Journal of
APPLIED PHILOSOPHY

Journal of Applied Philosophy
doi: 10.1111/japp.12274

No Such Thing as Killer Robots

MICHAEL ROBILLARD

ABSTRACT *There have been two recent strands of argument (one offered by Rob Sparrow and the other offered by Duncan Purves, Ryan Jenkins, and Bradley Strawser) arguing for* the pro *tanto impermissibility of fully autonomous weapon systems (AWS). On Sparrow's view, AWS are impermissible because they generate a morally problematic 'responsibility gap'. According to Purves* et al.*, AWS are impermissible because moral reasoning is not codifiable and because AWS are incapable of acting for the 'right' reasons. I contend that these arguments are flawed and that AWS are not morally problematic* in principle. *Specifically, I contend that these arguments presuppose an incoherent conception of an AWS as somehow making genuine decisions but not being morally responsible for those very same decisions. Rather than conceiving of AWS in this way, I argue that an AWS is* either *a socially-constructed institution that has been physically instantiated* or *it is a genuine agent. If it is the former, then we should treat AWS as we do any other collective action problem. If it is the latter, then we should treat AWS as responsibility-bearers, but also as bearers of rights and/or interests. To reject this disjunction is not only conceptually incoherent but also potentially morally dangerous.*

Introduction

There has been much debate recently, both within academia and without, regarding the ethics of autonomous weapons systems (AWS) and the debate does not appear to be waning anytime soon. While many of the more popular arguments against the use of AWS have focused on more contingent or downstream worries, three prominent arguments within the contemporary ethics literature have emerged for why AWS are morally problematic *in principle*.[1] One argument, advanced by Rob Sparrow, is based on the idea that AWS are *pro tanto* impermissible because their use necessarily generates a 'responsibility gap' that is morally troubling. The other two arguments, advanced by Duncan Purves, Ryan Jenkins, and Bradley Strawser are based on the ideas that AWS are *pro tanto* impermissible because moral reasoning resists algorithmic codification and because AWS are not the kinds of things that are capable of being motivated by the right kinds of reasons. In this article, I challenge all three of these arguments against the use of AWS. Indeed, my claim here is that AWS are not morally problematic *in principle*.[2] This is so, I argue, because there are no intentions or decisions *of the AWS itself* separate from or in addition to the set of previous human decisions and intentions physically instantiated in the form of the machine's software and implementation. For us to think otherwise is actually the greater moral danger since doing so discourages programmers, designers, and decision-makers from

carefully considering their actual moral responsibilities and causal contributions within the causal chain.

Despite this, I nonetheless believe there are very good reasons for states to refrain from using AWS in war, but that these reasons ultimately find moral grounding in familiar normative concepts of collective responsibility, collective action, and risk under epistemic uncertainty. In Section 1, I give a brief overview of what philosophers, policymakers, and military thinkers might mean when they refer to a weapon as being 'autonomous'. In Section 2, I give an explication of the three philosophical arguments for the impermissibility of AWS introduced above and offer several critical objections. In Section 3, I explore some of the common ontological assumptions about AWS and argue that much of the current AWS debate is ill conceived. I argue here that AWS are functionally isomorphic and functionally equivalent *to social institutions* (i.e. corporations, governments, etc.) and that we should therefore treat them no differently morally speaking. Call this the 'institutional view' of AWS. Lastly, in Section 4, in light of this institutional view, I conclude by briefly noting some of the moral dangers of ascribing moral responsibility based on a concept of AWS that is fundamentally incoherent.

1. Terminology

Let us first begin by getting clearer on what exactly we mean by the term 'autonomous weapon'. As Burri notes, presently, international humanitarian law has failed to give an explicit legal definition of AWS. However, the US Department of Defense, in a 2012 directive, offered the following definition:

> Autonomous weapon system. A weapon system that, once activated, can select and engage targets without further intervention by a human operator. This includes human-supervised autonomous weapon systems that are designed to allow human operators to override operation of the weapon system, but can select and engage targets without further human input after activation.[3]

Providing a somewhat similar account, Scharre and Horowitz further point out that AWS can be defined in terms of their causal relationship to human operators. Hence, weapons where a human is 'in the loop' (and thereby plays a necessary causal role in the weapon's proper functioning) do not count as AWS. However, weapons where, once activated, a human is 'on the loop' (causally unnecessary for the weapon's proper function, but monitoring and able to intervene) and 'out of the loop' (causally unnecessary for the weapon's proper function and unable to intervene) will count as AWS.[4]

Offering a more succinct definition, Sparrow says of AWS that, 'their actions originate in them and reflect their ends. Furthermore, in a fully autonomous agent, these ends are ends that they have themselves, in some sense, chosen.'[5]

