Journal of Applied Philosophy, Vol. 24, No. 1, 2007

Killer Robots

ROBERT SPARROW

ABSTRACT *The United States Army's Future Combat Systems Project, which aims to manufacture a 'robot army' to be ready for deployment by 2012, is only the latest and most dramatic example of military interest in the use of artificially intelligent systems in modern warfare. This paper considers the ethics of the decision to send artificially intelligent robots into war, by asking who we should hold responsible when an autonomous weapon system is involved in an atrocity of the sort that would normally be described as a war crime. A number of possible loci of responsibility for robot war crimes are canvassed: the persons who designed or programmed the system, the commanding officer who ordered its use, the machine itself. I argue that in fact none of these are ultimately satisfactory. Yet it is a necessary condition for fighting a just war, under the principle of* jus in bellum, *that someone can be justly held responsible for deaths that occur in the course of the war. As this condition cannot be met in relation to deaths caused by an autonomous weapon system it would therefore be unethical to deploy such systems in warfare.*

Introduction

The United States Army's Future Combat Systems Project, which aims to manufacture a 'robot army', to be ready for deployment by 2012, is only the latest and most dramatic example of the military's interest in the use of artificially intelligent systems in modern warfare. A number of 'autonomous weapon systems' are already in use in armed forces around the world, including cruise missiles, torpedoes, submersibles, robots for urban reconnaissance, Uninhabited Aerial Vehicles and Uninhabited Combat Aerial Vehicles; more are in production or on the drawing board. The military remains one of the largest sources of funding for research into robotics and artificial intelligence technology. As these technologies improve it seems inevitable that they will be used more and more in warfare. This paper considers the ethics of a decision to send artificially intelligent robots into war, by asking who we should hold responsible when an autonomous weapon system is involved in an atrocity of the sort that would normally be described as a war crime.

The Real World of Robots at War

The idea of killer robots may seem somewhat farfetched — something out of science fiction. In fact research into military robotics and military applications of artificial intelligence is already well advanced.[1] I shall devote the next few paragraphs to describing autonomous weapon systems (AWS) that are already in production and that those

that are, according to reputable sources, likely to be deployed in the next decade, in order to show that the issues discussed here are more pressing than is generally thought.

Existing autonomous weapons systems include cruise missiles, torpedoes, submersibles, robots for urban reconnaissance, Uninhabited Aerial Vehicles (UAVs) and Uninhabited Combat Aerial Vehicles (UCAVs).

Cruise missiles and torpedoes have for a long time exercised a (very limited) degree of autonomy in determining their approach to their target and this is steadily increasing. The Boeing Corporation's SLAM-ER cruise missile, for instance, now has an Automatic Target Recognition capability that allows it to choose a target once it reaches its area of operations, whereas the Northrop Grumman BAT, which homes in on enemy tanks using their acoustic signature, is described by one source as 'a fully autonomous weapons system that can locate, identify, attack and destroy armoured vehicles'.[2] Of particular note in this context is the US Air Force's Low Cost Autonomous Attack System (LOCAAS). LOCAAS is a turbine powered 'stand off' munition, which is designed to 'autonomously search for, detect, identify, attack and destroy theatre missile defence, surface to air missile systems, and interdiction/armour targets of military interest'.[3] It will be equipped with a Laser Radar system and an Autonomous Target Recognition capability that will allow it to search for and identify targets within a 33 sq. mile area. LOCAAS will also carry a warhead that is capable of three different configurations for use against different types of targets, with the machine itself choosing which to employ in its attack.[4] The perceived success of 'smart weapons' in military conflicts in Iraq, Kosovo and Afghanistan means that they are likely to play a large role in any future wars.[5] Improvements in computing power, component miniaturization, and software engineering mean that they are also likely to keep getting 'smarter'.

