The Ethics of Killer Robots

https://www.e-ir.info/2014/07/23/the-ethics-of-killer-robots/

RYAN JENKINS, JUL 23 2014

For many, autonomous weapons or 'killer robots' are the stuff of science fiction, the stuff of nightmares, or both. Autonomous weapons, though, may soon become fact. Thankfully, careful ethical explorations are now underway to prevent killer robots from ushering in the dystopian horror that has stirred Hollywood and others for decades – notably in 2001: A Space Odyssey, The Terminator, I, Robot, The Matrix, and Virtuosity. These discussions are the topic of this article – is there anything morally wrong with deploying a killer robot in war?

Defining Autonomous Robots

We should first be clear on what we mean by autonomous weapons. To do this, we must distinguish them from the term 'drone,' which has been in popular use for some time. 'Drone' implies an autonomy that the drones in use today do not have. The most famous drones in the US arsenal – the Predator and Reaper drones – have very low levels of autonomy, and most importantly, a human remains in the loop for lethal decisions. This means there is always a person making the decision to engage a target. Thus, the Predator and the Reaper are more properly called uninhabited aerial vehicles (UAVs) or remotely piloted aircraft (RPAs). The autonomous robots we are concerned with here are those *without* a human in the loop: robots that are free to choose their targets and fire on humans on their own accord.

Some weapons already demonstrate rudimentary forms of autonomy. For example, there are cruise missiles that can direct their own flight over terrain through radar guidance. Demonstrating more advanced levels of autonomy, some cruise missiles can choose their own targets from within a geographic territory by comparing what they see against preloaded target signatures. Finally, some are equipped with a variety of ordinance packages that they can choose from when finally engaging their target (see Sparrow, 2007:63 for a survey of weapons with limited autonomy under development, including the US Air Force's Low Cost Autonomous Attack System).

Spurred by advances in computer vision and processing power (see Adams, 2001 for a discussion of how computer decision-making is coming to replace human decision-making in warfare), we may soon see the development offully autonomous weapons systems (AWS): weapons that are capable not only of flying on their own, but also of*choosing and engaging their own targets*, without any human input. The actions they perform can be said to have originated within the machines themselves, rather than being guided from outside. They are capable of setting their own goals or, more philosophically, they are capable of *legislating their own ends* to themselves (Kant, 2012).

Creating fully autonomous weapons is a daunting task. If they turn out to be technically unfeasible, then questions about the ethics of their deployment will be moot. However, it seems the current advances in technology may be inexorable, instead lending these moral questions a real urgency.

Three Kinds of Mistakes

Set aside for the moment the technical difficulty of constructing an autonomous weapons system and consider the morality of their use. Many people would find their use unacceptable because they would be worried that autonomous weapons might make serious mistakes. It will be useful to frame our discussion as an examination of the kinds of mistakes that autonomous weapons could make – that is, the many ways they could go wrong. By evaluating the

likelihood of each of these kinds of mistakes, we can either sharpen our critique of AWS or assuage popular fears.

Any time an autonomous weapon makes the decision to engage a target, it makes two judgments: (1) certain targets are legitimate targets and (2) the object it is looking at is one of those targets. It then infers the conjunction that the object it is looking at is a legitimate target. Finally, it engages the target, at which point it could make another mistake and fail to achieve its goal. There are, then, *three ways* in which an autonomous weapon can make a mistake. Let's use the following language: mistakes in identifying targets are *empirical mistakes*; mistakes in judging a target to be a legitimate target are *moral mistakes*; and mistakes in executing this judgment are *practical mistakes*. This threefold distinction is discussed in Purves, Jenkins, and Strawser (manuscript).

Let's begin with empirical mistakes, that is, a robot being in error about what it 'sees.' These are mistakes in making a judgment of fact. For example, an autonomous weapon could mistake a reporter carrying a camera for a soldier carrying a rocket-propelled grenade. Currently, there are significant technical difficulties in developing computer vision: the cameras and processing architectures that would enable a machine to see reliably. Empirical mistakes will become less and less likely as this technology improves. Surely, it would be wrong to deploy autonomous weapons unless we could be confident in their abilities to at least make sense of the world. (This is even *before* we trust them to make complex moral decisions.) In fact, we should hope that autonomous weapons would not be deployed unless they would be excellent – *at least as good as a typical soldier* – in this capacity.

