27

Morally Repugnant Weaponry?

Ethical Responses to the Prospect of Autonomous Weapons

Alex Leveringhaus

I. INTRODUCTION

In 2019, the United Nations (UN) Secretary General *Antonio Guterres* labelled lethal autonomous weapons 'as political unacceptable and morally repulsive'.[1] 'Machines', *Guterres* opined, 'with the power and discretion to take lives without human involvement are politically unacceptable, morally repugnant and should be prohibited by international law'.[2] The Secretary General's statement seems problematic. Just because something is morally repugnant does not entail that it should be banned by law. Further, it is not clear what exactly renders autonomous weapons systems (AWS hereinafter) morally abhorrent.[3] The great danger is that statements such as the Secretary General's merely rely on the supposed 'Yuck' factor of AWS.[4] But Yuck factors are notoriously unreliable guides to ethics. While individuals might find things 'yucky' that are morally unproblematic, they might not be repulsed by things that pose genuine moral problems.

In response to the Secretary General's statement, the purpose of this chapter is twofold. First, it seeks to critically survey different ethical arguments against AWS. Because it is beyond the scope of this chapter to survey every ethical argument in this context, it outlines three prominent ones, (1) that AWS create so-called responsibility gaps; (2) that the use of lethal force by an AWS

[1] N Werkhauser, 'UN Impasse Could Mean Killer Robots Escape Regulation' DW News (20 August 2018) www.dw.com/en/un-impasse-could-mean-killer-robots-escape-regulation/a-50103038 (hereafter Werkhauser, 'Killer Robots').

[2] Secretary-General, *Machines Capable of Taking Lives without Human Involvement are Unacceptable, Secretary-General Tells Experts on Autonomous Weapons Systems* (United Nations Press Briefing, 25 March 2019), www.un.org/press/en/2019/sgsm19512.doc.htm.

[3] To avoid any misunderstanding at the outset, autonomy, in the debate on AWS, is not understood in the same way as in moral philosophy. Autonomy, in a moral sense, means to act for one's own reasons. This is clearly not the case in the context of AWS. These systems, as I point out shortly, require programming by a human individual. In quasi-Kantian parlance, then, AWS are heteronomous, rather than autonomous, in that they do not act for their own reasons. As I shall explain later, in the context of the debate on AWS, autonomy essentially describes a machine's capacity, once it has been programmed, to carry out tasks independently of, and without further guidance from, a human individual. This is, of course, not sufficient for moral autonomy in a meaningful sense. In the chapter, I use the term autonomy according to its technological meaning, rather than its moral one.

[4] The term Yuck factor describes a strong emotional reaction of revulsion and disgust towards certain activities, things, or states of affairs. The question is whether such visceral emotional responses are a reliable guide to ethics. Some activities or things – for example, in vitro meat or a human ear grown on a mouse for transplantation – might seem disgusting to some people, and sometimes this can indeed have normative significance. That being said, the feeling of disgust does not always explain why something is ethically undesirable. One problem is that our emotional responses are often shaped by social, economic, and political factors that can cloud our ethical judgement. Especially in the context of emerging technologies, the danger is that the Yuck factor might prevent the adoption of technologies that might be genuinely beneficial.

https://doi.org/10.1017/9781009207898.036 Published online by Cambridge University Press

is incompatible with human dignity; and (3) that AWS replace human agency with artificial agency. The chapter contends that neither of these arguments is sufficient to show that AWS are morally repugnant. Second, drawing upon a more realistic interpretation of the technological capacities of AWS, the chapter outlines three alternative arguments as to why AWS are morally problematic, as opposed to morally repugnant.

In the second part of the chapter, I write more about definitional issues in the debate on AWS. In the third part, I critically analyse, respectively, the notion of a responsibility gap, the relationship between AWS and human dignity, and role of human agency in war. In the fourth part, I outline a brief alternative account of why AWS might be morally problematic and explain how this intersects with other key issues in contemporary armed conflict.

Before I do so, I need to raise three general points. First, the chapter does not discuss the legal status of AWS. The focus of this chapter is on ethical issues only. The question whether, as suggested by the Secretary General, the alleged moral repugnancy of AWS justifies their legal prohibition is best left for a different occasion. Second, the chapter approaches AWS from the perspective of contemporary just war theory as it has developed since the publication of *Michael Walzer's* seminal "Just and Unjust Wars: A Moral Argument with Historical Illustrations" in 1977.[5] Central to *Walzer's* work, and much of just war theory after it, is the distinction between the normative frameworks of *jus ad bellum* (justice in the declaration of war) and *jus in bello* (justice in the conduct of war). As we shall see, the ethical debate on AWS has mainly been concerned with the latter, as it has tended to focus on the use of (lethal) force by AWS during armed conflict. Third, in addition to the distinction between *jus ad bellum* and *jus in bello*, *Walzer*, in *Just and Unjust Wars*, defends the distinction between combatants (who may be intentionally killed) and non-combatants (who may not be intentionally killed) during armed conflict. The former are typically soldiers, whereas the latter tend to be civilians, though he acknowledges the existence of grey zones between these categories.[6] In recent years, this distinction has come increasingly under pressure, with some theorists seeking to replace it with a different one.[7] For the sake of convenience and because these terms are widely recognised, the chapter follows *Walzer* in distinguishing between combatants and non-combatants. However, many of the issues highlighted in the following sections will also arise for theories that are critical of *Walzer's* distinction.