Robot surveillance drones/UAVs, reconnaissance robots, and submersibles demonstrate more autonomy, as they must be capable of a wider variety of tasks in more challenging environments. The US Navy is developing an Unmanned Underwater Vehicle (UUV), called the Long-Term Mine Reconnaissance System, which can be launched from a submarine whereupon it can seek out and map enemy beach defences and minefields. The Navy is also researching a much more ambitious UUV called MANTA, which will be capable of fully autonomous operation in order to seek out, attack and destroy enemy submarines.[6]

The majority of military robots in existence today however are UAVs. These high tech descendants of Remotely Piloted Vehicles differ from RPVs in having the ability to take off, fly to their objective, and return, with minimal, or even no, human control. UAVs play an increasingly important role in modern air forces. For example, the uninhabited Global Hawk spy plane is now a vital part of the US Air Forces global surveillance capacity. Unpiloted aircraft were used extensively in NATO military operations in Kosovo, both for purposes of surveillance and target designation. By its own reckoning the US Department of Defense has invested over $3 billion (US) in UAVs in the decade 1990–2000 and will spend some $4 billion in the coming decade.[7]

The launching of air-to-ground missiles from armed 'Predator' robot drones in the course of the US invasion of Afghanistan, received extensive publicity. Although a modified version of a pilotless surveillance plane, the Predator may now be classed as a Uninhabited Combat Aerial Vehicle. UCAVs are designed to attack enemies on the ground and in the air, and are widely predicted to be the future of air power.[8] The use

of Predator drones, as well as other UAVs, has become a routine part of combat operations in the current conflict in Iraq.[9]

A recent survey of UCAVs cites four as already in production and another seven under development, the most impressive of which is Boeing's X-45, a futuristic unmanned fighter aircraft.[10] Military interest in UCAVs stems from two main sources. Firstly, uninhabited systems offer the prospect of achieving military objectives without risking the politically unacceptable cost of friendly casualties. Secondly, they are expected to be substantially cheaper than the piloted systems they are intended to replace.[11]

Land warfare involves a more challenging terrain and environment for autonomous systems and so the military has been slower to adopt robots designed for ground combat. Nevertheless, the US Army employs a number of reconnaissance robots for use in urban combat (as does the Israeli army) and is developing more.[12] The US army has recently announced its intention to deploy an armed version of a bomb disposal robot, designed to serve as a remotely operated weapons platform, in Iraq.[13] Both UAVs and UCAVs have an obvious role in ground support; the US Marine Corps Warfighting Laboratory is testing a backpack portable UAV called Dragon Eye and a larger, more capable, rotor winged UAV called Dragon Warrior, for use by their infantry.[14]

However, the most ambitious envisioned deployment of AWS is actually intended for land warfare — the United States Army's Future Combat Systems (FCS) project. Developed in collaboration with the Defence Advanced Research Projects Agency, the FCS program has as its main objective the replacement of the US Army's main battle tank with a system (or 'system of systems') that will be capable of rapid deployment to any location in the world in the hold of a C130 transport plane and of integrating the most advanced battlefield technologies currently available, as well as those likely to become available in the next two decades. It is envisioned that many, perhaps all, components of this system will be capable of unmanned operation. Both manned and unmanned components are expected to operate in conjunction with battlefield surveillance provided by robot reconnaissance units, both in the air and on the ground, and fire support provided by robot aircraft and robot artillery platforms.[15] The emphasis on the role of robotics in the FCS concept is so heavy that media descriptions of the project as involving the creation of a 'robot army' do not seem unreasonable.[16]

Most existing versions of these technologies still require some level of human supervision.[17] However the need for such supervision is steadily decreasing as computing and sensor technology improves. It appears increasingly likely that robots will eventually be entrusted with decisions about target identification and destruction.[18]

Beyond these weapons, which are already on the drawing board, lies the prospect of weapons systems that possess genuine 'artificial intelligence'. According to a number of writers in the field, before the end of the century — and according to some, well before this — machines will be conscious, intelligent, entities with capacities exceeding our own.[19] Given that the military constitute a major source of funding for research into artificial intelligence, it seems inevitable that such AIs will be put to work in military contexts.

How 'Autonomous' are Autonomous Weapon Systems?

