REASONS TO PUNISH AUTONOMOUS ROBOTS

Zac Cogley
zaccogley@gmail.com

1. Introduction

Deploying autonomous robots in military contexts strikes many people as terrifying and morally odious. What lies behind those reactions? One thought is that if a sophisticated artificial intelligence were causally responsible for some harm, there will be no one to punish for the harm because no one—not programmers, not commanders, and not machines—would be morally responsible. Call this the *no appropriate subject of punishment* objection to deploying autonomous robots for military purposes. The objection has been discussed by several authors (Matthias 2004; Lucas 2013; Danaher 2016), but is most fully developed in Robert Sparrow's paper "Killer Robots" (2007).[1]

There have been other attempts to address the objection (Kershnar 2013; Simpson and Müller 2016), but to my knowledge no one has tried to do so by taking seriously the idea of the robots, themselves, being both morally responsible and appropriate subjects of punishment. Perhaps that's because most theorists find punishing robots to be "utterly ludicrous," as noted by George R. Lucas, formerly of the U.S. Naval Postgraduate School and the U.S. Naval Academy (Lucas 2013, 223). Lucas himself takes the concern behind lacking a subject of punishment to pose an "admittedly-formidable" design problem that robot engineers and programmers should be required to address. I won't be able, here, to do any substantive work toward solving the design problem. I hope instead to convince you that the concept of punishing robots isn't totally absurd.

In what follows, I first discuss the design and plausibility of punishable autonomous military robots. I argue that it is an engineering desideratum that these devices be sensitive to relevant moral considerations in their domain of operation and that they be responsive to human criticism and blame. In addition, I suggest that at some point in the future it will in fact be possible to build such machines, but that such machines will not be moral patients as they will lack the capacity for pain and only have domain-specific autonomy. To help fix intuitions and to have a relevant example for discussion, I describe a test case of an autonomous robot committing a war crime. Following that, I develop the no appropriate subject of punishment objection to deploying such a robot and discuss extant, not fully successful, replies. I then respond to the argument by defending the claim that future autonomous military robots can be morally responsible and blameworthy for their conduct. Does this give us reason to punish them? I hold that whether it does depends on why we find human punishment reasonable and discuss relevant options. Finally, I conclude by discussing an important moral implication of my argument concerning the permissibility of deploying autonomous military robots. Deploying future autonomous military robots is of true moral concern because of the possibility that such machines might be deployed *without* engineering them to be sensitive to moral considerations.

[1] The more common term in the literature is 'responsibility gap' or 'accountability gap' objection. I instead use 'no appropriate subject of punishment' because some objections that concern robot punishment are not due to worries about robot responsibility or accountability.

2. Autonomous Military Robots: Design and Plausibility

As Sparrow notes, ‘autonomy’ means different things to different authors (2007, 65). Some would use the term to characterize cruise missiles and torpedoes. In the context of this paper, that level of ‘autonomy’ is insufficient to characterize a truly autonomous military robot. Unless specified otherwise, the autonomous military robots under discussion here are machines that will be able to decide whether potential targets are friend or foe and combatants or noncombatants and then decide whether to attack, how to attack, and when to disengage. Following Sparrow, I hold that such future autonomous robots will be sophisticated enough that their actions will be based on the internal representational states (like beliefs, desires, and values) of the artificial intelligence guiding them. They will have some capacity to form and revise these states, themselves, and they will have the ability to learn from experience (Sparrow 2007, 65). As noted by Bertram Malle and Matthias Scheutz (2015), robots capable of acquiring and using information about the world to guide their actions in accord with their goals would display the faculties of choice and intentional action.

I follow Malle (2016, 252–53) in developing the following argument regarding the moral capacities of future autonomous social robots—call this the *Design Argument*. Engineering robots able to perform successfully in human social situations—including wartime—cannot be achieved by relying on static, rule-following programs. Consider that even the ‘simple’ human social interactions involved in buying groceries still need a human overseeing the self-check-out machines to handle unusual circumstances. Human behavior is creative, adaptable, and hard to predict, so any robot that successfully interacts with humans in social situations must be able to learn from experience and flexibly respond to new information. A key part of the new information we humans use to respond properly to novel social situations is the moral criticism and threat of social rejection presented by other humans. Thus, a robot responding well in human social situations must be able to properly interpret and respond to moral criticism and the threat of social rejection. It needs to be able to respond appropriately to human expressions of blame.

It follows, then, that it is an engineering desideratum of useful autonomous robots deployed in military contexts that are able to make moral discriminations and also properly interpret and respond to human blame expressions. Consider that for autonomous robots to be useful to us in military contexts that require telling friend from foe or combatant from noncombatant, those machines must first be able to make those very discriminations—the same ones human soldiers must make. If we could create robots that would never make mistakes, we would then never have reason to worry about robot mistakes and improving robot performance (Purves, Jenkins, and Strawser 2015). But if, as is certain, our robots will sometimes err, they will also need a mechanism that prompts them to revise their representations with the aim of not making those mistakes in the future. Call the internal state that prompts representation revisions ‘machine guilt’—it is the functional machine analogue of human guilt: the state of subjectively caring about having done wrong (D. W. Shoemaker 2003, 99).[2]

[2] I’m using the term ‘machine guilt’ as a placeholder for the relevant functional state and, in my view, it’s open whether ‘machine guilt’ could be operationalized by a natural extension of extant learning methods. Whatever machine learning mechanisms are required to operationalize machine guilt might deviate substantially from current approaches. Or, perhaps not – it will be interesting to find out!

Additionally, note that while I take it that human guilt involves subjectively caring about having done wrong, ‘machine guilt’ does *not* require that the machine be a conscious, Human guilt is, at least in part, an error correction mechanism. When prompted by the criticism and blame of others, guilt leads us to update our internal representations of situations where we have made the wrong choice (Damasio 2006; Baumeister et al. 2007; Giner-Sorolla and Espinosa 2010). Similarly, socially useful autonomous robots will use machine guilt for error correction purposes.[3] Thus, the Design Argument says that to be truly socially useful, future autonomous robots must have these features. But is making a robot with these characteristics even achievable?

Some deny the possibility of robots with these capacities, for example because they think robots will be unable to capture the meaning of information (Stahl 2004) or will only be able to follow pre-programmed rules and cannot appreciate reasons (Purves, Jenkins, and Strawser 2015).[4] These arguments and claims depend on an outdated picture of AI research, neglecting machine learning and deep neural networks. One of the most impressive advances, here, is AlphaGo, an AI developed by Alphabet's (formerly Google) DeepMind subsidiary (Silver et al. 2016) that plays the game of Go better than any human being. AlphaGo has now beaten the world number one player, Ke Jie (Anthony 2017), 18-time world champion Lee Sedol, as well as several other human grand masters.