Could a Robot Ever Act Morally?

The second kind of error is the one that probably underlies most of the worries about autonomous weapons: moral errors. These are mistakes in moral reasoning, separate from mistakes in judging the facts of a situation. For example, a machine that judged a soldier to be surrendering, and *then* judged that he was still a legitimate target would be making a moral error. The malicious machines in *The Terminator* or *The Matrix* correctly judge humans to be humans, and then kill them *because* they are humans, thus making a grave moral mistake. There are plenty of reasons why we might harbor qualms about machines' moral reasoning abilities.

First of all, we might think that morality is just not the kind of thing that can be codified in strict rules. This tradition in moral philosophy – rather uncreatively termed the non-codifiability thesis – stretches back to Aristotle and boasts supporters with a variety of views (Mill, 1863; McDowell, 1979; McNaughton, 1988; Little, 1997; Ross, 2002; McKeever and Ridge, 2005). While this thesis is controversial, it's clear why many find it plausible: for many people, morality is something that has to be felt or intuited, or at least has an essential affective component. If that's right, it seems like computers could never navigate the moral universe as well as (some) humans do (Dreyfus, 1992:199). There could be no computer sages or saints of the likes of the Buddha or Martin Luther King, Jr.

Another reason we might be disturbed by the possibility of a robot making moral mistakes is that there would be no one (or no thing) to hold accountable for that mistake. In perhaps the best-known philosophical argument against autonomous weapons, Sparrow (2007) argues that this problem makes deploying autonomous weapons immoral. Typically, in warfare, we are concerned with holding responsible those who perpetrate intentional wrongdoing. (Keep in mind that we're supposing AWS *are* capable of making intentional choices.) This adherence to the rule of law is what distinguishes disciplined armies from rampaging barbarians. However, in deploying an autonomous weapon, there would be no one to hold accountable should it go rogue, or commit war crimes. It would be unfair to blame the programmers or the commanding officer, since they could not have predicted how the robot would act. This is just what it means for a robot to be autonomous! And it would be bizarre to punish the robot itself – what would that even look like? Thus, there would exist a 'responsibility gap' in assigning blame. If we are concerned with showing respect to enemies by maintaining the guise of accountability, we would be wrong to deploy killer robots.

It may ultimately be impossible to reliably safeguard the software that controls military robots: there may always be bugs, or the threat of hacking or hijacking. Do we have any safeguards against such 'morally bad' robots? Hagerott (2014) suggests regulating autonomous military robots' *hardware* instead. In particular, he suggests regulating the size and fuel source of robots: this way, a hijacked or rogue robot could never be so small as to be practically undetectable, and we could be confident that a rogue robot would eventually run out of fuel, rather than continue its

erratic behavior indefinitely.

What If Autonomous Weapons Could Be Perfect?

Maybe, however, this is too quick. Maybe computers could learn to reliably imitate the moral decisions that humans make. Even if we think, for example, that IBM's Watson computer does not *understand* language, at least it can follow complex heuristics, or shortcuts, in order to make decisions about language as effectively and convincingly as a human. The original Watson 'learned' by reading the Internet. Perhaps there could be a *moral* Watson, trained by incorporating the aggregate moral experience of humanity. From there, perhaps it could guess how the wisest humans might react to hypothetical or unfamiliar cases.

If this were true, then we would have no good reason to worry about autonomous weapons making the wrong decisions or rebelling against their creators. And if autonomous weapons were morally perfect, we would never have a reason to assign blame for the bad decisions they would make, so we would never have to worry about responsibility gaps.

Are there any objections to autonomous weapons if we suppose they are as good as – or even *better* than – humans at making moral decisions?

