AI and Ethics
https://doi.org/10.1007/s43681-024-00567-7

ORIGINAL RESEARCH

Robot warfare: the (im)permissibility of autonomous weapons systems

Jack Madock[1]

Received: 7 May 2024 / Accepted: 22 August 2024

Abstract
This paper argues against prominent views of the impermissibility of autonomous weapons systems (AWS). It does so by assuming each theory is true and arguing towards contradiction. To arrive at a contradiction two assumptions are necessary. First, the theory of impermissibility in question is assumed. Second, a thought experiment called the ideal warfare scenario is assumed. The paper aims to demonstrate that in theory AWS could be deployed such that they bring about the best of possible warfare. However, even if AWS were deployed in the fairest and safest way, these accounts would still find them to be impermissible. This is deemed to be a failure of these accounts. Finally, this paper discusses the future of artificial intelligence (AI). It explores how AI may be used in the future of warfare and the challenges advanced AI would pose to the ethics of war and the relationship between human soldiers and AWS.

Keywords Autonomous weapons systems · Robert Sparrow · Ethics of warfare · Moral permissibility · Ethics of artificial intelligence

1 Introduction

In the coming years military technology will undergo a shift towards the design and implementation of autonomous weapons systems (AWS).[1] This has raised numerous concerns across practical and philosophical discourse over the decisions that these AWS will make in warfare. These concerns range from the ability (or lack thereof) of holding these systems morally responsible [11, 13], to their capability to exhibit the appropriate respect to fellow combatants [1, 14], to whether or not AWS can act for the right sort of reasons or even for reasons at all [8, 17].

This paper offers an argument by way of *reductio ad absurdum*. More specifically, it assumes a hypothetical thought experiment involving the most ideal warfare scenario possible (IWS). This is a thought experiment in which there is a war fought without the involvement of human combatants at all. This involves all human soldiers being replaced by AWS in a protected combat zone. Further, the paper makes the case that this IWS is a morally permissible situation. This is ensured by bracketing concerns that could render it impermissible such as concerns of an unjust war or of a tyrannical despot coming to power as a result of the combat. The argument then shows that when applied in the situation of the IWS, the foregoing accounts of the impermissibility of AWS would still deliver a verdict of impermissible. But this of course must be a contradiction, as the war scenario that involves no loss of life must be morally permissible and is in theory achievable by deploying AWS. The motivating idea of the paper is a simple one: The accounts of the impermissibility of deploying AWS that have been put forth so far cannot be correct. That is so because those

[1] United States. Dept. of Defense [16], Red Cross [10], Etzioni [2].

✉ Jack Madock
johnmadock@ufl.edu

[1] Department of Philosophy, University of Florida, Gainesville, FL 32601, USA

Published online: 05 September 2024

theories argue that AWS lack a morally relevant property (reasons, respect, or responsibility). But having these properties is at best derivatively related to fairness or justice in war and at worst irrelevant. The most important thing in any armed conflict, is to minimize harm and suffering to humans. If AWS could achieve this, then they should be deployed.[2] The problem is that those aforementioned theories would deem AWS impermissible to deploy even if they could bring about the safest conflict imaginable.

The plan for the paper is as follows. Section one contains clarification over the definitions and terms that are used throughout the paper. It further offers more detail on the thought experiment of the ideal warfare scenario introduced above. Section two explores the existing arguments of the impermissibility of AWS and presents challenges to those arguments. Section three highlights contradictions in the arguments put forth in previously offered theories. This data is used to encourage a reconsideration of those arguments. It finally explores and highlights the issue of human suffering and looks into potential implications of further expanding the use of artificial intelligence in warfare. Section four concludes.

1.1 Definitions and existing regulations

The term autonomous weapon system (AWS) refers to a broad range of weapons that have at least limited forms of agency over the decisions they take in the battlefield. The idea of autonomous weapons operating in the field of battle is relatively old and is mentioned in Defense Advanced Research Projects Agency documents from 1983 [6]. What many imagine when they think of AWS however remains at this stage to be the material of science fiction. It therefore remains the case that what is currently discussed in the literature over AWS are a range of highly automated control systems.[3] Indeed with any of the weapons about which this argument is concerned there remains a level of meaningful human control.