It is one thing to point out that these weapons exist, and that more are being developed, it is another to claim that they raise any new ethical issues. The idea that they

might derives from taking the claim that at some point in the future there might be *autonomous* weapons systems seriously.

Sometimes the claim that a weapon is 'autonomous' means only that it is capable of acting independently of immediate human control; typically it means that it is a 'fire and forget' system capable of determining its trajectory or pursuing its target to some limited extent. Many existing autonomous weapon systems are of this type. Weapons of this nature do not in themselves raise any ethical questions beyond those raised by other modern long range weapons.

However, much more than that is being claimed for the next generation of intelligent robots. According to a number of writers the robots of the future will be capable of acting on their own, in some more robust sense.[20] Artificially intelligent weapon systems will thus be capable of making their own decisions, for instance, about their target, or their approach to their target, and of doing so in an 'intelligent' fashion.[21] While they will be programmed to make decisions according to certain rules, in important circumstances their actions will not be predictable. However this is not to say that they will be random either. Mere randomness provides no support for a claim to autonomy. Instead the actions of these machines will be based on reasons, but these reasons will be responsive to the internal states — 'desires', 'beliefs' and 'values' — of the system itself. Moreover, these systems will have significant capacity to form and revise these beliefs themselves. They will even have the ability to learn from experience.[22] In practice, this is likely to mean that the actions of these machines will quickly become somewhat unpredictable. As we shall see below, this is in itself enough to raise difficult ethical questions about the use of these weapons.

How far the autonomy of these systems extends beyond this is not clear. In particular, it is unclear at what point a weapon system's autonomy means that *it* is, in some real sense, making the decisions with which it is entrusted. An important reason for the lack of clarity about this question is that the nature and extent of autonomy, even in people, is itself a controversial and poorly understood matter.

A full account of what autonomy consists in, and whether or not machines could possess it, would require answering a set of questions about the nature and limits of the will, which have puzzled philosophers for centuries. This is a task well beyond the scope of the current paper. Instead, I shall have to be content with two observations.

Firstly, a large number of influential writers have argued that future artificial intelligences will possess capacities equal to, and even exceeding those of human beings.[23] Should this occur then these machines will presumably have a strong claim to be autonomous as well.[24] At least some serious writers in the area are therefore committed to the claim that the dilemmas that I study here will arise.[25]

Secondly, that autonomy and moral responsibility go hand in hand. To say of an agent that they are autonomous is to say 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. Their ends result from the exercise of their capacity to reason on the basis of their own past experience. In both of these things, they are to be contrasted with an agent whose actions are determined, either by their own nature, or by the ends of others. Where an agent acts autonomously, then, it is not possible to hold anyone else responsible for its actions. In so far as the agent's actions were its own and stemmed from its own ends, others can not be held responsible for them.[26] Conversely, if we hold anyone else responsible for the actions

of an agent, we must hold that, in relation to those acts at least, they were not autonomous.

For the moment, I want to remain agnostic on the question of the extent to which existing or future AWS can truly be said to be autonomous. Ultimately, I will argue that even 'strong AI's' of the sort discussed by Kurzweil, Brooks, Moravec, and others, are likely to inhabit a 'grey area' in the middle of the range between systems which are determined and those which are full moral agents, precisely because it is difficult to see how it would be possible to hold them responsible for their actions. However, what I want to do here is to take seriously for the moment the possibility that they might exercise a substantial degree of autonomy and see what follows from that. I shall argue that the more these machines are held to be autonomous the less it seems that those who program or design them, or those who order them into action, should be held responsible for their actions. This itself is enough to suggest that the prospects for fully autonomous machines are more remote than is sometimes claimed.