II. WHAT IS AN AUTONOMOUS WEAPON?

Here, I offer a fourfold attempt to define AWS. First, it is self-evident that AWS are weapons. In this sense, they differ from other forms of (military) technology that are not classifiable as weapons. The following analysis assumes that weapons have the following characteristics; (1) they were specifically designed in order to (2) inflict harm on another party.[8] Usually, the harm is achieved via a weapon's kinetic effect. The harmful kinetic effect is not random or merely a by-product of the weapon's operation. Rather, weapons have been intentionally designed to produce a harmful effect. Non-weapons can be used as weapons – you could stab me with the butterknife – but they have not been deliberately designed to inflict harm.

[5] M Walzer, *Just and Unjust Wars: A Moral Argument with Historical Illustrations* (5th ed. 2015) (hereafter Walzer, *Just and Unjust Wars*).

[6] Ibid, 145.

[7] See J McMahan, *Killing in War* (2009).

[8] J Forge, *Designed to Kill: The Case against Weapons Research* (2013).

https://doi.org/10.1017/9781009207898.036 Published online by Cambridge University Press

Second, as stated by Secretary General *Guterres*, the crucial feature of AWS, accounting for their alleged moral repugnancy, is that their kinetic and potentially lethal effect is created by the weapon without human involvement.[9] However, AWS will require initial mission programming by a human programmer. Hence, there will be human involvement in the deployment of an AWS. The point, though, is that once an AWS has been programmed with its mission parameters, the weapon is capable of operating without any further guidance and supervision by a human individual. Crucially, it can create a harmful and potentially lethal kinetic effect by delivering a payload without direct or real-time human involvement. The technical term for such a weapon is an out-of-the-loop system. Unlike in-the-loop-systems in which the decision to apply kinetic force to a target is made by the weapon's operator in real-time, or on-the-loop systems where the operator remains on stand-by and can override the weapon, a genuine out-of-the-loop system will not involve an operator once deployed.[10]

Third, the notion of out-of-the-loop systems could be equally applied to automated and autonomous systems. Indeed, the literature is far from clear where the difference between the two lies, and any boundaries between automated and autonomous machine behaviour might be fluid. As a rule of thumb, autonomous systems are more flexible in their response to their operating environment than automated ones.[11] They could learn from their prior experiences in order to optimise their (future) performance, for example. They might also have greater leeway in translating the orders given via their programming into action. What this means in practice is that, compared to an automated system, any autonomous system (and not just weapons) is less predictable in its behaviour. That said, AWS would be constrained by particular targeting categories. That is, their programming would only allow them to attack targets that fall within a particular category. To illustrate the point, an AWS programmed to search and destroy enemy tanks would be restricted to attacking entities that fall into this category. Yet, compared to an automated weapon, it would be hard to predict where, when, and which enemy tank it would attack.

Fourth, as the quote from Secretary General *Guterres* suggests, AWS can produce a lethal kinetic effect without any human intervention post-programming. Here, the question is whether the alleged moral repugnancy of AWS only refers to AWS that would be deliberately programmed to attack human individuals. If so, this would potentially leave scope for the development and deployment of AWS that are not used for this purpose, such as the one mentioned in the 'enemy tank' example above. Moreover, it is noteworthy that any weapon can kill in two ways, (1) as an intended effect of its operation, and (2) as a side-effect of its operation. Presumably, the earlier quote by Secretary *Guterres* refers to (1), where a programmer would intentionally programme an AWS in order to attack human individuals, most likely enemy combatants.

The focus on this issue is problematic, for two reasons. First, it neglects lethal harm that might arise as a side effect of the operation of an AWS. As I shall show later, this category of harm is, in the context of AWS, more morally problematic than intended harm. Second, it is doubtful whether the intentional targeting of individuals through AWS is legally and morally permissible. To explain, as was noted in the introduction to this chapter, at the level of *jus in bello*, contemporary just war theory post-*Walzer* rests on the distinction between combatants and non-combatants. True, given advances in machine vision, an AWS could, with great reliability,

[9] Werkhauser, 'Killer Robots' (n 1)

[10] P Scharre, *Army of None: Autonomous Weapons and the Future of War* (2019).

[11] A Leveringhaus, *Ethics and Autonomous Weapons* (2006) 46 *et seq* (hereafter Leveringhaus, *Ethics and Autonomous Weapons*).

https://doi.org/10.1017/9781009207898.036 Published online by Cambridge University Press

distinguish between human individuals and non-human objects and entities. Yet, what it cannot do, at the present state of technological development at least, is to accurately determine whether an individual is a legitimate target (a combatant) or an illegitimate target (a non-combatant). It is, in fact, hard to see how a machine's capacity for such a qualitative judgement could ever be technologically achieved. As a result, the deployment of an AWS to deliberately kill human individuals would not be permissible under *jus in bello*.

If the above observation is true, it has two immediate repercussions for the debate on AWS. First, militaries might not be particularly interested in developing systems whose purpose is the autonomous targeting of human individuals, knowing that such systems would fall foul of *jus in bello*. Still, militaries may seek to develop AWS that can be programmed to attack more easily identifiable targets – for example, a tank, a missile, or a submarine. In this case, I contend that the ethical debate on AWS misses much of the actual technological development and restricts its own scope unnecessarily. Second, as I have argued elsewhere,[12] in order to assess whether programming an AWS to kill human individuals is morally repugnant, it is necessary to assume that AWS do not fall down at the normative hurdle of accurately identifying human individuals as legitimate or illegitimate targets. This assumption is a necessary philosophical abstraction and technological idealisation of AWS that may not reflect their actual development and potential uses. Bearing this in mind, the chapter continues by analysing whether it is morally repugnant to deliberately programme an AWS to kill human individuals in war.