Lastly, building on Sparrow's definition of AWS, Purves *et al.* cite the Tjerk de Greef's capabilities scale where the type of technology they are envisioning would be classified as having a 'High' level of autonomy on De Greef's scale. They write, 'we are imagining weapons that can act in such a way that "the computer decides everything, acts autonomously, ignoring the human."'[6]

From these four definitions of AWS, we can begin to see a general pattern and family resemblance among them. According to these views, AWS possess the features of being able to function independently of human operators, behave in ways that resist precise prediction, and are capable of learning and adapting to their environments in novel ways. Most importantly though, for the purposes of this article, and for the AWS debate in general, AWS are presumed to make authentic, *sui generis* decisions that are non-reducible to their formal programming and therefore uniquely their own. In other words, AWS are presumed to be genuine *agents*, ostensibly responsive to epistemic and (possibly) moral reasons, and hence not mere *mimics* of agency.[7] Indeed, were it that case that AWS were just mimicking authentic decision-making, but were not actually responsive to epistemic and (possibly) moral reasons, then I am not sure in what coherent sense we could then regard them as 'genuine' decision-makers. What's more, such mimicry would fail to generate the kinds of 'in principle' worries about the apportioning responsibility that ethicists like Sparrow et al. have been raising. That being said, much of the present AWS debate seems to tacitly treat AWS as sort of 'quasi-agents' capable of making (not just mimicking) genuine decisions but somehow not being responsive to epistemic and (possibly) moral reasons. This concept of a quasi-agent, both in general as well as specifically applied AWS, I contend, is fundamentally incoherent.

2. Gaps, Codification, and Reasons

Gaps

Let us now look at Sparrow's argument against the use of AWS. Borrowing from Purves *et al.*, we can summarise Sparrow's argument as follows.

1. Waging war requires that we are able to justly hold someone morally responsible for the deaths of enemy combatants that we cause.
2. Neither the programmer of an AWS nor its commanding officer could justly be held morally responsible for the deaths of enemy combatants caused by AWS.
3. We could not, as a matter of conceptual possibility, hold an AWS *itself* morally responsible for its actions, including its actions that cause the deaths of enemy combatants.
4. There are no other plausible candidates for whom we might hold morally responsible for the deaths of enemy combatants caused by AWS.
5. Therefore, there is no one whom we may justly hold responsible for the deaths of enemy combatants caused by AWS.
6. Therefore, it is impermissible to wage war through the use of AWS. To do so would be to 'treat our enemy like vermin, as though they may be exterminated without moral regard at all'.[8]

Sparrow argues that the capacity 'to be held responsible' is a necessary feature for moral responsibility. He further argues that for one to be capable of being held responsible one must have the capacity to be punished, and indeed, the capacity to suffer. Sparrow writes,

> To hold that someone is morally responsible is to hold that they are the appropriate locus of blame or praise and consequently for punishment or reward. A crucial condition of the appropriateness of punishment or reward is the conceptual possibility of these treatments. Thus in order to be able to hold a machine morally responsible for its actions it must be possible for us to imagine punishing or rewarding it. Yet how would we go about punishing or rewarding a machine? In order for a machine to be capable of being punished then, it must be possible for it to be said to suffer.[9]

The fact that an AWS cannot suffer and hence, cannot be punished, argues Sparrow, therefore makes it the case that AWS cannot be a proper bearer of moral responsibility. I am not sure why the capacity to suffer or to be punished should be a necessary feature for moral responsibility or why robots couldn't, in principle, be made to suffer.[10] I am similarly sceptical of the idea that robots couldn't be bearers of responsibility without having to be capable of being 'held' responsible. As Burri keenly points out,

> Once AWS have goals and desires of their own, why wouldn't they suffer if they had these thwarted?[11]

She continues,

> Nor am I convinced that an agent cannot be morally responsible unless it is conceptually possible to hold them responsible. Consider the case of a successful suicide bomber who wrongfully blows up ten innocent children along with herself. While it may well be impossible to hold dead people morally responsible, the suicide bomber seems nevertheless morally responsible and blameworthy for what she did.[12]

This seems to me a fair critique of Sparrow's view. Indeed, sometimes causal responsibility and moral responsibility come apart in cases which we would consider genuine accidents, where no reasonable person could have foreseen, predicted, or prevented a given harm. Furthermore, moral responsibility and the capacity to be 'held' responsible further come apart such as in the case of the suicide bomber. Sparrow's account does not seem to be sensitive to this further distinction.[13]