With this in mind, Schmitt has offered a definition of AWS which includes their ability to identify targets and choose to attack without any human intervention [12]. In keeping with the above-mentioned sentiment, this definition stops short of full autonomy:

> "The crux of full autonomy is a capability to identify, target, and attack a person or object without human interface. Although a human operator retain the ability to take control of the system, it can operate without any control being exercised. Of course, a fully autonomous system is never completely human-free. Either the system designer or an operator would at least have to program the system to function pursuant to specified parameters" (ibid, 1).

This definition nicely encapsulates the state of autonomous weapons as they exist and are likely to exist for years to come. Following Purves et al. we can consider these types of AWS as a sort of weak AI, that is these systems are confined to a narrow scope of decision-making [8], 853). This can be extended to thinking, as Sparrow does, about the form that these AWS take. His paper, titled: "Killer Robots," seems to offer at least one possibility of how these sorts of systems may appear in warfare. In this paper, when referring to AWS, what is being discussed is an automated system which can identify and attack combatants without human oversight of those decisions.

These limited autonomous machines are already the subject of a large debate over their permissibility outside of the academic community. Institutions such as the International Committee of the Red Cross (ICRC), the United Nations, and the Geneva Convention refer to such weapons. The ICRC explicitly states that the deployment of autonomous weapons brings risks of harm to civilians and combatants as well as: "raises fundamental ethical concerns for humanity" [5]. In Additional Protocol 1 of the Geneva Convention, Article 35 lays down the "Basic Rules" for the methods and means of warfare. This document further makes reference to prohibitions on weapons that may result in "superfluous injury or unnecessary suffering" [4]. This itself seems to indicate what is a fundamental concern of the Red Cross, that the weapons result in as few harmed people as possible. Further, Article 36 states that new weapons in development must be approved by meeting the criteria laid out in the basic rule from Article 35 (ibid). This state of the debate over autonomous weapons already highlights a divide between the discussion that goes on in high-level policy circles and that which exists in philosophy. The argument presented here can be seen as concurring with these policy documents over and above the philosophical positions offered so far. What is paramount in those documents as well as in the following argument is the prevention of unnecessary suffering, injury, and death.

1.2 The ideal warfare scenario

Three possible ways that warfare can advance exhaust the logical space of the topic: (1) Parties can maintain a human

[2] This outcome is unlikely, and this paper should not be read as a paper in support of the deployment of autonomous weapons. Instead, this argument is meant to motivate those working in the field to search more diligently for the wrong making feature of the deployment of AWS.

[3] Many thanks to reviewers from AI and Ethics for this point, and for this terminology. This is the first of many points in the paper for which I owe both reviewers a large debt of gratitude for their professionalism, knowledge, and interest in this project.

soldier exclusive military, or (2) parties can maintain a hybrid military of human soldiers and AWS, or (3) parties can maintain an exclusive AWS military. Obviously, this is a coarse picture of the possibilities of the future of warfare. There is indeed quite a bit of variance in the way that situation (2) may appear. For the purposes of this paper, those concerns will be bracketed. Theoretically speaking these three options exhaust the possibilities of the relationship between human soldiers and AWS and the thought experiment about ideal warfare relies on the strongest possible version of scenario (3), wherein no humans are involved and thus do not suffer any harm at all.

The basic premise is this: Take any war as it is now and imagine if all the humans could stay home. This is intuitively preferable to sending them off to die or suffer in a war. One way this is achievable, at least in theory, is by developing and deploying autonomous weapons. Further imagine that nations have agreed in advance that this is how wars are to be fought. This would take place in a "civilian free zone" which would ensure that harm only comes to AWS or other government property in that area. Imagine further that the criteria of *jus in bello* are met. That is so because if they were not in fact met, then the situation would be impermissible in a much more traditional sense. It does not matter if a human or a robot violates the *jus in bello* criteria, this constitutes a violation of the rules of warfare all the same.

Further imagine that in the IWS the eleven principles adopted by the United Nations can act as guidelines to the right conduct of the use of autonomous or semi-autonomous weapons [15]. This document offers principles for the (future) use of AWS and often overlaps with the philosophical arguments presented below. Above all it seems to prioritize a meaningful level of control by human agents.