These systems are programmed with machine learning algorithms – like regression – that apply across domains and, through trial and error, learn to extract and update relevant patterns from the data. The algorithmic process used by such systems sensitizes the resultant neural networks to relevant reasons that operate in their domain of deployment. For example, a machine learning algorithm that learns to play chess competently will become sensitive to representations of concepts like material advantage, space, and king safety. These representations would then play a role in guiding its choices and may also be modified in response to additional feedback – both roughly comparable to how human players deploy and modify strategic representations in choosing between moves. Deep learning representations are not encoded propositionally. Instead, these systems have morphological content: they retain information in their standing structure that is automatically accommodated during processing (Horgan and Potrč 2010). Morphological content likely undergirds a large portion of human moral decision-making (Horgan and Timmons 2007), as when we intuitively recoil at the thought of a puppy being abused without having to deliberate about whether such acts are wrong.

Noah Goodall has proposed a machine learning strategy for programming moral decision-making about crashing in autonomous vehicles (Goodall 2014, 63) that could be ported to the military context. The idea is to train a neural network on a data set of recordings of real crashes as well as near misses, in addition to simulations of both. Human beings would then score potential actions and results as more and less morally correct.[5] The

experiencing subject. I hold only that machine guilt plays the same functional role. More on this later in this section and beyond.

[3] In Section 6, I'll discuss the criteria that should govern how autonomous robots should modify their machine guilt in response to feedback.

[4] Additionally, some hold that robots will only be able to have a determined, and not 'free' will (Roff 2013). However, it's an open question whether human wills are free in the relevant sense (McKenna and Pereboom 2016), so we shouldn't fault autonomous robots for falling below a moral standard we, ourselves, might not reach.

[5] Obviously, human beings – and moral theories – differ at times on how morally correct/incorrect the very same actions are and what the proper grounds are for scoring correctness/incorrectness. Addressing these disagreements is another substantial challenge

neural network would then use this data to update its internal representations of which outcomes to pursue and avoid. For military use, we would use actual battlefield recordings as well as simulations, but the overall methods would be similar. Finally, Ronald Arkin and colleagues have already created a rudimentary software system integrating a simple version of ethical decision making using moral emotions like guilt to respond flexibly to battlefield information (Arkin, Ulam, and Wagner 2012).

These observations form the core of what I term the *Plausibility Argument*, the upshot of which is that it is reasonable to think that future autonomous robots can be engineered so they are responsive to moral considerations and sensitive to moral critique. I don't want to sugarcoat the engineering challenges here. Such robots are a long way from being developed.[6] My contentions are just that autonomous robots useful in social situations in the same ways that human beings are useful will, of necessity, be engineered to be responsive to moral considerations and moral critique (the Design Argument) and that developing such machines is an research challenge, not an in-principle impossibility (the Plausibility Argument).[7]

Let me also lay out some additional suppositions regarding future autonomous military robots. First, such robots will lack the capacity for pleasure and pain. We can likely avoid accidentally producing robots that feel pain by not building them with important mechanisms that undergird pain in sentient animals. For example, we might omit nociception mechanisms for extreme temperatures, noxious mechanical stimuli, and chemical agents (Julius and Basbaum 2001), which are necessary for our feeling pain in response to these mechanisms. Such robots will need to monitor the functioning of their parts via some feedback system, but we can likely program the robots so they can monitor their own functioning without pain or pleasure. Additionally, in human beings we can dissociate the sensory aspects of pain from its disagreeableness (Aydede 2013), so it would be surprising if we couldn't build robots with sensory and representational capabilities that don't experience unpleasantness. Of course, there are human representational states – like guilt – that are both unpleasant and representational. I'm just supposing that it's possible to build a machine that has a functional similar state that lacks the unpleasant/painful aspect.

It might be thought that if the robots in question won't feel pain/unpleasantness that would pose a barrier to robots possessing the machine guilt error correction mechanism I outlined above. Since guilt in human beings is, at least in part, an unpleasant sensation (Morris 1976, 101; Clarke 2016, 122), the worry is that a robot that doesn't feel pain couldn't thereby experience guilt in the way that human beings do. It might not. But I'm not interested in whether it would make sense to call the relevant state "guilt" or whether machine guilt and human guilt will possess all the same properties. Who's to say whether the

for the development of the kinds of systems under discussion. Still, these challenges can likely be addressed. For one, there is substantial agreement that certain kinds of actions – like killing innocents in an unprovoked attack – are morally wrong no matter how that wrongness is explained. Additionally, a system that is uncertain about what the right thing to do is in a situation because the system's training data was conflicting will accurately capture human ambivalence about that type of case.

[6] See (Arnold, Kasenberg, and Scheutz 2017) for a discussion of some of the relevant engineering difficulties.

[7] Readers who still doubt the in-principle possibility of such systems are of course still welcome to read the remainder of the paper in a conditional format: "if it were possible to develop such systems…."

functionally similar state is guilt, proper (Dennett 1997, 361)? What we need for the machine guilt state I'm describing is just that the robot has some way of representing having done the wrong action within a domain of activity, a way of representing the seriousness of the wrong, and a mechanism by which those representations cause the robot to update its representations of the moral valence of actions it could perform. Having the ability to feel pain/suffering/unpleasantness, I've suggested, isn't necessary for this process to occur.

Second, I suppose that the autonomy of future military robots will be domain-specific. Put another way, they will have domain-specific moral abilities without full moral agency. Why think this? The kinds of representations algorithms extract and update depends on the data and design decisions of the programmers. AlphaGo plays Go expertly, but would be no help at all in playing chess well. We can therefore expect that future military robots will have been trained to learn representations like *combatants* and *noncombatants*—as well as what objects fall under those representations—and not, for example, representations tracking whether someone is a *student* or *professor*.

In sum, then, I propose that future military robots should possess limited autonomy that enables them to make their own decisions regarding attack and engagement in the military theater. They will not possess the domain-general cognitive capacities that ground human autonomy or the capacity for pain, meaning that they will not be moral patients.[8] They will be sensitive to relevant moral considerations in their domain of operation. Finally, they will have an error correction mechanism—machine guilt—prompted by the moral criticisms of relevant personnel that causes them to update their representations when necessary.