Sparrow therefore faces a disjunction; either the AWS is an agent or it isn't. If it isn't, then premise 2 is either false and moral responsibility falls back on the commanders and programmers to the extent that they knew or should have known the harm they were contributing to and could have done otherwise *or*, premise 2 is true but vacuous because the harm was a result of a genuine accident. Accordingly, in such a case, there will in fact be a 'gap' between causal and moral responsibility, but it will be a non-problematic one. If the harm is the result of a genuine accident that no rational agent could have possibly foreseen, then it will therefore be no different in kind, morally speaking, from any other battlefield case involving unforeseeable malfunctions of weapons systems or equipment in war.[14] Admittedly, there might indeed be an intransitivity between the final output of the AWS and the individual intentions of each of the group's individual members, but this intransitivity will be wholly in virtue of the collective nature of the group that built, programmed, and implemented the AWS (hence a collective action problem *in general*) and not in virtue of the machine's

apparent autonomy.[15] Remove the collectivity from the equation, and all intransitivity and gappy-ness goes away. To see this, consider the following case,

> *Mastermind*
>
> *A computer mastermind assembles and programs his own personal autonomous weapons system completely from scratch and then launches it onto the battlefield. Despite his precautions and confidence in its targeting abilities, the AWS kills several innocent civilians.*[16]

Given such a case, if the harm was not the result of an unforeseeable accident (say, from some yet to be discovered causal feature at the level of quantum physics), do we still believe that there is any gap whatsoever in terms of moral responsibility simply because an AWS was involved? What's more, if the mastermind were put on trial and he claimed innocence due to the fact that he didn't expect that the robot would react the way that it did, would we find this to be a reasonable excuse?[17] Hardly. Certainly, the locus of moral responsibility for the harm, if it is to be located anywhere at all, is to be located fully with the mastermind and with no gap in responsibility left unaccounted for despite the weapon being 'autonomous'. Otherwise, if we do still think that a responsibility gap persists despite the removal of all collectivity and despite the harm not being the result of a genuine accident, then it would follow that an agent, in principle, could be blameless for a harm caused by an AWS *completely of his own making* were he to simply make the AWS sufficiently complex such that it outstripped his capacity to predict its future behaviours. This can't be right. Rather, when it comes to AWS as we normally conceive of them, any gaps in moral responsibility for a given harm, if not the result of a genuine accident, will be fully explainable in terms of the *collective nature* of how the AWS was built, programmed, and employed and not because of anything having to do with its supposed autonomy.

Otherwise, the AWS is an actual decision-maker and therefore a proper bearer of moral responsibility for its actual decisions. Consequently, premise 3 of Sparrow's argument would therefore be false. If the AWS is a genuine agent, then we need to take seriously the thought that it would therefore have rights or minimally, interests. Indeed, if an AWS were capable of making genuine decisions, then precisely in virtue of such agency and decision-making power we should thereby regard it as the new locus of moral responsibility for the given harm. Even if it turned out to only be a partial agent like an animal or a small child, then the AWS would still be the bearer of *some* degree of responsibility to the extent that it had some degree of agency, if only minimally. The bulk of blame would then fall back on the AWS's programmers, chain of command, etc. as it did before.

Indeed, if genuine agency were to emerge *ex nihilo* from the machine's programming structure, then such a state of affairs would not only mean that there would be no responsibility gap (provided, once again, that the harm was not the result of a genuine accident) it would also mean that the AWS would be a proper bearer of rights like a person, or minimally, a bearer of interests like some animals. Accordingly, this would raise difficult moral questions about what we owed to such machines and what rights violations or harms we might be unjustly bringing upon them when we shut them off, reprogram them, or make them fight our wars for us. These entailments, about

machine rights and interests and our concomitant duties to them are all implications one would necessarily have to accept were one to ascribe genuine agency to an AWS.

Codification and Reasons

Let us next move on to consider the two arguments against the use of AWS offered by Purves *et al.* Their first argument is based on the idea that authentic moral reasoning cannot, in principle, be reduced or codified into a set of finite decision-procedures or algorithms. Consequently, they propose what they refer to as the 'anti-codifiability thesis'. They write,

> The *codifiability thesis* is the claim that the true moral theory could be captured in universal rules that the morally uneducated person could competently apply in any situation. The *anti-codifiability thesis* is simply the denial of this claim, which entails that some moral judgment on the part of the agent is necessary.[18]

The second related argument proposed by Purves *et al.* is based on the idea that AWS are morally problematic since they are not capable, in principle, of acting for the right moral reasons. They write,