One may object already at this point by saying: This is all well and good, but we are talking about warfare. Is it not the case that the reason for going to war in the first place must be a repugnant aim of one nation against another? Moreover, are we to expect that the losing party in this scenario meekly accept the outcome of this idealized war?[4] Regarding the first point, that war must have a repugnant cause to begin with. It would be a stronger and perhaps more ambitious claim to try and argue that everything about a conflict between nations could be fair. It seems that this amount of fairness and cooperation would eliminate the need for warfare at all. What is meant to be captured by the IWS thought experiment is analogous to the practice of a duel. Within the duel, participants limit the risk of suffering to the two parties, as opposed to a fight in the street or in a tavern. There is also a convention of dueling in another's place, a champion. We can treat the IWS thought experiment as a duel of nations wherein the champions are the various AWS that they possess. A duel of this sort would also necessarily entail the respecting of the outcome by the involved parties. Additionally, respecting the outcome of agreed upon decision processes is a hallmark of any liberal society. This is of course a fanciful proposal, but the point is not that it must actually occur, but that were it to occur it would be just. What is more, were it to occur it would avoid human death and suffering in the resolution of conflicts. Finally, this argument relies on the notion that this would be a better outcome than a traditional war fought by combatants capable of respect, reasons, and responsibility.

Again, the point of this argument is not to argue for the *actual* deployment of AWS. Instead, the aim is to highlight that foregoing accounts of this as an impermissible action must have gone wrong. Indeed, if all the world's soldiers could stay home and let robots do the fighting for them, then who would object to that?

2 Challenges to existing arguments

This section outlines previous attempts to defend the moral impermissibility of deploying AWS. Each argument will be scrutinized for potentially leading to absurd conclusions. Finally, concluding remarks will be offered on the implications of artificial intelligence for the future of warfare.

2.1 Sparrow and responsibility

One of the most influential views on the impermissibility of AWS comes from Robert Sparrow who argues, in two different papers for two different reasons, that AWS will constitute a morally problematic weapon or tool. In his first paper, Sparrow argues that we should not deploy AWS because of the so called "responsibility gap" [13]. Sparrow argues that there are roughly three possible candidates whom one could hold responsible for a war crime committed by an AWS. These are the robot itself, the programmer(s), or the officer who authorized the deployment [13], 67). He argues that each of these come with their own problems as far as being an appropriate candidate for responsibility.

To make his case, Sparrow needs to make a detour through responsibility and just war theory. Sparrow employs the concept of *jus in bello* which requires meeting the conditions of discrimination, proportionality, and necessity (Geneva [3]. To these conditions, Sparrow adds that we should also be committed to a condition of minimal respect to our enemies (ibid). This respect would itself entail that someone should in principle accept responsibility for the deaths of enemy combatants: "The least we owe our enemies is allowing that their lives are of sufficient worth that someone should accept responsibility after their deaths. […] Ensuring

[4] These are strong objections for which I again thank the reviewers.

that someone can be held responsible for each death caused in war is therefore an important requirement of *jus in bello*" (ibid). Not only is holding agents responsible important for just war theory, but on Sparrow's account any sensible deontological or consequentialist account of morality should have this requirement as well. Responsibility lies at the heart of respect, which is a central tenet of a Kantian deontology. Moreover, a lack of holding war criminals responsible could have disastrous consequences for war (ibid).

First, Sparrow tries to avoid the responsibility gap by implementing a human overseer for all decisions an AWS makes. This solution however proves inadequate, as Sparrow argues that any constant human oversight is sure to become a disadvantage as the pace of war increases and enemies seek to target the communication links between AWS and humans [13], 69). Sparrow goes on to show how three candidates (programmer, robot, officer) each fail to be appropriate targets of moral responsibility for war crimes committed by an AWS.[5] The first candidate is the programmer. It is an intuitive notion to hold the programmer of an AWS responsible for a crime. Sparrow argues that this is not appropriate for two reasons. First the programmer is an inappropriate target of responsibility because if a programmer were to disclose the limitations of the system, and an officer accepted these limitations, then it would be the responsibility of the officer, not the programmer. It is likely that any purchase of an AWS would come with such a disclaimer as do most software that is in use today. Second, Sparrow argues that the concept of (even limited) autonomy entails that the human who created the program is not the author of the actions of that program. This is indeed entailed by the concept of AWS. If the opposite were the case, then the weapons would not be autonomous in the relevant sense at all. The entire purpose of developing AWS is that they can act autonomously (ibid), it would then be odd to hold the developer or programmer responsible for their actions.

