wrongful harm even when handled conscientiously.

([36]) Cf. Jeff McMahan, “The Basis of Moral Liability to Defensive Killing,” *Philosophical Issues* 15, no. 1 (2005): 394–395.

([37]) As Tom Parr and Andreas Carlsson have pointed out to me, there is an interesting question here about who is morally responsible for wrongful harm an agent causes in the conscientious pursuit not of his or her own purposes but of a morally obligatory aim. Suppose LARs are used in a morally required humanitarian intervention and the LARs inflict wrongful harm despite the fact that all reasonable precautions were taken. In such a case, who is morally responsible for the harm inflicted? Interesting as it is, I won’t pursue this question further here as it arises not from the use of LARs in particular but simply from the fact that things can sometimes go wrong when we pursue what we have reason to believe is a morally obligatory end.

([38]) McMahan, “The Basis of Moral Liability,” 394–404.

([39]) Sparrow, “Killer Robots,” 66, 70.

([40]) Nick Bostrom, “The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents,” *Minds and Machines* 22, no. 2 (2012): 71–85.

([41]) Sparrow, “Killer Robots,” 70–71.

([42]) Sparrow, “Killer Robots,” 70–73.

PRINTED FROM OXFORD SCHOLARSHIP ONLINE (www.oxfordscholarship.com). (c) Copyright Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a monograph in OSO for personal use (for details see http://www.oxfordscholarship.com/page/privacy-policy). Subscriber: Universitats- und Stadtbibliothek Köln; date: 11 June 2018

([43]) The organization People for the Ethical Treatment of Reinforcement Learners (PETRL) is based on the premise that machines might be capable of suffering. For an argument in favor of this premise, see Brian Tomasik, *Do Artificial Reinforcement-Learning Agents Matter Morally?* Cornell University Library, October 30, 2014, http://arxiv.org/abs/1410.8233. I thank Stefan Schubert for drawing my attention to PETRL.

([44]) Once it is no longer appropriate to see LARs as tools in the hands of human agents, an alternative to thinking of them as capable of inflicting wrongful harm is to see them as forces that inflict the type of amoral harm that an earthquake or a wild animal inflicts.

([45]) Sparrow, “Killer Robots,” 66.

([46]) “Banning Autonomous Killing,” 232.

([47]) *Deadly Decisions: 8 Objections to Killer Robots*, PAX, February 26, 2014, 7, http://www.paxforpeace.nl/stay-informed/news/stop-killer-robots-while-we-still-can.

([48]) The second argument against LARs that I discussed can be read as a special case of the fourth argument presented here. Read in this way, Purves et al. are claiming that the morally best killings are performed for the right reasons and that a killing performed by an LAR can never be performed for the right reasons.

([49]) The forcefulness of this dilemma depends on the truth of the revisionary approach to just war theory that is defended by Jeff McMahan and others. While I believe that the revisionary approach to just war theory is correct, I won’t try to argue for this here. According to revisionist just war theory, war is not a morally privileged activity: whatever moral principles govern the infliction of harm in ordinary life govern its infliction in the context of war as well. This implies that there is a difference in the moral status of just and unjust soldiers. The former are fighting justly for a just cause and are therefore not liable to be harmed. The latter are fighting unjustly or for an unjust cause and are therefore liable to harmed.

([50]) If someone is liable to some harm, he or she is not wronged if that harm is imposed on him or her.

([51]) What if the people to be killed are civilians who are not liable to be killed, but it is nevertheless morally permissible to bring about their deaths as a foreseen but unintended side effect? In such a case, it may well be morally appropriate not to prioritize the safety of the just

PRINTED FROM OXFORD SCHOLARSHIP ONLINE (www.oxfordscholarship.com). (c) Copyright Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a monograph in OSO for personal use (for details see http://www.oxfordscholarship.com/page/privacy-policy). Subscriber: Universitats- und Stadtbibliothek Köln; date: 11 June 2018

combatants over the interests of the civilians but to take the interests on both sides equally seriously. Yet even if this is correct, it might nevertheless turn out that the just combatants' interest in not being put at risk of harm outweighs whatever interest the civilians have in being killed by a human being as opposed to an LAR.

([52]) Cf. Bradley J. Strawser, "Moral Predators: The Duty to Employ Uninhabited Aerial Vehicles," *Journal of Military Ethics* 9, no. 4 (2010): 342–368.

([53]) AWS may help reduce the risk of physical harm to human soldiers in places where remotely piloted weapon systems cannot be deployed, for example, in environments where our adversaries are able to make use of electronic jamming to disrupt the wireless data links needed to run remotely piloted systems. I owe this point to Ryan Jenkins. See also Krishnan, *Killer Robots*, 37–42.

([54]) Jean L. Otto and Bryant J. Webber, "Mental Health Diagnoses and Counseling Among Pilots of Remotely Piloted Aircraft in the United States Air Force," *Medical Surveillance Monthly Report (MSMR)* 20, no. 3 (2013): 3.

([55]) It is possible to doubt that recent statistics on the mental health of US military members are reliable sources of information about the mental health risks that just soldiers are exposed to. Hans Pols and Stephanie Oak, "War and Military Mental Health: The US Psychiatric Response in the 20th Century," *American Journal of Public Health* 97, no. 12 (2007): 2132–2142, describe the mental and emotional harms suffered by American soldiers in World War II as serious and widespread. They report that "in total, there were more than 1 million neuropsychiatric admissions to the medical services of the US armed forces" (p. 2135). Together with the fact that the rates of mental health diagnoses of RPA pilots are similar to those of other pilots, this suggests that RPA pilots fighting for a just cause would be at significant risk of emotional and mental harm.

([56]) Nancy Sherman, *Afterwar: Healing the Moral Wounds of Our Soldiers* (Oxford: Oxford University Press, 2015), argues that soldiers' *moral* integrity is on the line as well: as soldiers often fall short of their own moral ideals, they have to somehow come to terms with this failure, as well as with the associated feelings of guilt and shame.

PRINTED FROM OXFORD SCHOLARSHIP ONLINE (www.oxfordscholarship.com). (c) Copyright Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a monograph in OSO for personal use (for details see http://www.oxfordscholarship.com/page/privacy-policy). Subscriber: Universitats- und Stadtbibliothek Köln; date: 11 June 2018

Access brought to you by:

PRINTED FROM OXFORD SCHOLARSHIP ONLINE (www.oxfordscholarship.com). (c) Copyright Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a monograph in OSO for personal use (for details see http://www.oxfordscholarship.com/page/privacy-policy). Subscriber: Universitats- und Stadtbibliothek Köln; date: 11 June 2018