III. PROGRAMMED TO KILL: THREE ETHICAL RESPONSES

The main ethical argument in favour of AWS is essentially humanitarian in nature.[13] More precisely, the claim is that AWS (1) ensure stricter compliance with *jus in bello*, and (2) reduce human suffering and casualties as a result.[14] Interestingly, the ethical counterarguments do not engage with this humanitarian claim directly. Rather, they immediately attack the notion of autonomous uses of force via an AWS. In this part of the chapter, I look at three ethical responses to the prospect of AWS being intentionally programmed to take human lives, (1) the argument that AWS create so-called responsibility gaps, (2) the claim that the intentional use of AWS to kill is incompatible with human dignity, and (3) the argument (made by this author) that, by replacing human agency with artificial agency at the point of force delivery, AWS render humans incapable of revising a decision to kill. As indicated above, the three arguments rely on a technologically-idealised view of AWS.

1. *Responsibility Gaps*

One of the earliest contributions to the ethical debate on AWS is the argument that these weapons undermine a commitment to responsibility. Put simply, the claim is that, in certain cases, it is not possible to assign (moral) responsibility to a human individual for an event caused by an AWS. This is especially problematic if the event constitutes a violation of *jus in bello*. In such cases, neither the manufacturer of the AWS, nor its programmer, nor the AWS itself (of course) can be held responsible for the event, resulting in a responsibility gap.[15] This gap

[12] Leveringhaus, *Ethics and Autonomous Weapons* (n 11).

[13] Ibid, 62–63.

[14] R Arkin, 'The Case for Ethical Autonomy in Unmanned Systems' (2010) 9(4) *Journal of Military Ethics*, 332–341.

[15] R Sparrow, 'Killer Robots' (2007) 24(1) *Journal of Applied Philosophy*, 62–77.

https://doi.org/10.1017/9781009207898.036 Published online by Cambridge University Press

arises from the inherent unpredictability of autonomous machine behaviour. No human programmer, it is claimed, could foresee every facet of emerging machine behaviour. Hence, it is inappropriate, the argument goes, to hold the programmer – let alone the manufacturer – responsible for an unforeseen event caused by an AWS. In a moral sense, no one can be praised or blamed, or even punished, for the event. Why should this pose a moral problem? Here, the claim is that for killing in war to be morally permissible, someone needs to be held responsible for the use of force. Responsibility gaps, thus, undermine the moral justification for killing in war.

Admittedly, the idea of a responsibility gap is powerful. But it can be debunked relatively easily. First, moral responsibility can be backward-looking and forward-looking. The responsibility gap arises from a backward-looking understanding of responsibility, where it is impossible to hold a human agent responsible for an event caused by an AWS in the past. The argument has nothing to say about the forward-looking sense of responsibility, where an agent would be assigned responsibility for supervising, controlling, or caring for someone or something in the future. In the present context, the forward-looking sense of responsibility lends itself to an on-the-loop system, rather than an out-of-the-loop system. Either way, it is not clear whether a gap in backward-looking responsibility is sufficient for the existence of a responsibility gap, or whether there also needs to be a gap in forward-looking responsibility. A backward-looking gap may be a necessary condition here, but not a sufficient one.

Second, it is contested whether killing in war is *prima facie* permissible if, and only if, someone can be held responsible for the use of lethal force. There are, roughly, two traditions in contemporary moral philosophy for thinking about the issue.[16] The first, derived from Thomism, is agent-centric in that it focuses on the intentions of the agent using lethal force. The second tradition is target-centric in that it focuses on the moral status of the target of lethal force. That is to say, the permissibility centres on the question whether the target has become liable to attack because it is morally and/or causally responsible for a (unjust) threat. On the target-centric approach, an agent who could not be held responsible for the use of lethal force may be allowed to kill if the target was liable to attack. In short, then, if the link between (agent) responsibility and the moral permission to use force is far weaker than assumed, the idea of a responsibility gap loses its normative force.

Third, the idea of a responsibility gap lets those who deployed an AWS off the hook far too easily.[17] True, given that autonomous systems tend to show unpredictable emerging behaviours, the individual (or group of individuals) who deploys an AWS by programming it with its mission parameters cannot know in advance that, at t_5, the AWS is going to do x. Still, the programmer and those in the chain of command above him know that the AWS they deploy is likely to exhibit unforeseen behaviour, which might, in the most extreme circumstances, result in the misapplication of force. Notwithstanding that risk, they choose to deploy the weapon. In doing so, they impose a significant risk on those who might come into contact with the AWS in its area of operation, not least non-combatants. Of course, the imposition of that risk may either be reasonable and permissible under the circumstances or unreasonable and reckless – more on this shortly. But generally, the claim that those deploying an AWS are not responsible for any unforeseen damage resulting from its operation appears counterintuitive.

Finally, even if it is hard to hold individuals responsible for the deployment of an AWS, it is worthwhile remembering that armed conflicts are (usually) fought by states. In the end, the buck stops there. Needless to say, this raises all sorts of difficult issues which the chapter cannot go

[16] S Uniacke, *Permissible Killing: The Self-Defence Justification of Homicide* (1994).