Next Sparrow considers the military commanding officer. It is common practice to hold commanding officers responsible for the misuse of weapon systems or for the misdeeds of their subordinates. Therefore, to hold those same officers responsible for the misuse of AWS would imply that AWS is no different than those traditional weapons systems or soldiers. It is a fundamental feature of AWS that they are at least semi-autonomous, and thus it seems inappropriate to hold the commanding officer responsible for its actions. "The use of autonomous weapons therefore involves a risk that military personnel will be held responsible for the actions of machines whose decisions they did not control" [13], 71). To take the argument even further than Sparrow's conclusion goes, holding the commanding officer responsible would risk defeating the theoretical reason for deploying AWS in the first place.

Finally, why not the machine itself? For Sparrow, this solution is likewise inadequate. This is because to hold an agent responsible it must be possible to punish or reward that agent. Yet it is obviously not possible to punish or reward a machine. Of course, if the AWS possessed anything like a human intellectual capacity, then punishing it should be possible in principle. One could merely apply the same sort of punishments assigned to guilty human soldiers to the guilty AWS. In other words, the punisher could frustrate its desires, remove its capabilities, or kill it [13], 72). The problem with this solution is that it is difficult to imagine the machine suffering as a result of the aforementioned punishment. A necessary condition of punishment, on Sparrow's view, is that the recipient be made to suffer. Until a machine can be made to suffer, they cannot be appropriate recipients of punishment (ibid). This leads to an interesting argument concerning a sort of trade-off between the usefulness and the permissibility of AWS. For now, it seems that solving the responsibility gap, by making the machine so human-like that it is punishable, may lead to more problems. This paradox will be discussed further in Sect. 3.4. Based on this argument, Sparrow defends his concept of the responsibility gap and concludes that the deployment of AWS is unethical and morally impermissible.

2.2 Sparrow and respect

In a separate paper, Sparrow outlines another reason for thinking AWS are impermissible. Sparrow first follows Nagel in arguing that even in wartime we must engage other human agents as "subjects" [14], 106). Nagel's account argues against a consequentialist ethics regarding warfare:

> A positive account of the matter must begin with the observation that war, conflict, and aggression are relations between persons. The view that it can be wrong to consider merely the overall effect of one's actions on the general welfare comes into prominence when those actions involve relations with others. A man's acts usually affect more people than he deals with directly, and those effects must naturally be considered in his decisions. But if there are special principles governing the manner in which he should treat people, *that* will require special attention to the particular persons toward whom the act is directed, rather than just to its total effect [7], 133 (Emphasis Original)).

[5] An account such as Robillard's [11] might argue that these actors together constitute the AWS which is more of a social construction than an independent autonomous agent at all. In this sense Robillard might argue that the appropriate target of responsibility is all of the actors who make up the system.

Sparrow concurs with Nagel in thinking that there are indeed special principles governing the manner of treating people. Hostile treatment, Nagel argues, can only be justified in terms of facts about the person themselves and not by derivative concerns (ibid). Sparrow follows Nagel, arguing that hostility must be directed at a person *qua* person and be based on features fundamentally about them as a person. Sparrow is happy to adopt a similar framework and say that the combination of *jus in bello* with Nagel's view results in a demand to establish an: "[…] Interpersonal relationship with those who are targets of a lethal attack […]" [14], 107). For Sparrow, what would constitute establishing an interpersonal relationship would be showing enemy combatants a certain level of respect. And moreover precisely what Sparrow means by respect will be determined by the shared social understanding between the two agents involved in the relationship [14], 107).

The problem is that an AWS or robotic soldier cannot engage in the appropriate sort of relationship with its targets. "[…] I have suggested that widespread public revulsion at the idea of autonomous weapons should be interpreted as conveying the belief that the use of AWS is incompatible with such respect"[14], 109). Like the case of punishment, until there is an AWS that can acknowledge another (human) person in the way Nagel and Sparrow have outlined, there remains a gap between the two types of combatants.

The argument put simply is that the requirements of *jus in bello*, combined with Nagel's account of interpersonal relationships, demand that when soldiers engage their enemies their treatment of them: "[…] should be compatible with respect for the humanity of our enemy and that the content of this concept is partially determined by shared social understandings regarding what counts as respectful treatment" [14], 109).[6] Thus, since we socially understand the deployment of AWS to be repulsive then this must indicate the absence of the possibility of such respect being present in the relationship between humans and machines.