Robot Warriors and Robot War Crimes

The prospect of the deployment of AWS raises many disturbing ethical questions. How will the use of robot weapons affect the ways in which wars are fought, the level and nature of casualties, and the threshold of conflict? What sort of decisions should they be allowed to make? How should they be programmed to make them? Should we grant non-human intelligent agents control of powerful weapons at all? If a system is intelligent enough to be trusted with substantial decisions making responsibility in battle, should it also be granted moral standing? Should they then be granted rights under the Geneva Conventions? These are just some of the difficult questions surrounding the ethics of the use of AWS.[27]

The question I am going to consider here is who should be held responsible if an AWS was involved in a wartime atrocity of the sort that would normally be described as a war crime. The reason I shall concentrate on this question is that it is arguably prior to the others suggested above. If, as I shall argue below, it turns out that no-one can *justly* be held responsible for the actions of these systems, then it will be unethical to use them in war. The other questions then need not arise.[28]

Let us imagine that an airborne AWS, directed by a sophisticated artificial intelligence, deliberately bombs a column of enemy soldiers who have clearly indicated their desire to surrender. These soldiers have laid down their weapons and pose no immediate threat to friendly forces or non-combatants. Let us also stipulate that this bombing was not a mistake; there was no targeting error, no confusion in the machine's orders, etc. It was a decision taken by the AWS with full knowledge of the situation and the likely consequences. Indeed, let us include in the description of the case, that the AWS had reasons for what it did; perhaps it killed them because it calculated that the military costs of watching over them and keeping them prisoner were too high, perhaps to strike fear into the hearts of onlooking combatants, perhaps to test its weapon systems, or because the robot was seeking to revenge the 'deaths' of robot comrades recently destroyed in battle.[29] However whatever the reasons, they were not the sort to morally justify the action. Had a human being committed the act, they would immediately be charged with a war crime.

Who should we try for a war crime in such a case? The robot itself? The person(s) who programmed it? The officer who ordered its use? No one at all? As we shall see below, there are profound difficulties with each of these answers.

Responsibility and Jus in Bello

The question of the attribution of responsibility in this situation matters because I take it that it is a fundamental condition of fighting a just war that someone may be held responsible for the deaths of enemies killed in the course of it. In particular, someone must be able to be held responsible for civilian deaths. The responsibility at issue here is moral and legal responsibility and not mere causal responsibility.

This condition may be thought of as one the requirements of *jus in bello*: it may also be thought of as a precondition of applying this idea at all.

It is a minimal expression of respect due to our enemy — if war is going to be governed by morality at all — that someone should accept responsibility, or be capable of being held responsible, for the decision to take their life. If we fail in this, we treat our enemy like vermin, as though they may be exterminated without moral regard at all. The least we owe our enemies is allowing that their lives are of sufficient worth that someone should accept responsibility for their deaths.[30] Similarly, their grieving relatives are entitled to an answer as to why they died, which includes both knowing who is responsible and what their reasons were. Ensuring that someone can be held responsible for each death caused in war is therefore an important requirement of *jus in bello*.

The assumption and/or allocation of responsibility is also vital in order for the principles of *jus in bello* to take hold at all. The principle of discrimination, for instance, which requires that combatants distinguish between legitimate and illegitimate targets, assumes that we can specify who is responsible for attacks that may violate it. More generally, application of the principles of *jus in bello* requires that we can identify the persons responsible for the actions that these principles are intended to govern.

Whichever way we understand it, the principle that we must be able to identify those responsible for deaths in war receives support from strong consequentialist and deontological arguments. As I set it out above, it is a necessary condition of the respect for persons that is at the heart of Kantian, and other deontological, ethics. However, it is also clearly supported by weighty consequentialist considerations. An inability to identify those responsible for war crimes would render their prosecution moot, for instance, with disastrous consequences for the ways in which wars are likely to be fought. Indeed the very same consequentialist reasons which motivate a concern for *jus in bello* in the first place ground our interest in the question of responsibility.

Sometimes, of course, there will be individual circumstances in which it is exceedingly difficult to determine who should be held responsible for certain deaths, or perhaps even where no-one could justly be held responsible. Accidents do happen, and even more so in war. However, these accidents represent regrettable, if inevitable, failures to live up to principles of justice in war fighting. If the nature of a weapon, or other means of war fighting, is such that it is *typically* impossible to identify or hold individuals responsible for the casualties that it causes then it is contrary to this important requirement of *jus in bello*. It will not be ethical to employ this means in war.