Many moral philosophers believe that whether a person acts rightly or wrongly depends in part on her intentions. For example, there is a big difference between the saintly philanthropist who lovingly donates her personal fortune to charity, and the grudging miser who donates the same amount while hating every second of it. So, as many people think, if we really want to do the right thing, maybe we need to have the right intentions or the right emotional state when we're acting. If that's right, then autonomous weapons may fail this moral test. This is because, even assuming AWS can perfectly imitate human moral behavior, they would never be acting with *the right state of mind* (see, for example, St. Augustine's Letter 189 to Boniface, 2004). As an analogy, many people would find it problematic to deploy a sociopathic soldier – one who kills mechanically, without any emotional response or empathy for her victims. Deploying autonomous weapons would be just as unsettling. This argument appears in Purves, Jenkins, and Strawser (manuscript).

In a similar vein, we might worry that autonomous weapons cannot show respect to their enemies (Sparrow, manuscript; O'Connell, 2014; Nagel, 1972:172). While it may seem odd to say that there is something essentially respectful about killing another human being, this kind of killing might require at least 'acknowledging the morally relevant features that render them combatants or otherwise liable for being subject to a risk of being killed' (Sparrow, manuscript). It means understanding and acknowledging that a soldier has made a choice that makes him a soldier rather than a noncombatant. Since AWS cannot make any such acknowledgement, their killings would be profoundly disrespectful.

This category of mistake – moral mistakes – is central to determining the permissibility of autonomous weapons. The arguments in this section are attempts at crystallizing the nebulous unease that strikes many people when considering killer robots. We should consider whether any of these reasons constitutes a significant moral difference between a killer robot and a soldier that execute the very same task.

Practical Mistakes

A practical mistake is a mistake in executing a decision once it's been made. A machine that correctly identified a soldier, and correctly judged that soldier to be a legitimate target, but then misjudged her trajectory and fired a missile into an orphanage would have made a practical mistake. Practical mistakes, like empirical mistakes, are a mere technical difficulty. We should expect that machines would make fewer practical mistakes as technology improves.

Practical mistakes are problematic because they inhibit an autonomous weapon from doing the right thing. However, note that this worry cuts both ways: some practical mistakes could be good because they could prevent a machine

from successfully executing what *would* have been the wrong action. Thus, the moral importance of practical mistakes might be a wash.

In the near future, robots might be slightly worse at making moral decisions than humans, but much betterat executing those decisions. In this case, deploying an autonomous weapon would be like deploying a skilled sharpshooter who occasionally chooses the wrong target. If only the consequences of our actions matter, and autonomous weapons are overall better than the typical soldier, then we would have an obligation to deploy them.

Keep in mind that policymakers and commanders have an obligation to protect their own soldiers, as long as they are carrying out a just mission. Strawser (2010) points out, correctly, that it is wrong for commanders to subject their soldiers to unnecessary risk unless they have a good reason. As an analogy, Strawser asks you to imagine you are a police chief charged with defusing a bomb. If you have a perfectly good robot bomb defuser, it would be wrong to deploy your human subordinate instead, precisely because it would subject him to unnecessary risk. Thus, if autonomous weapons turn out to be roughly as reliable as human soldiers, we may have obligations to protect our human soldiers by deploying autonomous weapons in their place. (This is separate from the clear strategic advantage of conducting missions without risking human life.)

Would Deploying Killer Robots Be Unfair?

Many people are concerned that deploying autonomous weapons would be unfair or dishonorable. Asymmetrical war is nothing new, but AWS seem to represent an altogether new kind of warfare – one where our enemies are totally helpless to fight back, where our actions more closely resemble 'pest control' than war (Steinhoff, 2006:7).

But, if our cause for going to war is justified, it's unclear why we should choose to make war more difficult on ourselves than it has to be. This is like choosing to forego cruise missiles and sending in human soldiers instead when there's no obvious benefit. It may be that autonomous weapons, like cruise missiles and artillery, are 'merely an extension of a long historical trajectory of removing a warrior ever farther from his foe for the warrior's better protection' (Strawser, 2010:343, in the context of UAVs).

Would Deploying Killer Robots Make War Too Easy?