> Even if the anti-codifiability thesis is true, our first objection to AWS succeeds only if we place disproportionate disvalue on genuine moral mistakes compared with empirical and practical mistakes. Our second objection to the deployment of AWS supposes that AI could become as good as or better than humans at making moral decisions, but contends that their decisions would be morally deficient in the following respect: they could not be made for the *right reasons*. This provides the missing theoretical basis for the disproportionate disvalue that our first argument places on genuine moral mistakes.[19]

Accordingly, Purves *et al.* suggest that were we to have the choice of either deploying, for a just cause, a racist soldier, who abided by *jus in bello* constraints but was motivated to kill for reasons of racial bias, or a non-racist soldier, who abided by *jus in bello* constraints because of genuine regard for their moral import, all other things being equal, we would have a strong moral reason to choose the non-racist soldier over the racist soldier. The likely explanation for this intuition, argues Purves *et al.* is that, while the racist soldier abides by the constraints of *jus in bello*, he is *acting for the wrong reasons*.[20] This judgment, they argue, can be extended to AWS.

I do not disagree with any of the above claims about the nature of both moral reasons and moral reasoning. Indeed, I agree with Purves *et al.* that morality is not codifiable. I also agree with them insofar as I believe that fully just actions must follow from the right moral reasons. Where I take issue with them is in their thinking that the actions *of the AWS* should be the proper object for which these moral concerns actually apply. Put another way, while I do not think morality is subject to formal codification, I do not think that the apparent 'decisions' of the AWS stand as something metaphysically distinct from the set of prior decisions made by its human designers, programmers and implementers, decisions that ostensibly *do* satisfy the conditions for counting as genuine moral decisions. Even if the codified decision-procedures of the AWS amount to only a truncated or simplified version of the programmers' moral

decision-making for anticipated future contexts the AWS might some day find itself in, *the act of codifying* those decision-procedures into the machine's software will itself still be a genuine moral decision. The same goes for the condition of being motivated by the right kinds of reasons.[21]

Accordingly, concerns about the morality of AWS would all amount to *contingent* worries at best (worries about apportioning moral responsibility among a collective of agents, determining epistemic responsibility, weighing risks in conditions of epistemic uncertainty, etc.). There would, however, be nothing wrong *in principle* about using AWS. That being said, were the moral stakes high enough, use of AWS over the use of soldiers, under some set of conditions, could be conceivably permissible as well as conceivably obligatory.

What obfuscates the situation immensely is the highly collective nature of the machine's programming, coupled with the extreme lag-time between the morally informed decisions of the programmers and implementers and the eventual real-world actions of the AWS. Indeed, the decision to deploy an AWS in a particular context, *in principle*, is not any more or any less insensitive to moral reasons than the decision to place a land mine in a particular context. These actions, at base, are both human decisions, responsive to moral reasons through and through. The only difference that separates these actions from more familiar actions on the battlefield is that there is a pronounced lag-time between the latent human decisions built into the causal architecture of the weapons system itself and the anticipated combat effect of that weapon system that later eventuates.

This pronounced lag-time combined with the AWS's collective nature and complex interface has led philosophers to mistake the set of human decisions instantiated in the form of the machine's programming and implementation as summing up into an *additional* set of decisions (the AWS's decisions) that is metaphysically distinct. However, after we have summed the total set of human decisions instantiated in the machine's software and implementation, I am hard pressed to see what genuine decisions would be left over. Indeed, the AWS might and likely will behave in ways that we cannot predict, but these actions will not fail to be logical entailments of the initial set of programming decisions encoded in its software combined with the contingencies of its unique, unanticipated environment.[22]

As a final point worth noting, despite the arguments here given, one might still think that we can coherently conceive of the relationship between programmers and AWS as analogous to a parent's relationship to a small child, insofar as both the AWS and the child can be seen as *partial* agents. This analogy, however, does not hold. For instance, imagine a radical terrorist parent who trains a small child to carry a bomb into a marketplace, to locate a person whose appearance looks least ethnically similar to their own, and to then press the detonator button. While the child's individual decisions regarding where and when to detonate the bomb would still count as genuine decisions, we would not say that the parent was somehow absolved of moral responsibility for the resulting harm in virtue of the child's authentic, though partial, agency. In other words no responsibility 'gap' would be present. Accordingly, this analogy breaks down if we think the AWS's 'parents' (i.e. its programmers) should somehow be regarded any differently morally speaking.[23]

3. The Institutional View

Consider an adaptation of Block's famous *Chinese Nation* thought-experiment offered as a *reductio* against a functionalist account of mind.