Part of what is immoral about weapons of mass destruction, or other means of indiscriminate slaughter, such as anti-personnel mines, for instance, is that they violate

this condition. It is traditional to characterise the problem with these weapons as a failure to distinguish between legitimate and illegitimate targets, typically, soldiers and civilians.[31] However, another way of thinking about these weapons is that when they are used, no one is taking responsibility for the decision about who does and does not get killed.[32] It is mere accident that *these* particular civilian deaths occurred. The use of these weapons implies that this is not important and thus demonstrates a profound disrespect for the value of an individual human life.[33]

If it turns out that no one can properly be held responsible in the scenario described above, then AWS will violate this important condition of *jus in bello*.

Will Insisting on 'Human Oversight' Avoid the Problem?

The thought that a machine might be trusted to make the decision to take a human life is obviously a disturbing one. In order to avoid the ethical dilemmas this might pose, or the public outcry it is likely to provoke, we might wish to ensure that any decision to open fire, or to take action that could threaten human life, be considered and approved by a human operator.[34] When I contacted the Program Manager of the US FCS project in the course of researching this paper, he was very quick to insist that this would be the case.[35]

Requiring that human operators approve any decision to use lethal force will avoid the dilemmas described here in the short-to-medium term. However, it seems likely that even this decision will eventually be given over to machines. There is an obvious tension involved in holding that there are good military reasons for developing autonomous weapon systems but then not allowing them to fully exercise their 'autonomy'. The same pressures that are pushing for the deployment of military robots in the first place also push for them to be given control over which targets to attack and when to open fire.

Indeed, as AI technology improves, a human operator may prove not merely redundant but positively disadvantageous in such systems. 'Improvements' in the technology of war fighting mean that the tempo of battle is continually increasing. The development of long range and high speed projectiles, with ever improving guidance and target acquisition systems means that in some combat arenas, especially air combat, the window of opportunity in which to take evasive or effective counter action once hostile contact is made is contracting rapidly. It seems likely that sometime in the not-too-distant future, the time available to make survival critical decisions will often be less than the time required for a human being to make them.[36] When this occurs, then *only* robots will be capable of participating effectively in these forms of combat. Many military commentators have observed that the generation-after-next of fighter aircraft are likely to be uninhabited for precisely this reason. Furthermore, the communications infrastructure required in order to maintain human oversight is an obvious weak point in the operations of these systems.[37] The links between the weapon system and its human overseer may be threatened by electronic counter-measures by hostile forces, by environmental factors (such as difficulties in maintaining line-of-sight necessary for communications, atmospheric interference, etc) or other exigencies of the 'fog of war'. The communications infrastructure itself is likely to be an early target for enemy attack, with the prospect that success in such attacks could disable all robot forces. The survivability and range of operations of an AWS will therefore be greatly increased if it is capable of operating without human supervision.

Both the tendency of the tempo of battle to increase with technological developments and the costs associated with keeping a human 'in the loop' are likely to be greatly exacerbated as soon as autonomous weapons systems are deployed. Weapons that require human oversight are likely to be at a substantial disadvantage in combat with systems that can do without. Thus as soon as one nation is capable of deploying AWS that can operate without human oversight then all nations will have a powerful incentive to do so.

For all these reasons, there is likely to be strong pressure in the future to allow AWS to operate in a 'fully autonomous' mode — including the ability to make decisions about choice of targets and payload deployment.