3. Test Case

Now that we have entertained the design and plausibility of the future autonomous robots I want to consider, here's a test case from Sparrow to help us fix our intuitions about the kind of situation that might lead to the no subject of punishment objection:

> Imagine that an airborne AWS [Autonomous Weapon System], directed by a sophisticated artificial intelligence, deliberately bombs a column of enemy soldiers who have clearly indicated their desire to surrender. The AWS had reasons for what it did; perhaps it killed them because it calculated that the military costs of watching over them and keeping them prisoner were too high, perhaps to strike fear into the hearts of onlooking combatants, perhaps to test its weapon systems, or because the robot was seeking to revenge the 'deaths' of robot comrades recently destroyed in battle. Whatever the reasons, they were not the sort to morally justify the action. Had a human being committed the act, they would immediately be charged with a war crime. Who should we try for a war crime in such a case? The robot itself? The person(s) who programmed it? The officer who ordered its use? No one at all? As we shall see below, there are profound difficulties with each of these answers. (Sparrow 2007, 66–67)

Let me flesh out the case a little more. Let's suppose that all the soldiers in the enemy column are waving white flags and have laid down their arms. Given this, the autonomous

[8] Of course, we shouldn't just assume that future robots we create will lack these capabilities. We have a moral obligation to try to determine if they have them or not, in order to ensure we're not creating moral patients we then use in objectionable ways. For a proposal regarding how to test for machine consciousness, see (Schneider and Turner 2017).

robot is required to accept their surrender (Robertson Jr 1996, 543). Why, then, did the robot attack? Broadly speaking, there are two sorts of options.

One possibility is that the robot made a serious targeting error that resulted from not perceiving that all the soldiers were offering surrender. (If not all soldiers in a unit are attempting to surrender, there is no obligation on the part of the attacker to stop firing.) This possibility—that the attack resulted from an error—will surface again later when I discuss the extant replies to the no subject of punishment objection. Another possibility is that the robot acted intentionally; it did aim to kill the enemy soldiers. Suppose this is because the autonomous robot was aware of a recent incident in which friendly forces took serious losses after acquiescing to a plea for surrender that turned out to be a ruse. (The enemy forces attacked after the friendly unit stopped firing and was ready to accept their surrender.) The possibility that the robot acted intentionally will come up below in the 'Robot Punishment' section.

4. Understanding the No Subject of Punishment Objection

Set aside the different interpretations of the test case for a moment and return to Sparrow's question: is there any subject of punishment when, as above, an autonomous robot commits a war crime? There are two ways to understand the worry that there is no subject of punishment. The first is based on an empirical claim: that the human desire for punishment will be frustrated. This is a minor concern going forward, but discussing it will help make clear the more important issues below.

The idea is that human beings want to punish *something* when things go badly wrong and, if robots are causally responsible for doing wrong, we humans won't have anything satisfying to punish.[9] This is an aspect of John Danaher's arguments (2016).[10] Danaher's concern starts from the idea that people are innate retributivists—they want to punish wrongdoing based on the perceived deservingness of offenders. (And, of course, some moral theorists believe people are correct to be this way.) Danaher predicts that people won't desire to punish autonomous robots because the robots don't seem deserving of punishment. If so, people's desire for retribution will go unfulfilled.

Well, what if it does? One reason we might be concerned is the possibility of scapegoating (Danaher 2016, 307). If we really want to punish someone whenever something goes badly wrong and we don't feel it will be satisfying to punish the robot, we may look around to punish someone else. Maybe we'll punish the robots' programmers or manufacturers. If we care about justice, Danaher urges, this should give us pause. Additionally, if our desires for retribution are going unfulfilled, this presents an opportunity for anyone who believes that retributivism is not the correct account of punishment's justification. Too much robot-caused harm could upset the retributivist status quo, leading to other approaches to punishment being taken more seriously (Danaher 2016, 308).

As a first reply to these worries, we can wonder why it would be a serious concern if other, non-retributive, accounts of punishment's justification got a hearing. That might be a

[9] Daniel Wegner and Kurt Gray call this urge to find someone or something to blame when something goes seriously wrong 'dyadic completion.' Historically, it wasn't uncommon in the medieval era to hold nonhuman animals responsible for harms, including pigs executed for murder, locusts found guilty of crop destruction, and even a rooster given the death penalty for laying an egg (2016, 51).

[10] Strictly speaking, Danaher says his argument involves what he terms a 'dance' between the normative and descriptive. I've separated those aspects to more clearly present them.

good thing, especially given some of the serious doubts raised about retributivism (Dolinko 1991, 1991; Boonin 2008 Ch. 3). Additionally, Danaher himself raises responses that undermine the force of the argument. If what is really important to us is just to have someone to punish, we can insist that commanding officers who order the use of robots are strictly liable for any improper harms they derivatively cause (2016, 306–7). That gives us someone to punish and might also help make sure autonomous military robots are only used judiciously. Finally, and most interesting to me, is the possibility that humans dealing with sophisticated robots of this sort may anthropomorphize them (Danaher 2016, 305–6) and so actually want to, and be satisfied by, punishing the robots.

Evidence for this comes from multiple sources. In general, it appears that human perception of minds in other things primarily depends on two factors: whether a thing is taken to be a "thinking doer" (an agent), or a "vulnerable feeler" (an experiencer, or patient) (Robbins and Jack 2006; H. M. Gray, Gray, and Wegner 2007; Jack and Robbins 2012; Wegner and Gray 2016). Agents are seen as subjects of moral responsibility while patients are seen as subjects of moral rights (K. Gray and Wegner 2009). Robots are generally construed by humans to be agents that lack experience (K. Gray and Wegner 2012).[11] Being seen as deserving of punishment for wrongdoing is highly correlated with being seen as an agent (H. M. Gray, Gray, and Wegner 2007). Thus, in general we see robots as agents, not experiencing patients. Punishment is seen as deserved by beings with agency. Therefore, we are likely to find sophisticated robots deserving of punishment.

More specific evidence comes from a study (Kahn Jr et al. 2012) involving human participants playing an item-finding game judged by a robot named Robovie. The robot was controlled off-scene by the experimenters, but 71% of subjects thought Robovie was operating completely on its own. At the beginning of the experiment, participants were introduced to Robovie, who then gave them a brief tour of the room. During the tour, subjects asked Robovie follow-up questions and elaborated on themes being discussed with the robot before playing the game. The game required subjects to find seven items in a short time period and was constructed so everyone would find even more. No matter how many items were found, however, Robovie would insist that subjects only found five items. Many subjects became visibly annoyed and confronted Robovie [link to video of one interaction], insisting they had won the game. In surveys after the game, 65% of subjects credited Robovie with some level of accountability—as compared with no subjects attributing any accountability to a vending machine and most subjects ascribing full accountability to human beings.