[17] Leveringhaus, *Ethics and Autonomous Weapons* (n 11) 76–86.

https://doi.org/10.1017/9781009207898.036 Published online by Cambridge University Press

into. For now, it suffices to note that states have made reparations for the (wrongful) damage they caused in armed conflict. Most recently, for instance, the United States (US) compensated Afghan civilians for the deaths of (civilian) family members in the course of US military operations in the country as part of the so-called War on Terror.[18] The most notorious case is that of Staff Sergeant *Robert Bales* who, after leaving his base without authorisation, went on a shooting rampage and was later charged with the murder of seventeen Afghan civilians, as well as causing injury to a number of others. The US paid compensation to those affected by Sergeant *Bales*' actions, even though Sergeant *Bales* acted out of his own volition and outside the chain of command.[19]

In sum, the notion of a responsibility gap does not prove that AWS are morally repugnant. Either the existence of a (backward-looking) responsibility gap is insufficient to show that the deployment of AWS would be morally unjustifiable or there is no responsibility gap as such. Yet, there are elements of the responsibility gap that could be salvaged. The argument that it is necessary to be able to hold someone responsible for the use of force is motivated by a concern for human dignity or respect for individuals. It might, therefore, be useful to focus on the relationship between AWS and human dignity. That is the purpose of the next section.

2. *Dignity*

Are AWS morally repugnant because, as has been suggested by some contributors to the debate, they are an affront to human dignity?[20] This question is difficult to answer because just war theorists have tended to eschew the concept of human dignity. Perhaps for good reason. Appeals to dignity often do not seem to decisively resolve difficult moral issues. For instance, the case for, as well as against, physician-assisted suicide could be made with reference to the concept of dignity. That said, the concept enters into contemporary just war thinking, albeit in an indirect way. This has to do with the aforementioned distinction between combatants and non-combatants. The former group is seen as a legitimate target in armed conflict, which means that combatants lack a moral claim against other belligerent parties not to intentionally kill them. Non-combatants, by contrast, are immune to intentional attack, which means that they hold a negative moral claim against combatants not to intentionally kill them. However, *jus in bello* does not grant non-combatants immunity against harm that would be unintentionally inflicted. Here, the Doctrine of Double Effect and its conceptual and normative distinction between intended and foreseen harm comes into play. In his classic discussion of non-combatant immunity, *Walzer* argues that it is permissible to kill or harm non-combatants if, and only if, the harm inflicted on them is (1) not intended, (2) merely foreseen (by the belligerent), (3) not used as a (bad) means to a good effect, (4) proportionate (not excessive to the good achieved), and (5) consistent with a belligerent's obligations of 'due care'.[21]

Granted, but why should the distinction between intended and foreseen harm have any normative significance? According to the *Kantian* view, the Doctrine of Double Effect protects

[18] See M Gluck, 'Examination of US Military Payments to Civilians Harmed during Conflict in Afghanistan and Iraq' (*Lawfare*, 8 October 2020) www.lawfareblog.com/examination-us-military-payments-civilians-harmed-during-conflict-afghanistan-and-iraq.

[19] Associated Press, 'US Compensation for Afghanistan Shooting Spree' (*The Guardian*, 25 March 2012) www.theguardian.com/world/2012/mar/25/us-compensation-afghanistan-shooting-spree.

[20] See A Pop, 'Autonomous Weapon Systems: A Threat to Human Dignity?' (International Committee of the Red Cross, *Humanitarian Law & Policy*, 10 April 2018) https://blogs.icrc.org/law-and-policy/2018/04/10/autonomous-weapon-systems-a-threat-to-human-dignity/.

[21] Walzer, *Just and Unjust Wars* (n 5) 153–154.

https://doi.org/10.1017/9781009207898.036 Published online by Cambridge University Press

the dignity of innocent individuals by ensuring that belligerents comply with the second formulation of *Kant's* categorical imperative, which obliges them to treat (innocent) individuals not merely as means to an end but always also as ends-in-themselves.[22] To illustrate the point, if Tim intentionally bombs non-combatants in order to scare the enemy into surrender, Tim violates their status as ends-in-themselves, instrumentalising their deaths in order to achieve a particular goal (the end of the war). By contrast, if Tom bombs a munitions factory and unintentionally kills non-combatants located in its vicinity as a foreseen side-effect of his otherwise permissible (and proportionate) military act, Tom does not instrumentalise their deaths for his purposes. Counterfactually, Tom could destroy the munitions factory, even if no non-combatant was harmed. Unlike Tim, Tom does not need to kill non-combatants to achieve his goals. Tom's actions would not violate the ends-not-means principle – or so one might argue.

According to the *Kantian* View of the Doctrine of Double Effect, then, if Tam intentionally programmed an AWS to kill non-combatants, he would violate their dignity. Note, though, that there is no moral difference between Tam's and Tim's actions. The only difference is the means they use to kill non-combatants. As a result, this example does not show that AWS pose a unique threat to human dignity. Any weapon could be abused in the way Tam abuses the AWS. Hence, in the example, the use of the AWS is morally repugnant, not the weapon as such.