In any case, a version of the problem that concerns me here arises even when human beings have the last word on a decision, as long as human combatants are relying crucially on other decisions made by AI's. For instance, the pilots of modern fighter aircraft must make sense of a tremendous amount of rapidly evolving information from multiple sources, very quickly, in situations where reaching the correct decision is a life or death matter. In order to make this task easier, research is under way into systems in which a computer, using artificial intelligence technology, analyses and interprets the incoming data from various radar and other sensing systems and then re-presents it to the pilot in the form of moving icons indicating the probable nature of various objects (including 'friend or foe' identification), the level of threat they represent and their likely future trajectory. The pilot can then act on the basis of this information.[38] Where such a system is in use, the fighter pilot relies crucially on targeting information provided by an AI in making the decision to destroy a target. If this information turns out to be wrong — perhaps even 'deliberately' misleading — can we still hold the pilot responsible for the consequences of their decision? The question of who is responsible for the decisions on the basis of which the human decision is made, remains crucial.[39]

Responsibility for Robot War Crimes

Who should we hold responsible for a war crime in a situation that crucially involves a decision made by an AWS, such as that described above?

The Programmer?

Given that the weapon is presumably not supposed to behave in this way, it is tempting to insist that the fault lies with the person(s) who designed and/or programmed the weapon, and that they should be held responsible for its destructive result.[40] However, this will only be fair if the situation described occurred as a result of negligence on the part of the design/programming team.

This need not be the case for two reasons.

Firstly, the possibility that the machine may attack the wrong targets may be an acknowledged limitation of the system. If the manufacturers have made this clear to those who purchase or deploy the system, then it seems they can no longer be held responsible, should this occur. The responsibility is assumed by those who decide to send the weapon into battle regardless.

Secondly, and more importantly, the possibility that an autonomous system will make choices other than those predicted and encouraged by its programmers is inherent in the claim that it is autonomous. If it has sufficient autonomy that it learns from its experience and surroundings then it may make decisions which reflect these as much, or more than, its initial programming. The more the system is autonomous then the more it has the capacity to make choices other than those predicted or encouraged by its programmers. At some point then, it will no longer be possible hold the programmers/designers responsible for outcomes that they could neither control nor predict. The connection between the programmers/designers and the results of the system, which would ground the attribution of responsibility, is broken by the autonomy of the system. To hold the programmers responsible for the actions of their creation, once it is autonomous, would be analogous to holding parents responsible for the actions of their children once they have left their care.

The Commanding Officer?

The argument above suggests that the officer who ordered the deployment of the weapons system should instead be held responsible for the consequences of its use. The risk that it may go awry is accepted when the decision is made to send it into action. This is the preferred approach of the military forces seeking to deploy existing AWS.[41] It accords with the precedent set by our response to cases where other weapons that may kill people other than their intended target are used. In these cases we simply insist that those who use them should be held responsible for the deaths they cause, even where these were not intended.

If the autonomy of the weapon *merely* consists in the fact that its actions cannot always be reliably predicted and therefore that it may sometimes kill people whose deaths were not intended, then the analogy with existing weapons may be close enough. Employing AWS, then, is like using long-range artillery. The risk that shells may land off target is accepted when the decision to fire is made. If they do kill people other than their intended targets, responsibility for the decision to fire remains with the commanding officer.

However, this is a peculiar way to treat what are advertised as 'smart' weapons. It implies that there is no fundamental moral difference between them and more ordinary 'dumb' weapons. This way of resolving the problem therefore sits uneasily with the original claims about the 'autonomy' of such systems. What distinguishes AWS from existing weapons is that they have the capacity to choose their own targets. If we understand the autonomy they exercise in doing so only as a limit on our ability to predict how they will behave, then on the face of it this implies that the more autonomous they become the less confidence we can have that they will attack the targets that we intend. However, we normally think that smart bombs, and other AWS, are *more* reliable ways to attack the enemy. It is hoped that future AWS will even be capable of discriminating reliably between civilian and military targets. This has even led a number of critics to argue that the use of these weapons is morally *superior* to the use of ordinary 'dumb' weapons.[42]

So, the autonomy of the systems cannot be captured by the mere fact that they are unpredictable. Yet these weapons are more than just guided weapons, which attack targets that have been chosen for them. The more autonomous these weapons are, the

more it is possible that they might attack the wrong target. That is, the more it is true that they *could* do so. The precise nature and significance of this 'could' must remain somewhat mysterious. Again, the nature of free will is too large a topic for me to treat here. At this stage of my discussion it will have to suffice to note that the autonomy of the machine implies that its orders do not determine (although they obviously influence) 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. The more autonomous the systems are, the larger this risk looms. At some point, then, it will no longer be fair to hold the Commanding Officer responsible for the actions of the machine. If the machines are really choosing their own targets then we cannot hold the Commanding Officer responsible for the deaths that ensue.