The upshot is that we should expect people to treat robots as moral agents if they perceive them to have a system of moral norms that guide their actions in conjunction with at least one of these traits: moral cognition and affect, a moral vocabulary, moral decision-making and action, or moral communication (B. F. Malle and Scheutz 2015; Bertram F. Malle 2016). These latter traits, in particular moral decision making and action, will help prompt people to attribute moral agency to robots. Note that, with respect to assessing Danaher's empirical claims about whether we'll want to punish robots, it's not required that we actually build robots that have these attributes. It would be enough, instead, to simply build robots so people attribute these characteristics to them. This suggests we can build robots so that people's desires for retributive punishment can be satisfied.

[11] Additionally, robots that have human facial features or are described as having experiences like hunger or fear are seen as unnerving, likely because those features are associated more with being a patient, not an agent (K. Gray and Wegner 2012).

A more interesting way to understand the no subject objection is that if we deploy autonomous robots in wartime and they commit a war crime, punishment will not be *morally reasonable*—after such an event, there will be no one who *deserves* punishment: a significant moral concern (Matthias 2004; Sparrow 2007; Danaher 2016). Here is an argument motivated by the concern:

(1) Some action contexts—including war—are so morally serious that it is unjust not to punish agents who commit grave wrongs in those contexts.

(2) Deploying autonomous robots in these contexts—including war—will mean that there will be no one deserving of punishment for harms the robots cause.

(3) Therefore, deploying autonomous robots in these contexts is unjust.

Some brief comments on the argument: what contexts count as so morally serious? At least war and policing, but perhaps also medicine and law. These are all professional contexts where professionals hold great power over 'ordinary' people. Why would no one be deserving of punishment? It's not clear that the designers or programmers of autonomous robots would be deserving, nor is it clear that commanders of the robots would be deserving, nor does it seems the robots, themselves, would be. Sparrow calls this 'the trilemma.' If we cannot escape the trilemma and so there is no one deserving of punishment, yet we deploy a robot that commits a serious wrong, then we act unjustly.[12] As Sparrow puts it, "The least we owe our enemies is allowing that their lives are of sufficient worth that someone should accept responsibility for their deaths" (Sparrow 2007, 67).

Let's explore the trilemma. Would it make sense to hold the designers or programmers of autonomous robots responsible for the robot-caused harms (Sparrow 2007, 69–70)? Not if the programmers have made clear the possibility that the robots' may attack the wrong targets and have taken all reasonable care to program and train the robots not to do so.[13] Additionally, it is arguably not the programmers' responsibility to decide to use this technology. Finally, autonomous systems of this sort will be able to make choices that go beyond those predicted or encouraged by programmers. What about the commanding officer (Sparrow 2007, 71)? Importantly, orders to an autonomous robot by the commanding officer will not wholly determine the robots' actions. If these machines really choose their own targets, the commanding officer doesn't deserve to be held responsible for the deaths. We can imagine in our test case that the autonomous robot was ordered directly only to do reconnaissance, came under fire from the enemy troops, and then returned fire, leading to the attempted surrender.[14] Finally, what about holding the robot, itself, responsible? Sparrow is right that, "To hold that someone is morally responsible is to hold that they are the appropriate locus of blame or praise and consequently for punishment or reward" (Sparrow 2007, 71). Could a robot be deserving of blame or praise and so punishment or reward?

[12] As readers might expect, my main strategy of responding to the trilemma is to push back on the claim that robots can't be appropriate subjects of punishment, but I'll say something about each horn of the trilemma in what follows.

[13] The notion of 'all reasonable care' is further developed in section 5, when discussing Simpson and Müller's reply to the no subject of punishment objection.

[14] Note that if the programmers or commanding officers have acted unjustly or have not taken all reasonable care they could bear some responsibility for negative outcomes that result. And, it would be possible for multiple parties to bear responsibility. I follow Sparrow in setting these cases aside to focus on the interesting possibility presented by the situation where the programmers and commanding officers have not acted badly – what then?

Sparrow accepts that advanced artificial intelligences may have desires and goals that go beyond those of their military role. If so, we could then frustrate those desires by restricting the robot's liberty or destroying it. But Sparrow denies that these could actually be punishment because just frustrating the robot's desires wouldn't mean that the robot is *suffering*. Sparrow holds that the suffering of those we punish must be morally compelling for us in the sense that, if the suffering were unnecessary—if the one punished was innocent—we would feel we had committed a serious wrong. But, Sparrow continues, if we were actually able to build robots like that, we wouldn't have achieved our aims. Our purpose in building and deploying such robots was to fight wars without risking our soldiers, but now we are simply putting other morally salient beings at risk (Sparrow 2007, 73). In a nutshell, the idea is that if a robot can't suffer in the right way, it won't be an appropriate subject of punishment, so we act unjustly to those it targets. If it is really a proper subject of punishment, we have reasons not to deploy it in war.

5. Extant Replies

It will be helpful to briefly explore two extant replies to Sparrow's argument, as how they fail is instructive. The first argument, due to Stephen Kershnar, starts with the claim that having someone to hold accountable typically doesn't affect the morality of defensive violence (2013, 237). For example, suppose someone is about to die from natural causes. That person is still permitted to use violence to defend others from attack, even though her imminent death means there will be no one to hold accountable for mistakes she might make. Kershnar considers the likely response: that the defender can still be *deserving* of blame or punishment if she's dead. His reply is then that if we want to be sure to have someone to hold accountable, then the person deploying the autonomous robot can similarly be held accountable (perhaps via a strict liability regime). However, the question is about whether we will have someone who really deserves blame and punishment, not whether we can in fact produce someone to hold accountable. Kershnar suggests the latter without answering the former. In responding to Sparrow, we need to avoid the same oversight.

A more plausible response to the argument comes from Thomas W. Simpson and Vincent C. Müller. They hold that autonomous robots are engineered products and so deploy the general moral framework used for dealing with potentially engineered risks. The idea is that autonomous robots can be justly deployed in case (2016, 316):

(1) The risks that such robots pose to non-combatants are less than those posed by all-human armies, and
(2) The amount of risk is as low as technologically feasible.