2.3 Purves et al. and reasons

Another theory that argues for the impermissibility of AWS concerns whether and what sorts of reasons for action by which AWS might be motivated. The authors ask the question of whether an automated system could ever adequately recreate the sort of human moral deliberation that is expected from soldiers. In their paper, Purves et al. outline a reasons-based account for the impermissibility of deploying AWS. Purves et al. have two arguments that independently support the conclusion that it is morally impermissible to deploy AWS. First, they argue that even an advanced sort of system is not capable of replicating the moral judgment of humans [8], 851). This is so because human moral judgment is not codifiable, and all machines can do is follow codes. Second, they argue that even if the true moral theory were codifiable, the decisions made by an AWS still could not be made for the right sorts of reasons. This is so, because AWS will not make decisions based on any reasons at all (ibid).[7] Each argument is taken up in turn.

The authors' first argument, the anti-codifiability argument, is essentially a denial of the claim that an accurate moral theory could be codifiable in a discrete list of rules [8], 856). This rejection consists in showing that there is popular support among ethicists from many different ethical theories (deontologist, consequentialists, virtue-ethicists) who are committed to the importance of moral judgment and human intuition in any moral deliberation (ibid, 857). Even if one could enumerate a list of principles, that list would essentially consist of an itemization that amounted to an attempt to implement human intuition [9], or exercising good judgment, which is not an act that an AWS could be capable of in principle. Even if there could be a list of moral rules, the application of those rules requires a special sort of deliberation [8], 856). This is the element of the process which would be impossible for an AWS to recreate. Coming up with the list is dubious enough itself, what's more is that even with access to this sort of list a machine is still not capable of deliberating the way a human soldier is.

Purves et al.'s. second argument stands on its own. That is, we can feel the pull of this argument whether or not we are convinced of the first. Suppose a moral theory could be codified contra the anti-codifiability thesis, and AWS could engage in the proper sort of moral deliberation to follow that code. In that case the author's contest that even so the decisions of an AWS could not be made for the right reasons, namely because it would not be made for any reasons at all [8], 860).

The authors offer an interesting thought experiment to demonstrate our reliance on the right sorts of reasons for action even in war, these are the "Racist Soldier" and "Sociopathic Soldier" cases. In these cases, we are asked to imagine first the deployment of a racist soldier whose motivation for being involved in a just war (say a war waged in self-defense) is his vile hatred of the enemy. He thus kills scores of enemy soldiers under the just war conditions, but

[6] As he himself admits, the more recent paper by Brand "goes further" than Sparrow's [1]. For this reason, I have not included it in the above argument, but do acknowledge that Brand's reciprocity functions similarly to Sparrow's respect. An argument against Sparrow's paper, must therefore capture any stronger argument of similar stripes.

[7] This is the position of the authors of the paper being considered in this section. This position is, of course, strictly speaking incorrect. Machines, including AWS, will have reasons for the decisions they take. Those reasons however will not be easy to determine, nor will they be 'human like'. I am grateful to an anonymous referee for this point.

for the reason of his hatred as opposed to out of a sense of duty and with regret for the necessity of the killing. The authors contend that we feel this to be an intuitively bad situation:

> "The likely explanation for this is that, while Racist Soldier abides by the constraints of *jus in bello*, he is acting for the wrong reasons. We believe this judgment can be extended to AWS. Just as it would be wrong to deploy the Racist Soldier, it would be wrong to deploy AWS to the theater of war because AWS would not be acting for the right reasons in making decisions about life and death" [8], 860).

If one finds the conduct of a racist soldier objectionable, then a parallel objection should similarly be considered concerning the deployment of AWS. The authors outline a second case, the Sociopathic Soldier, which is similar in nature. Instead of being a racist, this soldier is simply unmoved by the killing of his fellow human beings and thus has no reasons at all for his actions. Sociopathic soldier allows Purves et al. to make the move from acting for the wrong reasons to merely having no reasons at all. This second scenario is analogous to deploying AWS. If the sociopathic soldier, who kills indiscriminately without any reason, is morally objectionable then so too is the AWS, which in principle cannot have anything that resembles a human reason for action.

3 Automated war and impermissibility

Having covered these accounts of the impermissibility of deploying AWS, the task now is to evaluate them based on the ideal warfare scenario outlined above. The methodology in this section is to take as an assumption the IWS as well as the target account of moral impermissibility. Recall that the IWS thought experiment involves a war fought completely with autonomous weapon systems. This then necessarily entails that no humans suffer or die, which must be a morally permissible case. Further, this scenario is achievable through the deployment of AWS. It has already been seen in warfare that using AWS can lead to minimized human suffering if deployed under the right circumstances. If the thought experiment of the IWS is to be effective, it must also operate under the assumption that the outcome of the conflict would not lead to a tyrannical government inflicting great suffering on civilians, and that other concerns of injustice are bracketed.