The Machine?

The final possible loci of responsibility, then, is the machine itself. Perhaps we should try the machine for war crimes?

It is hard to take seriously the idea that a machine should — or could — be held responsible for the consequences of 'its' actions.[43] We can easily imagine a robot, or for that matter, any machine, being *causally* responsible for some death(s). A number of authors have even recently argued that machines may be properly be thought of *as* acting; that they are 'artificial agents'.[44] However, we typically baulk at the idea that they could be *morally* responsible.[45]

Why should it be so hard to imagine holding a machine responsible for its actions? One reason is that it is hard to imagine *how* we would hold a machine responsible — or, to put it another way, what would follow from holding it to be responsible. 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.[46] Yet how would we go about punishing or rewarding a machine?

For some people, the intuitive implausibility of punishing or rewarding a machine will be enough to establish that the AWS could never be held morally responsible for its actions. However, I want to investigate the matter a little bit further in order to show that what would need to be true in order for us to be able to hold a machine responsible is even more demanding than might at first sight appear. I shall concentrate on the question of punishment as the issues are starker here; but the arguments explored below will also apply to the question of the possibility of rewarding a machine.

Some people may not, in fact, share the intuition that we could not punish a machine, especially when it comes to robots that are sufficiently intelligent to have a good claim to be described as autonomous. It seems likely that any robot that is capable of 'intelligent' behaviour will have internal states which function like desires as well as an internal structure which motivates them to pursue these.[47] These cognitive states will be necessary in order to make it possible that the machine can be rewarded, or experience reward, when it achieves its (or perhaps its designer's) goals. Frustration of these desires might therefore function as a mechanism whereby the machine could be punished.

If we imagine an AI with intellectual capacities similar, or even equal to those of human beings, then it seems even more plausible that they could be punished. Such machines will presumably have goals and desires beyond those presupposed by their military role. If an AWS earned wages for its services to the nation, which it could spend on its own projects, then we could simply dock these for minor misdemeanours. For more serious crimes, we could imprison the culprit. One could, for instance, restrict a machine's liberty by restraining it physically or by imposing a restraint through its programming. Alternatively, we might administer corporeal punishment by damaging the machine in some way, or perhaps by administering electric shocks to those electrodes through which it senses damage in combat. Finally, we might institute 'capital' punishment for the most serious crimes, such as war crimes, and destroy the machines responsible for them.

However, although these courses of action might appear to satisfy our psychological need for revenge, it is not yet clear that they would count as punishment. In order for any of these acts to serve as punishment they must evoke the right sort of response in their object. The precise nature of this response will depend on our theory of punishment. I shall assume that the most plausible accounts of the nature and justification of punishment require that those who are punished, or contemplate punishment, should suffer as a result.[48] While we can imagine doing the things described above to a machine, it is hard to imagine it suffering as a result.

In order for a machine to be capable of being punished then, it must be possible for it to be said to suffer. Furthermore, its suffering must be of the sort that we find morally compelling. Talk of machine suffering which is most naturally expressed in inverted commas — 'suffering' — will not suffice here. If a grieving relative of one of the machine's victims questions whether the machine has been punished sufficiently, it will not do to point out that it is 'suffering' because there is friction in its gears as a result of not being oiled, or that it hasn't been able to log on to the web to play chess in its spare time. In order for our treatment of the machine to count as punishment, it must be capable of suffering in ways that might motivate the same set of responses that we have as a matter of course to human beings. It must be such that we could understand someone saying that they felt sympathy for it, or grief, or remorse, if this suffering turned out to be unnecessary.[49] Indeed, the suffering involved in a machine being punished must be such that if we discover that it was in fact innocent (perhaps the bomb that killed the soldiers was launched by another AWS near by) then we feel that we have done it a serious wrong and owe it recompense.