Note that we don't demand that other engineered projects function perfectly. We only demand that they operate within their proper risk tolerance—the likelihood of a harmful failure given their normal use, the resources we have to develop them, and the problem they are being developed to solve. If a bridge fails due to a "1,000-year" rain in combination with heavy traffic, but was correctly only engineered to withstand a "500-year" rain, no one is blameworthy for the failure. This means that—as considered in the 'error' version of the test case above—some robot killings will occur for which no one is blameworthy. But if the robots are responsibly engineered and regulated, this poses no moral problem. Non-combatants will actually be *safer* when robots that meet Simpson and Müller's conditions are deployed.

While Simpson and Müller's argument is stronger than Kershnar's, there are two reasons it doesn't answer the no subject of punishment objection. The first is that the level of autonomy Simpson and Müller envision for the autonomous robots they consider doesn't rise to what we might call ‘full’ or ‘human-level’ autonomy—the level of autonomy that motivates the no subject of punishment objection. In their paper, they consider autonomous robots that have algorithms for telling soft-skinned vehicles (like cars and trucks) from military ones (like armored vehicles and artillery pieces) or programs that target only pickup trucks with heavy weaponry mounted on the rear. But these autonomous robots differ only in degree from current systems, like the Phalanx Close-In Weapons System, which, in automatic mode, fires on all objects that fall into particular size, distance, velocity and maneuverability ranges. These systems’ capacities do not rise to the level of fully autonomous weapons and they do not make *moral* discriminations. Thus, Simpson and Müller haven’t offered an argument that addresses the kind of robots that generate the no subject of punishment objection.

In addition, if we were permitted to apply the engineering-risk framework to systems of any autonomy level, we should be able to reapply the framework to the use of human soldiers. Suppose, then, we deploy a battalion of “better engineered” human soldiers—their training was more rigorous than the training of the previous generation of soldiers—who meet both of Simpson and Müller’s requirements. Suppose one of the better trained soldiers commits a war crime. If Simpson and Müller are correct, we have no reason to even consider whether the soldier deserves blame or punishment for what she did. All the moral questions would have been answered using the risk-engineering framework. But they aren’t. The reason is that sufficiently autonomous human soldiers choose whether to stay within the risk tolerances of the mission, or whether to go beyond the aims of the operation. Soldiers who go beyond the risk tolerances of the mission can be blameworthy and deserve punishment for what they do. The same thing would be true of sufficiently autonomous weapons.

6. Robot Punishment

Back to Sparrow. I deny both of Sparrow’s claims regarding the proper deployment of autonomous robots. Robots that can’t suffer can be appropriate subjects of punishment. Robots that are appropriate subjects of punishment can also be sensibly deployed in war. Take the latter claim, first. Sparrow’s worry is that autonomous robots would be beings to whom we have moral obligations, just like we bear obligations to protect human soldiers, when feasible. But, as argued above, we will be able to make these robots insensitive to pain and they will have limited projects and aims, meaning they will have reduced, or no, moral status. Thus, deploying them will be preferable to using human soldiers in war and there will be no moral barrier to doing so.

Can such robots be appropriate subjects of punishment? The upshot of the Design and Plausibility arguments is that future autonomous robots can and will be designed so that they are sensitive to moral considerations, as well as moral critique and blame. So, within the limited domain in which they are trained to operate, they will be morally responsible—they will deserve blame when they act wrongly—for what they do. Again, the robots under discussion are those that will guide their actions via knowledge they have acquired, thereby displaying the capacity for choice and intentional action (B. F. Malle and Scheutz 2015). Such robots will have the ability to learn from experience and thereby form and revise internal representations—their beliefs, desires, and values—themselves (Sparrow 2007, 65).

Suppose we have a robot that responds to moral criticism, social rejection, and the blame of relevant human interlocutors. It is capable of machine guilt and modifying its representations in response. Then, I hold that it would be morally responsible for its conduct in its domain of operation, deserving blame for its wrongful conduct.[15] Why? Although there are different conceptions of moral responsibility, one prominent account holds that agents are morally responsible just in case they deserve blame or credit for actions they perform (Feinberg 1970; Zimmerman 1988; Pereboom 2001, 2008, 2014; Bennett 2002; Strawson 2002; Sommers 2007; McKenna 2012). So, determining when someone is morally responsible requires examining what it is to deserve blame.

In general, someone deserves blame when blame's psychological functions are appropriately directed at that person (Cogley 2013, 2016). So, for example, since one function of blame is to appraise actions as wrong, blame aimed at an actual wrongdoer is apt and thus deserved by the wrongdoer. A robot that can perform wrongful actions can thus deserve blame in this sense. Another function of blame is to communicate to the wrongdoer that her act was wrongful with the aim of her acknowledging fault (Walker 2006; Darwall 2006; Smith 2007; D. Shoemaker 2007; Macnamara 2013). Thus, blame is felicitous and therefore deserved when aimed at a wrongdoer capable of acknowledging her conduct was wrong and giving interpersonal expression to that fact. A robot that can feel machine guilt in response to the blame of others and can inform them that it is modifying its representations in response can also deserve blame in this sense. So, an autonomous robot with the capacities in question will be morally responsible for its conduct because it will deserve blame for what it does.

We need to now ask a question Sparrow does not. Does a robot's moral responsibility entail that we have reason to punish it? In considering this question, we should canvas some of the standard reasons we offer for punishing human beings. Punishment has been defended by citing its deterrent effects (Farrell 1985, 1995; Ellis 2003), that it can help restore trust to a society (Dimock 1997), communicate condemnation of what has been done (Duff 2001), or help teach others that such acts are not to be done (Hampton 1984). The hope is that punishment of one agent serves to produce good effects, either in that agent, other agents, or for society in general. Could punishing a robot produce these same effects? Yes—so long, of course, as these effects are actually produced by punishing humans.

Recall that, for the robot, blame and social criticism provide data that will help it better navigate the social world. Punishment is an additional source of important information about what acts should not be done, so autonomous robots should additionally be engineered to learn from the punishment of themselves, other humans, and other robots. Additionally, human beings have active agency detection modules in our brains that lead us to attribute agency even when it is not present (Atran 2002; Boyer 2002). Already, soldiers give the non-autonomous robots they work with names, like "Boomer," and see them as saving lives and having distinct personalities (Garber 2013). This means that humans working alongside robots that do, in fact, possess agency will also attribute agency to the robots. (Recall that most human subjects interacting with Robovie held this sham autonomous robot accountable.) Humans interacting with future autonomous robots in social situations will treat them as moral agents to the same degree they treat human beings with comparable capacities as ethical partners. Thus, to the extent that deterrence, restoring trust, communicating condemnation, or providing education provide good reasons for punishing human agents, they also provide reasons to punish autonomous robots.