What about combatants? If Tam intentionally programmed an AWS to kill enemy combatants, would he violate their dignity? That question is hard to answer conclusively. First, because combatants lack a moral claim not to be killed, Tam does not violate their moral rights by deploying an AWS against them. Second, unlike non-combatants, it is usually morally permissible and necessary to instrumentalise combatants. One does not need to go quite as far as Napoleon who remarked that 'soldiers are made to be killed'.[23] But *Walzer* is right when he observes that combatants are the human instruments of the state.[24] As a result, combatants enjoy far lower levels of protection against instrumentalization than non-combatants. In a nutshell, it needs to be shown that, although combatants (1) lack a moral claim not to be intentionally attacked [during combat], and (2) do not enjoy the same level of protection against instrumentalization as non-combatants, the use of an AWS in order to kill them would violate their dignity.

The dignity of combatants, critics of AWS may argue, is violated because a machine should not be left to decide who lives or dies. At the macro-level of programming the argument is certainly wrong. Tam, the programmer in the above example, makes the decision to programme an AWS to detect and eliminate enemy combatants. In this sense, the machine Tam deploys does not make a decision to take life. Tam does. At the micro-level of actual operations, though, the argument has some validity. Here, the machine has some leeway in translating Tam's instructions into actions. Within the target category of enemy combatants, it could 'decide' to attack Combatant$_1$ rather than Combatant$_2$ or Combatant$_3$. It might, further, not be possible to ascertain why the machine chose to attack Combatant$_1$ over Combatant$_2$ and Combatant$_3$. The resulting question is whether the machine's micro-choice, rather than Tam's macro-choice, violates Combatant$_1$'s dignity.

Arguably not. This is because killing in war tends to be impersonal and to some extent morally arbitrary. Why did a particular combatant die? Often, the answer will be that he was a combatant. Armed conflict, as *Walzer* observes, is not a personal relationship. To wit,

[22] TA Cavanaugh, *Double Effect Reasoning: Doing Good and Avoiding Evil* (2006).

[23] Walzer, *Just and Unjust Wars* (n 5) 136.

[24] Ibid, 36–45.

https://doi.org/10.1017/9781009207898.036 Published online by Cambridge University Press

combatants are not enemies in a personal sense, which would explain the choices they make. They are the human instruments of the state. They kill and die because they are combatants. And often because they are in the wrong place at the wrong time. That is the brutal reality of warfare. Consider a case where an artillery operator fires a mortar shell in the direction of enemy positions. Any or no enemy combatant located in the vicinity might die as a result. We might never know why a particular enemy combatant died. We only know that the artillery operator carried out his orders to fire the mortar shell. By analogy, the reason for an AWS's micro-choice to target Combatant$_1$ over Combatant$_2$ and Combatant$_3$ is, ultimately, that Combatant$_1$ is a combatant. Combatant$_1$ was simply in the wrong place at the wrong time. It is not clear why this micro-choice should be morally different from the artillery operator's decision to fire the mortar shell. Just as the dignity of those combatants who were unlucky enough to be killed by the artillery operator's mortar shell is not violated by the artillery operator's actions, Combatant$_1$'s dignity is not violated because a machine carried out its pre-programmed orders by micro-choosing him over another combatant. So, the argument that human dignity is violated if a machine makes a micro-choice over life and death seems morally dubious.

But perhaps critics of AWS may concede that the micro-choice as such is not the problem. To be sure, killing in war, even under orders, is to some extent random. The issue, they could reply, is that the artillery operator and those whom he targets have equal skin in the game, while the AWS that kills Combatant$_1$ does not. In other words, the artillery operator has an appreciation of the value of (his own) life, which a machine clearly lacks. He is aware of the deadly effects of his actions, whereas a machine is clearly not. Perhaps this explains the indignity of being killed as a result of a machine's micro-choice.

This argument takes us back to the *Thomistic* or agent-centric tradition in the ethics of killing outlined previously. Here, the internal states of the agent using force, rather than the moral status of the target, determines the permissibility of killing. To be allowed to kill in war, a combatant needs to have an appreciation of the value of life or at least be in a similar situation to those whom he targets. Naturally, if one rejects an agent-centric approach to the ethics of killing, this argument does not hold much sway.

More generally, it is unclear whether such a demanding condition – that an individual recognises the value of life – could be met in contemporary armed conflict. Consider the case of high altitude bombing during NATO's war in Kosovo. At the time, *Michael Ignatieff* observed that NATO was fighting a 'virtual war' in which NATO did the fighting while most of the Serbs 'did the dying'.[25] It is hard to imagine that NATO's bomber pilots, flying at 15,000 ft and never seeing their targets, would have had the value of human life at the forefront of their minds, or would have even thought of themselves as being in the same boat as those they targeted. The pilots received certain target coordinates, released their payloads once they had reached their destination, and then returned to their base. In short, modern combat technology, in many cases, has allowed combatants to distance themselves from active theatres, as well as the effects of their actions, to an almost unprecedented degree. These considerations show that the inability of a machine to appreciate the value of life does not pose a distinctive threat to human dignity. The reality of warfare has already moved on.

But there may be one last argument available to those who seek to invoke human dignity against AWS. To be sure, combatants, they could concede, do not hold a moral claim against other belligerents not to attack them. Nor, as instruments of the state, do they enjoy the same level of protection against instrumentalization as non-combatants. Still, unless one adopts

[25] M Ignatieff, *Virtual War* (2000).

https://doi.org/10.1017/9781009207898.036 Published online by Cambridge University Press

Napoleonic cynicism, there must be some moral limits on what may permissibly be done to combatants on the battlefield. There must be some appreciation that human life matters, and that humans are not merely a resource that can be disposed of in whatever way necessary. Otherwise, why would certain weapons be banned under international law, such as blinding lasers, as well as chemical and biological weapons?