This scenario, wherein AWS are used to save lives and turn warfare into a duel, must entail that the deployment of AWS is permissible. This is so because in these cases deploying AWS will have solved the war's wrong-making feature: the death and suffering of those involved. If the foregoing theories still predict that the deployment of AWS is impermissible, this argument contends that this is a contradiction. This contradiction exposes that the theory of impermissibility under consideration picks out another regrettable feature of warfare and not the most important one, namely preventing death and suffering.

3.1 Respect and the IWS

Recall that Robert Sparrow's respect account deems AWS as incapable of showing respect to enemy combatants. Sparrow's account relies on a shared social understanding of what respect entails and the fact that an AWS would be unable to show that respect to a human soldier. In the IWS, wherein there are no human soldiers, the only candidate for respect would be other automated systems. On this account the deployment of AWS is impermissible as it would still fail to show the relevant respect to enemy combatants. This is so because these AWS have no shared social understanding from which one could build an account of respect, and that AWS, per Sparrow's own contention, are incapable of demonstrating that respect.

Taking the ideal warfare scenario and the respect account as assumptions, this argument reaches an obvious contradiction. The argument in this case is as follows:

(1) Assume IWS;
(2) Assume the Respect account is true;
(3) A war without death is morally permissible (1);
(4) So, IWS is morally permissible (from 1,3);
(5) The IWS involves deploying AWS.
(6) The respect account deems deploying an AWS impermissible as it cannot respect its fellow combatants (2);
(7) So, IWS is impermissible (2,5).

Reductio Ad Absurdum (4,7).

As expected, this argument has delivered a contradiction. In light of a contradiction, one must reject one of the premises or an assumption. Defenders of the respect account want to reject (1). Instead, this paper argues that while (1) is unlikely to ever occur, it is certainly logically possible. What's more, (3) and (4) which follow directly from (1), seem to be intuitively correct. What makes war wrong and horrible is the fact that humans suffer and die, if the world could undertake these war exercises without anyone being harmed, then this is obviously preferable to humans suffering and dying. One could further imagine that wars are fought in a simulation, this would also be unobjectionable as it does not involve suffering and death. In this way, (3) follows directly from (1). To reject (3) is really just to reject (1). To solve the problem, one should instead reject (2). We can

draw on intuitions about warfare as it exists compared with the IWS to make this clear. Imagine the argument taking the opposite direction, from having AWS on a battlefield to having none. We would not think that eliminating all AWS from a war thereby rendered warfare morally permissible. In other words, war was a bad thing during wars that involve very limited or rudimentary technology, so why should one think that new technology is really what makes the moral difference. With regard to addressing the contradiction, one should reject (2).

Notice as well that the argument never makes the claim that deploying AWS must be permissible. It only makes this claim derivatively from the notion of achieving the ideal warfare scenario. The absurdity in this conclusion comes from having defined the IWS as morally permissible and the Respect account still rendering the verdict of impermissible. As mentioned above, the issue is that the respect account does indeed pick out a regrettable feature of AWS, but it does not pick out the morally relevant feature that makes their deployment in warfare permissible or impermissible.

3.2 Reasons and the IWS

On the reasons account, offered by Purves. et al., an AWS cannot act for the right reasons, namely because it cannot act for any reasons at all. The capacity of an automated system to act for reasons should remain unchanged by the removal of all human combatants. So, one can assume in this scenario that AWS are still in principle incapable of acting for the right sorts of reasons. In this case the deployment of AWS would still be morally impermissible. But that cannot be so, because for the sake of argument the ideal warfare scenario is assumed and must be morally permissible. The argument in this case is analogous to the respect view:

(1) Assume IWS;
(2) Assume the Reasons account is true;
(3) A war without death is morally permissible (1);
(4) So, IWS is morally permissible (from 1,3);
(5) The IWS involves deploying AWS.
(6) The reasons account deems deploying an AWS impermissible, as it cannot act for the right reasons (2);
(7) So, IWS is impermissible (2,5).

Reductio Ad Absurdum (4,7).

Like the argument, the justification provided is analogous to that of the respect account.