[15] Building on what I said above in FN 14: punishment should be assigned to all wrongful actors commensurate with their responsibility. If programmers/commanders have acted wrongfully, they can be on the hook, too. I say more about the considerations that should guide assigning commensurate punishment, below.

More interesting, for our purposes, are reasons for punishment based directly on the capacities of the putative subjects of punishment. As noted above, punishing robots can lead the robots, themselves, to update their representations of situations, leading them to be educated so that they will not commit similar acts in the future. At this point, we can anticipate Sparrow objecting that if such machines do not have the capacity to be significantly harmed they are still not appropriate subjects of punishment, even if they can be morally responsible, blameworthy, and respond effectively to blame. But we should now ask why it is important that those we punish be harmed. One reason may be strictly definitional. Nothing that we do to a person that fails to cause harm could plausibly count as punishment, proper (Boonin 2008; Bedau and Kelly 2015; Duff and Hoskins 2017). I admit this conceptual point. My interest is in whether we have reason to do something punishment-like to autonomous robots. Their not being moral patients may mean we can't, strictly speaking, *punish* them. But we can do the very same sort of things to them—destroying or disabling them in a way that expresses condemnation of their actions—as we do to human beings. Call these kinds of things, when done to an autonomous robot, punishment*. Would we still have reason to punish* an autonomous robot, even if we were punishing* an agent we couldn't seriously harm?

One reason that we punish human beings may be that we hope to produce a certain kind of harm—the *pain* of guilt—that is necessary for the moral education of offenders so they do not behave similarly in the future. If so, certain harms are necessary in human beings for other good effects we truly want out of punishment. Given my suppositions about the capacities of future autonomous robots, however, these harms are not necessary for their moral improvement. But, if the robots are engineered to be sensitive to their punishment*. and possess machine guilt, these good effects we want out of human punishment are still possible with robot punishment*.

Sparrow does acknowledge that autonomous robots with internal desire-like states can be harmed in one way: by preventing them from acting as those desire-like states prompt. His skepticism about machine punishment stems from doubting that it will be able to experience pain in a manner that is morally compelling for us. Rather than skepticism, this is a reason to think that punishing* machines does not raise the same serious moral concerns of human punishment. If we accept harm in the case of human punishment because it is necessary for getting the good effects of punishment and we can or do make autonomous robots that have the same moral functionality but lack other abilities to be harmed, so much the better. If we additionally assume that future autonomous robots will be better at fighting and less susceptible to damage, we would have reason to deploy them.

Another possibility is that, in punishing human beings, the reason we care about the subject being harmed is just that we want to *hurt* the wrongdoer—we enjoy making agents who have caused harm experience pain. Without the ability to hurt the robot significantly, or only being able to harm it to some lesser degree, that desire might be frustrated. From a moral perspective, however, so much the worse for such desires. If we discover that we have no reason to punish* autonomous robots because we cannot satisfy sadistic desires, that should lead us to question our justification for punishing human beings. It should not lead us to think it is a substantive ethical concern regarding deploying autonomous robots in combat.

In sum, then, so much the better for the morally laudatory or defensible reasons we accept harming those we punish and so much the worse for the morally suspect ones. If human punishment is reasonable and ethically defensible, punishing* autonomous robots will be reasonable and ethically defensible. This is because there will be a lower ethical bar to punishing* such robots, as they will not be the moral equals of human soldiers we punish. And, if punishing human soldiers really does secure important goods, those goods can also be secured by punishing* robots. We should acknowledge, though, that reflecting on whether it makes any sense to punish* robots might prompt us to reexamine our human punishment practices in interesting, challenging, or helpful ways. We should not just assume, as Sparrow seems to, that our current punishment practices are morally in the clear. If it is not reasonable or ethically defensible to punish* autonomous robots, we should look hard at whether it is reasonable and ethically defensible to punish human beings.

7. Conclusion

I've here explored requisite future robot moral capacities, examined two ways the no subject of punishment worry has been developed, surveyed extant replies to the objection, and finally argued that we will have reasons to punish* future autonomous robots. Alternatively, if we lack such reasons, that means we should doubt the reasonability of punishing human beings. I suspect, however, that reservations about my argument may remain.

Perhaps some worries may be assuaged by emphasizing, again, that my concern in this paper has been with autonomous military robots of the future. My arguments do not bear on the appropriateness of deploying the current 'automatic' weaponry we have, or the kinds of more advanced systems we are likely to have in the near future. Simpson and Müller are correct that the 'reasonable risk' framework is sufficient for such machines. The machines under consideration by my argument will be cognitively sophisticated enough that it will make sense to trust them with decisions about how to attack, whether to attack, and when to disengage. Sparrow and the others pressing the no subject of punishment objection are right to think that there is a serious objection to deploying such robots, but they fail to accurately locate it. What would be objectionable is deploying robots that are *insensitive* to moral considerations and blame in situations where moral discriminations must be made in order to fight justly. For example, if morally innocent people will be present in a particular military theater, it would be at least a significant pro tanto wrong to deploy a robot that cannot recognize a person's innocence and take it as a stringent consideration against attacking her. Because of the inevitability of error, it would also be wrong to deploy a robot in such contexts that can make moral discriminations but lacks machine guilt, and so has no capacity to update its representations.

We rightly find the idea of cognitively sophisticated agents insensitive to moral considerations terrifying. Indeed, it is this very possibility that makes psychopaths so unnerving (K. Gray and Wegner 2012). Philosophers have also taken up this theme, arguing that it would be wrongful to deploy such machines (Purves, Jenkins, and Strawser 2015). I concur. But I part ways with theorists who hold that it is impossible to develop robots that are sensitive to moral considerations and moral blame. Developing robots that respond appropriately to moral wrongdoing, blame, and punishment would allow us to secure the moral goods of human punishment.