Part of the answer is that these weapons are likely to have an indiscriminate and disproportionate effect on non-combatants. But intuitively, as the case of blinding lasers illustrates, there is a sense that combatants deserve some protection. Are there certain ways of killing that are somehow cruel and excessive, even if they were aimed at legitimate human targets? And if that is the case, would AWS fall into this category?

There is a comparative and a non-comparative element to these questions. Regarding the comparative element, as macabre as it sounds, it would certainly be excessive to burn a combatant to death with a flamethrower if a simple shot with a gun would eliminate the threat he poses. That is common-sense. With regard to the non-comparative element, the issue is whether there are ways of killing which are intrinsically wrong, regardless of how they compare to alternative means of killing. That question is harder to answer. Perhaps it is intrinsically wrong to use a biological weapon in order to kill someone with a virus. That said, it is hard to entirely avoid comparative judgements. Given the damage that even legitimate weapons can do; it is not clear that their effects are always morally more desirable than those of illegitimate weapons. One wonders if it is really less 'cruel' for someone to bleed to death after being shot or to have a leg blown off from an explosive than to be poisoned. Armed conflict is brutal and modern weapons technology is shockingly effective, notwithstanding the moral (and legal) limits placed on both.

Although, within the scope of this chapter, it is impossible to resolve the issues arising from the non-comparative element, the above discussion provides two main insights for the debate on AWS. First, if AWS are equipped with payloads whose effects were either comparatively or non-comparatively excessive or cruel, they would certainly violate relevant moral prohibitions against causing excessive harm. For example, an autonomous robot with a flamethrower that would incinerate its targets or an autonomous aerial vehicle that would spray target areas with a banned chemical substance would indeed be morally repugnant. Second, it is hard to gauge whether the autonomous delivery of a legitimate – that is, not disproportionately harmful – payload constitutes a cruel or excessive form of killing. Here, it seems that the analysis is increasingly going in circles. For, as I argued above, many accepted forms of killing in war can be seen analogous to, or even morally on a par with, autonomous killing. Either all of these forms of killing are a threat to dignity, which would lend succour to ethical arguments for pacifism, or none are.

To sum up, AWS pose a threat to human dignity if they were deliberately used to kill non-combatants, or were equipped with payloads that caused excessive or otherwise cruel harm. However, even in such cases, AWS would not pose a distinctive threat. This is because some of the features of autonomous killing can also be found in established forms of killing. The moral issues AWS raise with regard to dignity are not unprecedented. In fact, the debate on AWS might provide a useful lens through which to scrutinise established forms of killing in war.

3. *Human and Artificial Agency*

If the earlier arguments are correct, the lack of direct human involvement in the operation of an AWS, once programmed, is not a unique threat to human dignity. Yet, intuitively, there is something morally significant about letting AWS kill without direct human supervision.

https://doi.org/10.1017/9781009207898.036 Published online by Cambridge University Press

This author has sought to capture this intuition via the Argument from Human Agency.[26] I argue that AWS have artificial agency because they interact with their operating environment, causing changes within it. According to the Argument from Human Agency, the difference between human and artificial agency is as follows. Human agency consists in refusing to carry out an order. As history shows, soldiers have often not engaged the enemy, even when under orders to do so. An AWS, by contrast, will kill once it has 'micro-chosen' a human target. We might not know when, where, and whom it will kill, but it will carry out its programming. In a nutshell, by removing human agents from the point of payload delivery, out-of-the-loop systems make it impossible to revise a decision to kill.

While the Argument from Human Agency captures intuitions about autonomous forms of killing, it faces three challenges. First, as was observed above, combatants do not hold a moral claim not to be killed against other belligerent parties and enjoy lower levels of protection against instrumentalization than non-combatants. Why, then, critics of the Argument from Human Agency might wonder, should combatants sometimes not be killed? The answer is that rights do not always tell the whole moral story. Pity, empathy, or mercy are sometimes strong motivators not to kill. Sometimes (human) agents might be permitted to kill, but it might still be morally desirable for them not to do so. This argument does not depend on an account of human dignity. Rather, it articulates the common-sense view that killing is rarely morally desirable even if it is morally permissible. This is especially true during armed conflict where the designation of combatant status is sufficient to establish liability to attack. Often, as noted above, combatants are killed simply because they are in the wrong place at the wrong time, without having done anything.

The second challenge to the Argument from Human Agency is that it delivers too little too late. As the example of high-altitude bombing discussed earlier showed, modern combat technology has already distanced individuals from theatres in ways that make revising a decision to kill difficult. The difference, though, between more established weapons and out-of-the-loop systems is that the latter systems remove human agency entirely once the system has been deployed. Even in the case of high-altitude bombing, the operator has to decide whether to 'push the button'. Or, in the case of an on-the-loop system, the operator can override the systems' attack on a target. Granted; in reality, an operator's ability to override an on-the-loop system might be vanishingly small. If that is the case, there might be, as the Argument from Human Agency would concede, fewer reasons to think that AWS were morally unique. Rather, from the perspective of the Argument from Human Agency, many established forms of combat technology are more morally problematic than commonly assumed.