Finally, we might worry that deploying autonomous weapons might make warfare too easy, or too tempting. This is because autonomous weapons, like the drones currently in use, lower the political costs and the psychological threshold for going to war. People who appeal to this objection, the threshold objection, worry that new technologies will make war seem so precise, easy, or sanitary that it becomes much more common. This is perhaps happening today with a greater reliance on Predators and Reapers in ongoing small-scale conflicts around the globe. Ultimately, though, this objection relies on weighing a potential cost against a known benefit (Strawser, 2010:358–361).

Conclusion

Autonomous weapons offer clear advantages while simultaneously raising powerful, but somewhat nebulous, moral concerns. For example, many find it dubious that asymmetrical war is unjust, or that soldiers have to *respect* their enemies or act with good intentions. We may ultimately have to decide how to weigh these vague moral concerns against the concrete promise of improving the outcomes of warfare in terms of successful missions and discrimination.

The line between peace and war is being increasingly blurred, and the advancing technology and tempo of warfare create powerful incentives for armies to automate killing. The technology behind these weapons may soon render them as effective as a typical soldier. In that case, our task would be to carefully examine the moral importance of the mental attitude of the soldier – or at least the *thing* – doing the killing. The answer to this question could ultimately determine the permissibility of autonomous weapons.

References

Adams, Thomas. 2001. 'Future Warfare and the Decline of Human Decision-making'. *Parameters: US Army War College Quarterly* (Winter, 2001-2): 57-71.

Augustine. 2004. "Letter 189 to Boniface." In *Letters 156-210 (II/3) Works of Saint Augustine*. New York: New City Press

Dreyfus, Hubert L. 1992. *What computers still can't do: a critique of artificial reason*. MIT Press.

Hagerott, Mark. 2014. 'The Cyber-Robotic Revolution and Implications for the Future: Offering a Framework and Suggestions'. Presentation at 'Cyberwarfare, Ethics, and International Humanitarian Law' workshop. International Committee for the Red Cross. Geneva, Switzerland.

Kant, Immanuel. 2012. *Groundwork of the Metaphysics of Morals*. Mary Gregor and Jens Timmerman (trans.). Cambridge: Cambridge University Press.

Little, Margaret. 1997. 'Virtue as Knowledge: Objections from the Philosophy of Mind'. *Nous* 31(1): 59-79.

McDowell, John. 1979. 'Virtue and Reason'. *The Monist* 62(3): 331-350.

McKeever, Sean, and Michael Ridge. 2005. 'The Many Moral Particularisms'. *Canadian Journal of Philosophy* 35.1 (2005): 83-106.

McNaughton, David. 1988. *Moral Vision*. Wily-Blackwell.

Mill, John Stuart. 1863. *Utilitarianism*. London: Parker, Son, and Bourn.

Nagel, Thomas. 1972. 'War and Massacre'. *Philosophy and Public Affairs* 1:123-144.

O'Connell, M. E. 2014. 'Banning Autonomous Killing', in *The American Way of Bombing: How Legal and Ethical Norms Change,* ed. M. Evangelista and H. Shue. Cornell University Press: Ithaca, NY.

Purves, Duncan, Ryan Jenkins, and Bradley Strawser. 'Autonomous Machines, Moral Judgment, and Acting for the Right Reasons'. Unpublished manuscript.

Ross, W. D. 2002. *The Right and the Good*. Oxford: Oxford University Press.

Sparrow, Robert. 2007. 'Killer Robots'. *Journal of Applied Philosophy* 24.1: 62-77.

Sparrow, Robert. 'Robots and Response: Assessing the Case against Autonomous Weapon Systems'. Manuscript.

Steinhoff, Uwe. 2006. 'Torture: The Case for Dirty Harry and against Alan Dershowitz'. *Journal of Applied Philosophy*, 23.3: 337–353.

Strawser, Bradley Jay. 2010. 'Moral Predators: The Duty to Employ Uninhabited Aerial Vehicles'.*Journal of Military Ethics* 9.4: 342-368.

About the author:

Ryan Jenkins has a PhD in philosophy from the University of Colorado Boulder and will begin teaching at California Polytechnic State in San Luis Obispo, California in January. He studies military ethics and the ethics of emerging

technologies.