3.3 Responsibility and the IWS

The final candidate is Robert Sparrow's second account of impermissibility, the responsibility gap. Here the wrong-making feature of AWS is that in the event of a malfunction, each of the candidates for punishment would be inappropriate. In this case there is no appropriate agent to hold responsible and therefore the deployment of AWS is impermissible. This case is slightly different from the aforementioned accounts. One way to show this is contradictory would be to expand the definition of the ideal warfare scenario by saying that, by definition, nothing will go wrong. But that seems question begging, and generally weak. It does seem like in the event that a miniscule chance of a mistake happening actualizes, there still would not be anyone to hold responsible for that bad thing. In this case the only reasonable agent to hold responsible would be the AWS itself, which has already been shown to be absurd by Sparrow.

There are two options in responding to this challenge. First, simply accept that the ideal warfare scenario has a responsibility gap. This seems a small price to pay for saving countless lives. It is also not intuitively different from the way warfare works today. Accountability is (unfortunately) rarely at the forefront in the aftermath of mistakes and crimes committed during warfare. Alternatively, one could argue that the IWS should only be implemented with robots that could be held responsible, and thus punished for their crimes. This, however, leads to more problems than it solves.

3.4 Responsibility gaps and the utility-permissibility paradox

This line of inquiry leads to more general questions about the future of artificial intelligence and warfare. As noted at the outset of this paper, extraordinarily advanced autonomous weapons are still a ways off in the future of development. Likewise, this is the case regarding strong or general AI. The question of what advanced artificial intelligence might be useful for in warfare is a separate but interesting one. Moreover, would stronger AI generate more problems than it solves. As Sparrow notes, in order for it to be appropriate to hold machines responsible: "They must make the same sorts of moral and empathic demands upon us as do other (human) people" [13], 72). This points to an emerging Utility-Permissibility paradox.

The main thrust is that the very moment that an AWS would become capable of being held responsible is the moment that it would become analogous to deploying a human soldier, and thus useless as a weapon or tool. If deploying an AWS were analogous to deploying a human soldier in the morally relevant respects, then they would require the same consideration as human soldiers. Moreover, part of the utility of AWS is that they do not require the same consideration as human soldiers. Below is a formalized version of this paradox.

(1) For the deployment of an AWS to be permissible, one must be able to hold that AWS responsible (Sparrow).
(2) Being held responsible involves being capable of human level moral reasoning.
(3) Any agent capable of human level moral reasoning deserves full moral consideration.
(4) An AWS is useful because it can carry out complex tasks without needing to be the subject of full moral consideration.

This paradox involves plausible premises which cannot be mutually true. By Sparrow's account, we are committed to the idea that for the use of an AWS to be permissible it must be subject to responsibility and punishment, and to be subject to responsibility it must be able to recreate the morally relevant human features of decisions, actions, and suffering. Yet it is also agreed that much of the utility of AWS and AI based systems lies in the fact that they can perform tasks that humans perform without receiving the respect that human agents require. In this case then, the moment that AWS become permissible to employ in a warfare scenario on Sparrow's account is also the moment it will become completely inutile as it will then need to be included in the human moral community as a moral agent. This would make it akin to deploying a human soldier, a type of agent who is already deserving of full moral consideration. This paradox shows that one of the only feasible ways to solve the responsibility gap involves creating a much more serious problem. If AWS had the same moral worth as humans, it would obviously be permissible to deploy them. The trouble then, is that it would be utterly pointless.

The central notion of this paradox applies to all AI based or automated systems. These systems are valuable because they allow for the offloading of dangerous or menial labor onto a system which lacks the relevant psychological states to experience unfairness, danger, or boredom. Were these systems to become so advanced that they were indistinguishable from a human agent (at least psychologically speaking) then society would no longer be justified in their deployment. That is so because deploying advanced and agential systems would be analogous to deploying human beings. This leads to the conclusion that there are two options moving forward: Accept that automated and AI systems as they exist fall short of being full agents, and thus accept the limitations that fact carries (responsibility gaps, reasons gaps, respect gaps, etc.); Or work to develop systems which are fully agential and treat these systems as morally analogous to human beings.

4 Conclusion

This paper has argued that the foregoing attempts to describe the impermissibility of autonomous weapon systems lead us to absurd conclusions. Under an assumption about the ideal warfare scenario, each of these accounts lead to the verdict that deploying an AWS would remain impermissible. It was then shown how this led to contradiction, as the ideal warfare scenario must be morally permissible. This led to the rejection of the accounts of the moral impermissibility of deploying autonomous weapons that are currently on offer.