The third challenge is a more technical one for moral philosophy. If, according to the Argument from Human Agency, not killing is not strictly morally required because killing an enemy combatant via an AWS does not violate any moral obligations owed to that combatant, there could be strong reasons in favour of overriding the Argument from Human Agency. This would especially be the case when the deployment of AWS, as their defenders claim, led to significant reductions in casualties. Here, the Argument from Human Agency is weaker than dignity-based objections to AWS. In non-consequentialist or deontological moral theory, any trade-offs between beneficial aggregate consequences and dignity would be impermissible. The Argument from Human Agency, though, does not frame the issue in terms of human dignity. There might, thus, be some permissible trade-offs between human agency (deployment of human soldiers), on the one hand, and the aggregate number of lives saved via the deployment

[26] Leveringhaus, *Ethics and Autonomous Weapons* (n 11) 89–117.

https://doi.org/10.1017/9781009207898.036 Published online by Cambridge University Press

of AWS, on the other. Still, the Argument from Human Agency illustrates that there is some loss when human agency is replaced with artificial agency. And that loss needs to clear a high justificatory bar. Here, the burden of proof falls on defenders of AWS.

To conclude, while the Argument from Human Agency captures intuitions about autonomous killing, it is not sufficient to show that it is categorically impermissible to replace human with artificial agency. It merely tries to raise the justificatory bar for AWS. The humanitarian gains from AWS must be high for the replacement of human agency with artificial agency to be morally legitimate. More generally, neither of the three positions examined above – the responsibility gap, human dignity, and human agency – serve as knockdown arguments against AWS. This is partly because, upon closer inspection, AWS are not more (or less) morally repugnant than established, and more accepted, weapons and associated forms of killing in war. In this light, it makes sense to shift the focus from the highly idealised scenario of AWS being deliberately programmed to attack human targets to different, and arguably more realistic, scenarios. Perhaps these alternative scenarios provide a clue as to why AWS might be morally problematic. The fourth and final part of the chapter looks at these scenarios in detail.

IV. THREE EMERGING ETHICAL PROBLEMS WITH AWS

As was emphasised earlier, for technological reasons, it is hard to see that the intentional programming of AWS in order to target combatants could be morally (or legally) permissible. As a result, the intended killing of combatants via AWS is not the main ethical challenge in the real world of AWS. Rather, AWS will be programmed to attack targets that are more easily and reliably identifiable by a machine. It is not far-fetched, for instance, to imagine an autonomous submarine that hunts other submarines, or an autonomous stealth plane programmed to fly into enemy territory and destroy radar stations, or a robot that can detect and eliminate enemy tanks. While these types of AWS are not deliberately programmed to attack human individuals, they still raise important ethical issues. In what follows, I focus on three of these.

First, the availability of AWS, some critics argue, has the potential to lead to more wars. Surely, in light of the destruction and loss of life that armed conflicts entail, this is a reason against AWS. If anything, we surely want fewer wars, not more. Yet, in the absence of counterfactuals, it is difficult to ascertain whether a particular form of weapons technology necessarily leads to more wars. If, for instance, the Soviet Union and US had not had access to nuclear weapons, would they have gone to war after 1945? It is impossible to tell. Moreover, it is noteworthy that a mere increase in armed conflict does not tell us anything about the justness of the resulting conflicts. Of course, if the availability of AWS increased the willingness of states to violate *jus ad bellum* by pursuing unjust wars, then these weapons are not normatively desirable. If, by contrast, the effect of AWS on the frequency of just or unjust wars was neutral, or if they increased the likelihood of just wars, they would, *ceteris paribus*, not necessarily be morally undesirable.

Yet, while it is not self-evident that AWS lead to an increase in unjust wars, their availability potentially lends itself to more covert and small-scale uses of force. Since the US's targeted killing campaign against suspected terrorists in the late 2000s, just war theorists have increasingly been concerned with uses of force that fall below the threshold for war and thus outside the regulatory frameworks provided *jus ad bellum* and *jus in bello*. Using the US-led War on Terror as a template, force is often used covertly and on an *ad hoc* basis, be it through the deployment of special forces or the targeting of alleged terrorists via remote-controlled aerial vehicles ('drones'), with few opportunities for public scrutiny and accountability. AWS might be ideal

https://doi.org/10.1017/9781009207898.036 Published online by Cambridge University Press

for missions that are intended to fall, literally, under the radar. Once deployed, an AWS in stealth mode, without the need for further communication with a human operator, could enter enemy territory undetected and destroy a particular target, such as a military installation, a research facility, or even dual-use infrastructure. Although AWS should not be treated differently from other means used in covert operations, they may reinforce trends towards them.

Second, there is an unnerving analogy between AWS, landmines, and unexploded munitions, which often cause horrific damage in post-war environments. As we just saw, AWS can operate stealthily and without human oversight. With no direct human control over AWS, it is unclear how AWS can be deactivated after hostilities have been concluded. Rather unsettlingly, AWS, compared to landmines and unexploded munitions, could retain a much higher level of combat readiness. The moral issue is trivial and serious at the same time: does the very presence of autonomy in a weapon and the fact that it is an out-of-the-loop system make it difficult to switch it off? In other words, the central question is how, once human control over a weapon is ceded, it can be reasserted. How, for example, can a human operator re-establish control over an autonomous submarine operating in an undisclosed area of the high seas? There might eventually be technological answers to this question. Until then, the worry is that AWS remain a deadly legacy of armed conflict.