Why, we might finally ask, should we program a future robot so that punishing* it causes machine guilt and thus leads it to revise its representations? Why not program the robots, instead, so that when we determine they have done wrong and communicate that fact to them, the robots immediately suffer machine guilt and revise their representations? Which of these possibilities we go for depends on how committed we are to a practice that looks very much like how we punish humans, or whether we are open to other practices that accomplish the same aims but deviate from the standard script. Autonomous robots of the future will be able to be morally blameworthy and hence deserving of punishment*. But we might design them so we don't actually need to punish* them. So long, however, as they can be deserving of punishment*, and punishing* them can secure the important moral goods that human punishment does, the no subject of punishment worry can be successfully addressed.

To see this, return to the fundamental moral issue animating Sparrow's version of the objection: being just to our enemies. As he puts it, "The least we owe our enemies is allowing that their lives are of sufficient worth that someone should accept responsibility for their deaths" (Sparrow 2007, 67). Punishing deserving human soldiers for war crimes both helps us take responsibility for the misconduct of the soldiers, as well as demonstrates to our enemies that, though we are adversaries, we take still take their lives seriously. Creating autonomous robots that deserve blame and punishment* when they act wrongly and then actually punishing* the robots when they commit war crimes confirms that we still take sufficient responsibility for enemy deaths. In this case, we do it by deploying robots that have been engineered responsibility, can be morally responsible for their conduct, and deserve punishment* when they commit serious wrongs.

Bibliography

Anthony, Sebastian. 2017. "DeepMind's AlphaGo Takes on World's Top Go Player in China." Ars Technica. April 10, 2017. https://arstechnica.com/information-technology/2017/04/deepmind-alphago-go-ke-jie-china/.

Arkin, Ronald Craig, Patrick Ulam, and Alan R. Wagner. 2012. "Moral Decision Making in Autonomous Systems: Enforcement, Moral Emotions, Dignity, Trust, and Deception." *Proceedings of the IEEE* 100 (3): 571–589.

Arnold, Thomas, Daniel Kasenberg, and Matthias Scheutz. 2017. "Value Alignment or Misalignment -- What Will Keep Systems Accountable?" In *AAAI Workshop on AI, Ethics, and Society*.

Atran, Scott. 2002. *In Gods We Trust: The Evoluntionary Landscape of Religion*. Oxford: Oxford University Press.

Aydede, Murat. 2013. "Pain." The Stanford Encyclopedia of Philosophy. Spring Edition 2013. https://plato.stanford.edu/archives/spr2013/entries/pain/.

Baumeister, Roy, K. D Vohs, C. Nathan DeWall, and Liqing Zhang. 2007. "How Emotion Shapes Behavior: Feedback, Anticipation, and Reflection, Rather than Direct Causation." *Personality and Social Psychology Review* 11 (2): 167–203.

Bedau, Hugo Adam, and Erin Kelly. 2015. "Punishment." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2015. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2015/entries/punishment/.

Bennett, Christopher. 2002. "The Varieties of Retributive Experience." *The Philosophical Quarterly* 52 (207): 145–163.

Boonin, David. 2008. *The Problem of Punishment*. Cambridge, Mass.: Cambridge University Press.

Boyer, Pascal. 2002. *Religion Explained*. London: Vintage.

Clarke, Randolph. 2016. "Moral Responsibility, Guilt, and Retributivism." *The Journal of Ethics* 20 (1–3): 121–37. https://doi.org/10.1007/s10892-016-9228-7.

Cogley, Zac. 2013. "Basic Desert of Reactive Emotions." *Philosophical Explorations* 16 (2): 165–77.

———. 2016. "Basic Desert of Reactive Emotions." In *Basic Desert, Reactive Attitudes and Free Will*, edited by Maureen Sie and Derk Pereboom, 69–81. New York: Routledge.

Damasio, Antonio R. 2006. *Descartes' Error*. New York: Penguin.

Danaher, John. 2016. "Robots, Law and the Retribution Gap." *Ethics and Information Technology* 18 (4): 299–309. https://doi.org/10.1007/s10676-016-9403-3.

Darwall, Stephen. 2006. *The Second-Person Standpoint: Morality, Respect, and Accountability*. Cambridge, Mass.: Harvard Univ Press.

Dennett, Daniel C. 1997. "When HAL Kills, Who's to Blame?: Computer Ethics." In *HAL's Legacy: 2001's Computer as Dream and Reality*, edited by D.G. Stork, 351–65. Cambridge, Mass.: MIT Press.

Dimock, Susan. 1997. "Retributivism and Trust." *Law and Philosophy* 16 (1): 37–62. https://doi.org/10.1023/A:1005765126051.

Dolinko, David. 1991. "Some Thoughts About Retributivism." *Ethics* 101 (3): 537–59.

Duff, Antony. 2001. *Punishment, Communication, and Community*. Oxford University Press, USA.

Duff, Antony, and Zachary Hoskins. 2017. "Legal Punishment." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2017. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2017/entries/legal-punishment/.

Ellis, Anthony. 2003. "A Deterrence Theory of Punishment." *The Philosophical Quarterly* 53 (212): 337–351. https://doi.org/10.1111/1467-9213.00316.

Farrell, Daniel M. 1985. "The Justification of General Deterrence." *The Philosophical Review* 94 (3): 367–94. https://doi.org/10.2307/2185005.

———. 1995. "Deterrence and the Just Distribution of Harm." *Social Philosophy and Policy* 12 (02): 220–40. https://doi.org/10.1017/S0265052500004738.

Feinberg, Joel. 1970. *Doing & Deserving; Essays in the Theory of Responsibility*. Princeton, NJ. http://philpapers.org/rec/FEIDD.

Garber, Megan. 2013. "Funerals for Fallen Robots - The Atlantic." *The Atlantic*, September 20, 2013. https://www.theatlantic.com/technology/archive/2013/09/funerals-for-fallen-robots/279861/.

Giner-Sorolla, Roger, and Pablo Espinosa. 2010. "Social Cuing of Guilt by Anger and of Shame by Disgust." *Psychological Science*, December. https://doi.org/10.1177/0956797610392925.

Goodall, Noah. 2014. "Ethical Decision Making during Automated Vehicle Crashes." *Transportation Research Record: Journal of the Transportation Research Board*, no. 2424: 58–65.

Gray, Heather M., Kurt Gray, and Daniel M. Wegner. 2007. "Dimensions of Mind Perception." *Science* 315 (5812): 619–619. https://doi.org/10.1126/science.1134475.

Gray, Kurt, and Daniel M. Wegner. 2009. "Moral Typecasting: Divergent Perceptions of Moral Agents and Moral Patients." *Journal of Personality and Social Psychology: Attitudes and Social Cognition* 96 (3): 505–20. http://dx.doi.org/10.1037/a0013748.