> *China-Bot*
>
> *Suppose that the entire nation of China was reorganised to perfectly simulate the functional structure of the software of an AWS (to include its learning programs). Each Chinese person follows a finite set of rules spelled out on a piece of paper comparable to that of one of the AWS's subprograms and then communicates by special two-way radio in the corresponding way to the other Chinese people who are doing the same thing. The software program (being realised by the entire nation of China) would then be connected via radio to an actual AWS in the real world, one that provided the sensory inputs and behavioural outputs of the program.*[24]

Imagine that the AWS was released onto the battlefield where it then killed an innocent civilian. Given such a case, how would we assign moral responsibility for such harm? For one, we would want to know what each Chinese person knew and was capable of knowing about his or her causal contribution to the potential harm and what he or she could have causally affected. We would likewise want to know what the designers of the program knew or could have reasonably known or done differently with regard to what eventualities their software entailed once it was implemented. However, once we did so, would there be any additional decision-making or moral responsibility left to account for? I think the answer here is clearly no. What would we say then about a second case where each Chinese citizen, for reasons of fatigue, built a simple mechanical device that could perform the identical low-level physical tasks as their bodies were performing and then swapped themselves out with the device? Would we ascribe moral responsibility any differently in this case simply because the mechanical device built by each citizen was performing the low-level tasks rather than the Chinese citizen himself? I see no reason why we should. Even if there is a significant lag-time between when the citizens swap themselves out for the devices and when the devices actually perform their actions, there still seems to be no shift in the moral landscape whatsoever.

How then is this second scenario any different from when software programmers program their software (i.e. build their mechanical devices) and then step away? Once again, moral responsibility seems fully exhausted here once we ascribe it to the individual low-level programmers and to the macro-level designers. Accordingly, there simply isn't any need, reason, or room to then ascribe additional responsibility *to the AWS itself*.

In the same way that there is no additional metaphysical group agent 'China' over and above the individual Chinese people whose actions serve to functionally realise the AWS's software program, so too, I argue, there is no additional metaphysical agent 'the AWS'. Indeed, were we to have the language of god, we could speak and understand the exceptionally long conjunction of 'citizen A's decision *and* citizen B's decision *and* citizen C's decision. . .etc.' without any need whatsoever to refer to the set of decisions made by 'the nation of China', or alternatively, by 'the AWS'.

One objection that could be given here is that I have not actually given an argument yet against collectivism. All I have pointed out thus far is that if one is to ascribe genuine agency to a 'normal' AWS program, then they must also ascribe agency to the entity over and above of the collection of Chinese citizens realising the AWS's program. Consequently, one would thereby inherit all of the metaphysical baggage and bizarre mereological problems entailed by collectivism. But most collectivists would likely be completely fine with such implications.

That being said, even if one were to ascribe to a collectivist view of group agency, he or she would still face difficulties with our previous case. While collectivists argue that moral responsibility for a group action does not reduce to a mere aggregation of the moral responsibility of the group's individual members, most if not all collectivists will nonetheless argue that individuals in the group will at least bear *some* degree of responsibility for their causal contributions to the group act.[25] Conversely, those who posit that the AWS is an agent over and above its human programming and implementation argue that *no* responsibility whatsoever obtains for the programmers and implementers with respect to the AWS's genuinely independent actions. Consequently, the collectivist who ascribes to *both* the notion that there is a metaphysical agent over and above the group of individuals in a collective and the notion that there is a metaphysical entity over and above the AWS's programming and implantation, will yield a contradiction when trying to account for our *China-Bot* case. Indeed, they will not be able to have it both ways.

All this being said, put crudely, a fully autonomous weapon system is nothing more than a socially constructed institution that has been instantiated in physical form. In other words, much like a corporation or a government, an AWS is a socially constructed finite set of decision-procedures, made by human agents, for a finite set of intended contexts.[26] The only difference is that rather than having human bodies realising such decision-procedures, we have gotten the world to cooperate and conform such that those decision procedures are now realised by the laws of physics within microprocessors. Accordingly, the AWS is not a separate agent unto itself.

But such a difference between AWS and standard social institutions is not something to be taken lightly. This is so precisely because social institutions are, in a sense, 'social fictions' that only find causal efficacy in the world when presently-existing human agents decide to implement them and regard them as legitimate. Accordingly, social institutions (like governments, corporations, banks, economic currency, etc.) can be quickly revised and/or terminated altogether simply by the present intentions of participating agents. With AWS, however, since the institution's decision procedures have been concretised in physical form earlier in the past during its initial time of programming, its causal efficacy in the world can persist despite the absence of present intentions from present agents. In other words, the AWS's efficacy in the world is not dependent at all upon the intentions and participation of present agents (hopefully it is in the sense that we have a kill switch or ways to tell it to stop doing what it does), though its efficacy is dependent upon the intentions of programmers in the past. Having said that, this difference still would not change how we apportioned moral responsibility for the AWS's actions.