Third, while just war theorists have invested considerable energy into disambiguating the distinction between intended harm and unintended but foreseen harm, unintended and unforeseen harms, emanating from accidents and other misapplications of force, have received less attention. These harms are more widespread than assumed, leading to significant losses of life among non-combatants. Naturally, the fact that harm is unintended and unforeseen does not render it morally unproblematic. To the contrary, it raises questions about negligence and recklessness in armed conflict. One hypothesis in this respect, for instance, is that precision-weaponry has engendered reckless behaviour among belligerents.[27] Because these weapons are seen as precise, belligerents deploy them in high-risk theatres where accidents and misapplications of force are bound to happen. Here, abstention or the use of non-military alternatives seem more appropriate. For example, the use of military-grade weaponry, even if it is precise, over densely populated urban areas is arguably so risky that it is morally reckless. Belligerents know the risks but go ahead anyway because they trust the technology.

The conceptual relationship between precision-weaponry and AWS is not straightforward, but the question of recklessness is especially pertinent in the case of AWS.[28] After all, AWS not only create a significant kinetic effect, but they are unpredictable in doing so. As the saying goes, accidents are waiting to happen. True, in some cases, it might not be reckless to deploy AWS – for example, in extremely remote environments. But in many instances, and especially in the kinds of environments in which states have been conducting military operations over the last twenty-five years, it is morally reckless to deploy an inherently unpredictable weapon. Even if such a weapon is not deliberately programmed to directly attack human individuals, the threat it poses to human life is all too real. Can it really be guaranteed that an autonomous tank will not run over a civilian when speeding towards its target? What assurances can be given that an autonomous submarine does not mistake a boat carrying refugees for an enemy vessel? How can we be certain that a learning mechanism in a robotic weapon's governing software does not 'learn' that because a child once threw a rock at the robot during a military occupation, children in general constitute threats and should therefore be targeted? These worries are compounded

[27] B Cronin, *Bugsplat: The Politics of Collateral Damage in Western Armed Conflict* (2018).

[28] A Leveringhaus, 'Autonomous Weapons and the Future of Armed Conflict', in J Gailliot, D McIntosh, and JD Ohlin (eds), *Lethal Autonomous Weapons: Re-examining the Law and Ethics of Robotic Warfare* (2021) 175.

https://doi.org/10.1017/9781009207898.036 Published online by Cambridge University Press

by the previous point about re-establishing control over an AWS. After control is ceded, it is not clear how it can be re-established, especially when it becomes apparent that the system does not operate in the way it should.

Advocates of AWS could mount two replies here. First, eventually there will be technological solutions that reduce the risk of accidents. Ultimately, this necessitates a technological assessment that ethicists cannot provide. The burden of proof, though, lies with technologists. Second, humans, defenders of AWS could point out, are also unpredictable, as the occurrence of war crimes or reckless behaviour during armed conflict attests. But the reply has three flaws. The first is that AWS will not be capable of offering a like-for-like replacement for human soldiers in armed conflict, especially when it comes to operations where the targets are enemy combatants (who would need to be differentiated from non-combatants). In this sense, the scope for human error, as well as wrongdoing, in armed conflict remains unchanged. The second flaw is that, although human individuals are unquestionably error-prone and unpredictable, AWS are unlikely, at the present stage of technological development, to perform any better than humans. The final flaw in the response is that, in the end, a fully armed weapons system has the capacity to do far more damage than any single soldier. For this reason alone, the deployment of AWS is, with few exceptions, morally reckless.

Taking stock, even if one turns from the highly abstract debate on AWS in contemporary philosophy to a more realistic appreciation of these weapons, moral problems and challenges do not magically disappear. Far from it, AWS potentially reinforce normatively undesirable dynamics in contemporary armed conflict, not least the push towards increasingly covert operations without public scrutiny, as well as the tendency for high-tech armies to (sometimes) take unreasonable, if not reckless, risks during combat operations. The key question of how control can be re-established over an out-of-the-loop system has not been satisfactorily answered, either. While these observations may not render AWS morally distinctive, they illustrate their *prima facie* undesirability.

V. CONCLUSION

Perhaps more than any other form of emerging weapons technology, AWS have been met with moral condemnation. As the analysis in this chapter shows, it is hard to pin down why they should be 'morally repugnant'. Some of the central ethical arguments against AWS do not withstand critical scrutiny. In particular, they fail to show that AWS are morally different from more established weapons and methods of warfighting. Still, the chapter concludes that AWS are morally problematic, though not necessarily morally repugnant. The main point here is that, for the foreseeable future, AWS are not safe enough to operate in what is often a complex and chaotic combat environment. This is not to say that their technological limitations might not eventually be overcome. But for now, the deployment of a weapon whose behaviour is to some extent unpredictable, without sufficient and on-going human oversight and the ability to rapidly establish operator control over it, seems morally reckless. True, other types of weapons can be used recklessly in armed conflict, too. The difference is that the technology underpinning AWS remains inherently unpredictable, and not just the use of these weapons. Furthermore, while AWS do not appear to raise fundamentally new issues in armed conflict, they seem to reinforce problematic dynamics in the use of force towards ever more covert missions. AWS might make it considerably easier for governments to avoid public scrutiny over their uses of force. Hence, for democratic reasons, and not just ethical ones, the arrival of AWS and the prospect of autonomous war fighting should be deeply troubling.

https://doi.org/10.1017/9781009207898.036 Published online by Cambridge University Press