———. 2012. "Feeling Robots and Human Zombies: Mind Perception and the Uncanny Valley." *Cognition* 125 (1): 125–30. https://doi.org/10.1016/j.cognition.2012.06.007.

Hampton, Jean. 1984. "The Moral Education Theory of Punishment." *Philosophy & Public Affairs* 13 (3): 208–38. https://doi.org/10.2307/2265412.

Horgan, Terry, and Matjaž Potrč. 2010. "The Epistemic Relevance of Morphological Content." *Acta Analytica* 25 (2): 155–73. https://doi.org/10.1007/s12136-010-0091-z.

Horgan, Terry, and Mark Timmons. 2007. "Morphological Rationalism and the Psychology of Moral Judgment." *Ethical Theory and Moral Practice* 10 (3): 279–95. https://doi.org/10.1007/s10677-007-9068-4.

Jack, Anthony I., and Philip Robbins. 2012. "The Phenomenal Stance Revisited." *Review of Philosophy and Psychology* 3 (3): 383–403. https://doi.org/10.1007/s13164-012-0104-5.

Julius, David, and Allan I. Basbaum. 2001. "Molecular Mechanisms of Nociception." *Nature* 413 (6852): 203–210.

Kahn Jr, Peter H., Takayuki Kanda, Hiroshi Ishiguro, Brian T. Gill, Jolina H. Ruckert, Solace Shen, Heather E. Gary, Aimee L. Reichert, Nathan G. Freier, and Rachel L. Severson. 2012. "Do People Hold a Humanoid Robot Morally Accountable for the Harm It Causes?" In *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, 33–40. ACM. http://dl.acm.org/citation.cfm?id=2157696.

Kershnar, Stephen. 2013. "Autonomous Weapons Pose No Moral Problem." *Killing by Remote Control: The Ethics of an Unmanned Military*, 229–245.

Lucas, G. R. 2013. "Engineering, Ethics, and Industry: The Moral Challenges of Lethal Autonomy." *Killing by Remote Control: The Ethics of an Unmanned Military. Oxford University Press, Oxford*, 211–228.

Macnamara, Coleen. 2013. "'Screw You!' & 'Thank You.'" *Philosophical Studies* 163 (3): 893–914.

Malle, B. F., and M. Scheutz. 2015. "When Will People Regard Robots as Morally Competent Social Partners?" In *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 486–91. https://doi.org/10.1109/ROMAN.2015.7333667.

Malle, Bertram F. 2016. "Integrating Robot Ethics and Machine Morality: The Study and Design of Moral Competence in Robots." *Ethics and Information Technology* 18 (4): 243–56. https://doi.org/10.1007/s10676-015-9367-8.

Matthias, Andreas. 2004. "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata." *Ethics and Information Technology* 6 (3): 175–83. https://doi.org/10.1007/s10676-004-3422-1.

McKenna, Michael. 2012. *Conversation and Responsibility*. New York: Oxford University Press.

McKenna, Michael, and Derk Pereboom. 2016. *Free Will: A Contemporary Introduction*. New York: Routledge.

Morris, Herbert. 1976. "Guilt and Suffering." In *On Guilt and Innocence: Essays in Legal Philosophy and Moral Psychology*. Berkeley: Univ of California Press.

Pereboom, Derk. 2001. *Living Without Free Will*. Cambridge: Cambridge University Press.

———. 2008. "A Hard-Line Reply to the Multiple-Case Manipulation Argument." *Philosophy and Phenomenological Research* 77 (1): 160–70.

———. 2014. *Free Will, Agency, and Meaning in Life*. New York: Oxford University Press.

Purves, Duncan, Ryan Jenkins, and Bradley J. Strawser. 2015. "Autonomous Machines, Moral Judgment, and Acting for the Right Reasons." *Ethical Theory and Moral Practice* 18 (4): 851–72. https://doi.org/10.1007/s10677-015-9563-y.

Robbins, Philip, and Anthony I. Jack. 2006. "The Phenomenal Stance." *Philosophical Studies* 127 (1): 59–85. https://doi.org/10.1007/s11098-005-1730-x.

Robertson Jr, Horace B. 1996. "The Obligation to Accept Surrender." *International Law Studies* 68 (1): 6.

Roff, Heather M. 2013. "Responsibility, Liability, and Lethal Autonomous Robots." R*outledge Handbook of Ethics and War: Just War Theory in the 21st Century. Routledge*, 352–364.

Schneider, Susan, and Edwin Turner. n.d. "Is Anyone Home? A Way to Find Out If AI Has Become Self-Aware." Scientific American Blog Network. Accessed September 29, 2017. https://blogs.scientificamerican.com/observations/is-anyone-home-a-way-to-find-out-if-ai-has-become-self-aware/.

Shoemaker, David. 2007. "Moral Address, Moral Responsibility, and the Boundaries of the Moral Community." *Ethics* 118 (1): 70–108.

Shoemaker, David W. 2003. "Caring, Identification, and Agency." *Ethics* 114 (1): 88–118.

Silver, David, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, et al. 2016. "Mastering the Game of Go with Deep Neural Networks and Tree Search." *Nature* 529 (7587): 484–489.

Simpson, Thomas W., and Vincent C. Müller. 2016. "Just War and Robots' Killings." *The Philosophical Quarterly* 66 (263): 302–22. https://doi.org/10.1093/pq/pqv075.

Smith, Angela M. 2007. "On Being Responsible and Holding Responsible." *The Journal of Ethics* 11 (January): 465–84.

Sommers, Tamler. 2007. "The Objective Attitude." *The Philosophical Quarterly* 57 (July): 321–41.

Sparrow, Robert. 2007. "Killer Robots." *Journal of Applied Philosophy* 24 (1): 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x.

Stahl, Bernd Carsten. 2004. "Information, Ethics, and Computers: The Problem of Autonomous Moral Agents." *Minds and Machines* 14 (1): 67–83.

Strawson, Galen. 2002. "The Bounds of Freedom." In *The Oxford Handbook of Free Will*, edited by Robert Kane, 441–60. New York: Oxford University Press.

Walker, Margaret. 2006. *Moral Repair*. Cambridge: Cambridge University Press.

Wegner, Daniel M., and Kurt Gray. 2016. *The Mind Club*. New York, NY: Viking.

Zimmerman, Michael J. 1988. *An Essay on Moral Responsibility*. Totowa, NJ: Rowman & Littlefield.