Moving on, one response to my categorisation of AWS might be to say that while it is true that AWS do not presently have genuine agency of their own, they at least have what amounts to a sort of quasi-agency that could serve to mediate and possibly

distort the original decisions and intentions of its human programmers. Sparrow says something to this effect when he states,

> Even if the machine is not a full moral agent, it is tempting to think that it might be an 'artificial agent' with sufficient agency, or a simulacrum of such, to problematize the 'transmission' of intention.[27]

I still do not think this is correct. Once again, if we stop to consider our *China-Bot* case, we see that the AWS's programming is fully constitutive of the programmers' decisions and intentions. Hence, there is nothing left of what makes up the AWS's programming to *mediate* the original decisions and intentions of its programmers. Rather, the AWS's programing simply *is* the original decisions and intentions of its programmers in total. And while some philosophers might respond here by pointing to the fact that there is an *intransitivity* that exists between the individual decisions of the programmers and the eventual behavioural output of the AWS, such intransitivity, I argue, just seems to be a run of the mill problem that faces any collective action whatsoever and is not therefore one that is at all unique to just AWS.

4. Moral Dangers

Much of the current debate surrounding the permissibility of AWS seems to be taking place in a rather peculiar metaphysical space whereby the AWS somehow counts as a genuine decision-maker but somehow does *not* count as a bearer of responsibility for its genuine decisions. For the reasons I've given in this article, I believe that conceiving of AWS in this way is ultimately incoherent. What's more, I believe that for us to conceive of AWS in this way is also morally dangerous. If the AWS is a genuine agent, then we are doing something significantly wrong by not taking seriously the notion that AWS, in virtue of being agents, could indeed have rights and/or interests.[28] Accordingly, if AWS are in fact genuine agents with rights and/or interests, then turning them off, reprogramming them, or having them fight our wars for us would count as a serious wrong. What's more, it would then be logically conceivable to imagine a future world, 250 years from now, that succeeded in creating such robot agents, complete with the full set of rational, moral, and epistemic faculties as any human today. In other words, morally speaking, they would be just like human soldiers today. Consequently, the permissibility of their deployment would be *exactly* the same.

If, however, AWS are not genuine agents, then moral responsibility for a potential harm an AWS might cause would therefore fall back on the complex system of human programmers and implementers and would thereby warrant closer examination of the system's organisational and causal structure so as to prevent future harms from occurring. This odd, in-between grey zone in which the current AWS debate largely finds itself kills any and all regard for further examination of either of these sets of moral considerations.

If we think prevention of harm is at all important, then whenever a harm does actually occur, its occurrence demands of us that we re-examine more closely our own actions, assumptions, and behaviours as well as those of others in order to find an explanation for why things occurred as they did. Ostensibly by discerning such an explanation, we can then take active corrective measures to see that the same harm

does not repeat itself. If instead, we view an AWS's actions as being something radically distinct from the human decision-making that led up to and went into it, then doing so seems to absolve all potential contributors to the harm of any personal responsibility to reflect upon their actions and the actions others and to consider whether or not there was any negligence or wrongdoing that occurred. This is indeed *not* the attitude we want to be inspiring in designers, programmers, and implementers of AWS. However, by conceiving of AWS in the problematic way that I have here critiqued, this radical detachment from both responsibility and outcome is just such an attitude that the present AWS debate is inspiring.

What's more, viewing AWS in this way seems to similarly conceptually render potential contributors to the harm immune to all considerations of liability and/or desert. Intentionally benevolent programming, intentionally wicked programming, caution, negligence; all of these human inputs going into the AWS's creation seem to not matter at all, morally speaking, if the output (the AWS's authentic decision) is something metaphysically distinct and totally separate.

This would then mean that the supposed responsibility gap would show itself to be more of a responsibility black hole, negating all responsibility of those who contributed to its creation.[29] This would also entail that standardly unjust wars could be made just if simply fought totally by means of AWS, since the *real* deciders of violence would then be the AWSs themselves and not the human programmers or implementers of the state sending them forward.[30] Ironically, by conceiving of AWSs as being decision-makers unto themselves, but not actual agents capable of being bearers of moral responsibility, what follows then is a moral permissiveness that is far more dangerous than any potential *in bello* mishap. As such, those who worry about the potential moral dangers of AWS (and to be certain there are many) should therefore *embrace* the idea an AWS is not an authentic decision-maker, not shy away from it.