This account may also be read as arguing that there must always be a human decision-maker behind the lethal deployment of AWS. This would be another way to solve the contradictions presented above and would dovetail with the existing literature. In this way then, the arguments above could be read as consistent with, and in a way defending Sparrow's account of the responsibility gap. Moreover, this recommendation coheres with the aforementioned 11 principles of the United Nations and with other policy documents that have informed this paper.

The thrust of the arguments here is not to argue for the actual deployment of autonomous weapons systems. The notion of the IWS which has informed much of the argument here is necessarily an abstraction. This abstraction at best idealizes single instances of battles that occur in broader conflicts. The intention has not been to use the IWS to advocate for the *actual* deployment of AWS on a large scale. The IWS is employed as a thought experiment to demonstrate notable failures of foregoing accounts. Instead, this paper must be read as an attempt to demonstrate, contra the analyses on offer, that the academic community has so far failed to understand the feature of AWS which renders their deployment morally impermissible.

Declarations

Conflict of interest All authors have no conflicts of interest.

References

1. Brand, J.L.M.: Why reciprocity prohibits autonomous weapons systems in war. AI Ethics **3**(2), 619–624 (2023). https://doi.org/10.1007/s43681-022-00193-1
2. Etzioni, A.: Pros and Cons of autonomous weapons systems. In: Happiness is the wrong metric: a liberal communitarian response to populism, pp. 253–263. Springer International Publishing, Switzerland (2018)
3. Geneva Convention.: Geneva convention relative to the protection of civilian persons in time of war. OHCHR. 1949. https://www.ohchr.org/en/instruments-mechanisms/instruments/geneva-convention-relative-protection-civilian-persons-time-war. (1949)

4. International Committee of the Red Cross.: Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (Protocol I), 8 June 1977. (1987). https://ihl-databases.icrc.org/en/ihl-treaties/api-1977/article-35.
5. International Committee of the Red Cross.: International Committee of the Red Cross (ICRC) Position on Autonomous Weapon Systems: ICRC Position and Background Paper. (2022) http://international-review.icrc.org/articles/icrc-position-on-autonomous-weapon-systems-icrc-position-and-background-paper-915.
6. Müller, V.C.: Ethics of artificial intelligence and robotics. In: The stanford encyclopedia of philosophy, edited by Edward, N.Z., Uri Nodelman, F. Metaphysics Research Lab, Stanford University (2023). https://plato.stanford.edu/archives/fall2023/entries/ethics-ai/.
7. Nagel, T.: War and Massacre. Philos Public Aff **1**(2), 123–144 (1972)
8. Purves, D., Jenkins, R., Strawser, B.J.: Autonomous machines, moral judgment, and acting for the right reasons. Ethical Theory Moral Pract **18**(4), 851–872 (2015). https://doi.org/10.1007/s10677-015-9563-y
9. Rawls, J.: A theory of justice. Harvard University Press, Cambridge (1999)
10. Red Cross.: "What you need to know about autonomous weapons." July 22, (2022). https://www.icrc.org/en/document/what-you-need-know-about-autonomous-weapons.
11. Robillard, M.: No such thing as killer robots. J. Appl. Philos. **35**(4), 705–717 (2018). https://doi.org/10.1111/japp.12274
12. Schmitt, M.N.: Autonomous weapon systems and international humanitarian law: a reply to the critics. *Harvard National Security Journal*. https://harvardnsj.org/2013/02/05/autonomous-weapon-systems-and-international-humanitarian-law-a-reply-to-the-critics/. (2013)
13. Sparrow, R.: Killer robots. J. Appl. Philos. **24**(1), 62–77 (2007)
14. Sparrow, R.: Robots and respect: assessing the case against autonomous weapon systems. Ethics Int. Aff. **30**(1), 93–116 (2016). https://doi.org/10.1017/S0892679415000647
15. United Nations.: The convention on certain conventional weapons – UNODA. (1983). https://disarmament.unoda.org/the-convention-on-certain-conventional-weapons/.
16. United States. Dept. of Defense.: Unmanned Systems Roadmap 2007–2032. (2007) https://rosap.ntl.bts.gov/view/dot/18247.
17. Young, G.: Should autonomous weapons need a reason to kill? J. Appl. Philos.ppl. Philos. **39**(5), 886–900 (2022). https://doi.org/10.1111/japp.12597