Conclusion

Either an AWS is an agent or it is not. It is not both and it is not neither. Accordingly, persons within the current AWS debate must choose between one of these two disjuncts and then accept the conceptual and moral entailments that then follow. To conceive of an AWS as anything else is not only conceptually incoherent for the reasons here outlined, but it is also morally dangerous. It is morally dangerous precisely because it shifts focus off of the only two possible cases where responsibility could conceivably obtain and instead places the debate in a peculiar in-between space that is conceptually intractable. Indeed, often in philosophy, avoidance of sharp, binary thinking is both admirable and helpful. This however is not one of those times.

Put simply, if the AWS is an agent (either partial or full-fledged), then it is therefore capable of being the bearer of moral responsibility to a degree commensurate with its agency and capacity for knowing the consequences of its agency. This would mean that it would also be capable of possessing rights and/or interests; rights and interests that we would therefore be obligated to respect and honour. In other words, in being a moral agent, an AWS would necessarily be capable of being a moral *patient* as well. However, if the AWS is not an agent, then moral responsibility, if it is to be placed anywhere, must fall back *entirely* on the set of human designers, programmers, and

implementers. I have made the argument here that the latter description of an AWS is the correct one and that moral responsibility, if it is to be apportioned at all, must go to the set of humans responsible for its creation, programming, and deployment. That being said, I'm more than happy to concede that I could be altogether wrong about the AWS's metaphysical status, and that something about the machine's learning algorithms could make it such that its decisions were truly its own. But to be clear, if that were indeed the case, *then the AWS would then count as a genuine agent*, not as some mysterious ontological anomaly, and not as some weird moral chimera where all of our usual thinking about how agency and responsibility works somehow breaks down.

It is therefore important to note that either of the disjuncts here mentioned come with certain benefits as well as certain costs. If the AWS is an actual agent, then its creation would mean that the promise of strong A.I. had finally arrived and that knowledge of our metaphysical and moral universe had greatly expanded. This would entail however, a world of much greater moral demandingness than at present. If however, the alternative account of AWS that I have here proposed is correct, then it would mean that we would have to come to terms with a sobering truth that is at once harsh but also potentially empowering; namely, that there are, in fact, no machines who might one day conspire against us, no automata who might some day decide to revolt, in other words, no killer robots, just us.[31]

Michael Robillard, Uehiro Center for Practical Ethics, University of Oxford, 16-17 St. Ebbes St. Oxford, OX1 1PT, UK. michael.robillard@philosophy.ox.ac.uk

NOTES

1 I should note that Sparrow and Purves *et al.* are not the only philosophers who have recently argued against AWS for moral reasons. Tom Simpson and Mark Cockelburgh, for instance, also consider some of the morally problematic features of AWS. However, for reasons of scope, this article will focus upon what I believe to be the three best 'in principle' arguments for the *pro tanto* impermissibility of AWS.

2 For similar views, see Susanne Burri, 'In defense of killer robots' in R. Jenkins, M. Robillard & B.J. Strawser (eds) *Who Should Die: Liability and Killing in War* (Oxford: Oxford University Press, forthcoming) and Steve Kershnar, 'Autonomous weapons pose no moral problem', in B.J. Strawser (ed.) *Killing by Remote Control: The Ethics of an Unmanned Military* (New York: Oxford University Press, 2013) pp. 229–245.

3 United States of America Department of Defense (2012). Directive Number 3000.09 on Autonomy in Weapon Systems. Available online at http:// http://www.dtic.mil/whs/directives/corres/pdf/300009p.pdf.

4 Paul Scharre & Michael Horowitz, 'An introduction to autonomy in weapon systems' (2015). Available online at http://www.cnas.org/ intro-to-autonomy-in-weapon-systems#.VyTXlvkrKik. 8.

5 Rob Sparrow, 'Killer robots,' *Journal of Applied Philosophy* 24,1 (2007): 62–77, at p. 65.

6 Duncan Purves, Ryan Jenkins & Bradley Jay Strawser, 'Autonomous machines, moral judgment, and acting for the right reasons,' *Ethical Theory and Moral Practice* 18 (2015): 851–872 at p. 854.

7 I say 'possibly' moral reasons here since I want to remain agnostic about which is the correct theory of reasons (externalism versus internalism).

8 Sparrow's argument summarised in Purves, Jenkins & Strawser op. cit., p. 855. It is important to note here that while Purves *et al.* summarise Sparrow's account explicitly in terms of harms to enemy combatants, Sparrow's view is not limited to just this group of actors and indeed extends to consider potential AWS harms to non-combatants and friendly soldiers as well.

9 Sparrow 2007 op. cit., pp. 71–72.

10 The Society for the Ethical Treatment of Reinforcement Learners (http://petrl.org/) for instance proceeds from the assumption that algorithms are indeed capable of suffering.

11 Burri op. cit., p. 16. Also, technically Burri uses the term 'Lethal Autonomous Robot' (LAR) instead of 'AWS'.
12 Burri op. cit., p. 16.
13 For a plausible example of moral responsibility without blameworthiness, see McMahan's 'Conscientious Driver' case in Jeff McMahan, 'The basis of moral liability to defensive killing,' *Philosophical Issues* 15,1 (2015): 386–405.
14 For a good example of a non-problematic gap between causal responsibility and moral responsibility see Thomson's case, 'Day's End' in Judith Jarvis Thomson, *The Realm of Rights* (Cambridge, MA: Harvard University Press, 1990), p. 229.
15 For an example of such a collective action problem, see Phillip Pettit & Christian List, 'Aggregating sets of judgements: Two impossibility results compared', *Synthese* 140 (2002): 207–235, as well as Parfit's 'Harmless Torturers' case in Derek Parfit, *Reasons and Persons* (Oxford: Clarendon Press, 1984).
16 Let us stipulate that the mastermind has built the AWS for good moral reasons and has not merely created a violent autonomous weapon just for the fun of it.
17 I thank Ryan Jenkins for the development of this case.
18 Purves, Jenkins & Strawser op. cit., p. 858.
19 Purves, Jenkins & Strawser op. cit., p. 858.
20 Purves, Jenkins & Strawser op. cit., p. 866.
21 While the discussion of reasons internalism versus externalism as it relates to agency falls outside the scope of this article, one might think that if a Kantian internalist account of reasons was correct (something like Korsgaard's view) then the type of AWS that Purves *et al*. posit, one that only apprehends epistemic and practical reasons but not moral reasons, would be logically impossible. That is to say that on such a view of reasons, moral reasons would simply be a logical entailment falling out of mere practical reasons, thereby making it impossible for an agent to apprehend the latter without thereby apprehending the former as well. For more on this view see Christine Korsgaard, 'Skepticism about practical reasons,' *The Journal of Philosophy* 83,1 (1986): 5–25.
22 An important point worth noting here, one which Roff points out, is that the future AWS 'environment' that will likely be the most morally problematic will not be one of three-dimensional space, but rather one where the AWS's 'agency' is widely distributed across multiple servers, within cyberspace, and within future military command structures. Hence, the real worry regarding AWS will not simply be that of the tactical-level killer robot on the battlefield, but rather, that of the strategic-level cyber-general in cyberspace. See Heather M Roff, 'The strategic robot problem: Lethal autonomous weapons in war', *Journal of Military Ethics* 13,3 (2014): 211–227.
23 I thank Bill Lycan for the development of this case.
24 See Ned Block, 'Troubles with functionalism,' *Minnesota Studies in The Philosophy of Science* 9 (1978): 261–325.
25 Consider a real-world case like the BP oil spill of 2010. Even if moral responsibility for the collective act does not reduce to an aggregate of the moral responsibility for each of the group's individual members, certainly we would still think that we can hold the CEO of BP at least *partially* responsible for the disaster. What's more, we would also think that the CEO's partial responsibility would be greater than that of the average low-level shift worker.
26 For an in-depth investigation into how institutions are built up out of mutually agreed upon social and linguistic rules, see John Searle, *The Construction of Social Reality* (New York: The Free Press, 1995).
27 Rob Sparrow, 'Robots and respect,' *Ethics and International Affairs* 30,1 (2016): 93–116, at p. 108.
28 Some might argue that the idea of a robot having rights or interests is absurd, however such a knee-jerk reaction I believe to be ultimately unfounded and speciesist. Certainly the burden of proof would be on such an opponent to offer a principled explanation as to why human biology in particular should be the metaphysically privileged substrate to realise agency and hence, rights and interests.
29 While Purves *et al*. do not explicitly argue that AWS generate a 'responsibility gap', the fact that they conceive of the decisions of the AWS as being metaphysically distinct from that of the programmers and implementers nonetheless tacitly assumes such a gap.
30 Sparrow says that deploying AWS means 'treating our enemy as vermin'. Paradoxically, his claim might be that we would be blameworthy for deploying AWS but not subsequently blameworthy for what they did thereafter. While this might be an entailment of his view, I find this to be a very difficult bullet for him to bite.
31 I would like to especially thank Bradley Strawser, Ryan Jenkins, Susanne Burri, Andrew Parisi, Heather Roff, Jimmy Goodrich, Joseph Bowen, Lars Christie, David Rodin, Adil Haque, Helen Frowe, Holly Lawford-Smith, and Bill Lycan for their help and contributions to the development of this article.