What must be true in order for a machine to be held morally responsible is therefore much more demanding than at first appeared. Not only must machines have internal states of equivalent complexity and with analogous structure to those of human beings, but they must also have the capacity to express these in ways that will establish the moral reality of these states. They must make the same sorts of moral and empathic demands upon us as do other (human) people.

Again it might be thought that robots of the sort predicted by Brooks, Kurzweil, and Moravec would possess sufficient internal complexity and capacity for expressiveness for this to be the case.[50] That is, if we were to see them 'suffering' we would no longer be inclined to feel that we need the inverted commas — we would believe they really *were* suffering. Similarly, we would instinctively reach out to give them assistance, feel concern for their wellbeing, and grieve for their deaths. If, as a result, we felt of such

machines that they were appropriate sites of punishment, then they could be held responsible for their actions; but equally well we could be held responsible for our actions in relation to them. They would be full 'moral persons'. Paradoxically, the creation of such machines would not achieve the goal that motivated their development in the first place — that of allowing wars to be fought without risking our soldiers being killed. Our machines would have become our soldiers and we should be as morally concerned when our machines are destroyed — indeed killed — as we are when human soldiers die in war.

I will not attempt to settle here whether or not machines will ever be able to meet these conditions — although it is my belief that they will not.[51] It will serve here to point out the enormous gap between any system that might meet these conditions and any AWS currently in existence, or even on the drawing board. Even a machine that was capable of making exceedingly complex battlefield decisions, in a fraction of a second, on the basis of a knowledge of tactical and technical expertise way beyond that possessed by any human being would not meet these conditions. While some hypothetical AI that can establish a claim to moral personhood might be responsible for what it does, for the foreseeable future we will not be able to hold that machines are responsible for their actions.

We have reached an impasse. I have argued that as machines become more autonomous a point will be reached where those who order their deployment can no longer be properly held responsible for their actions. I have also suggested that machines could have capacities way beyond those necessary to reach this point without it being possible to hold them responsible. How can both these things be true — and how can we proceed from here?

Robot Warriors and Child Soldiers

We can better conceptualize this dilemma if we consider another case where attribution of responsibility in wartime is problematic; the use of child soldiers in war.[52] One of the things that is unethical about employing youth below a certain age in combat duties is that life and death decisions are placed in the hands of children who cannot be held responsible for their actions.[53] While they lack full moral autonomy — and therefore are not morally responsible for what they do — there is clearly *a* sense in which children are autonomous. They are capable of a wide range of decisions and actions. They are certainly much more autonomous than any existing robot. Yet they are not appropriate objects of punishment, as they are not capable of understanding the full moral dimensions of what they do — and therefore of understanding the connection between their punishment and their crime.

The limited autonomy that children do possess is enough, moreover, to ensure that those who order them into action do not control them. The fact that their commanders do not control them problematises the attribution of responsibility for their actions. Note that what makes the attribution of responsibility especially problematic here is not that child soldiers are necessarily unreliable or unpredictable. For instance, it might be a more reliable way to kill a particular enemy to send a child assassin to seek him or her out, than to drop a bomb from great height. Yet while they will *probably* kill the right person, they might not. This possibility is inherent in their capacity for autonomous action.

It is the prospect of intelligent actors without any moral responsibility that makes child armies especially terrifying. When child armies take to the battlefield, as they have in Angola and Liberia in recent years, no one is in control. If civilians are killed they are killed senselessly without anyone being responsible for their deaths. The deaths that occur, occur, in a sense, indiscriminately — without necessarily being random.

There seems to be a conceptual space in which children and (perhaps) machines are sufficiently autonomous to make the attribution of responsibility to an appropriate adult problematic, but not so autonomous as to be responsible for their own actions. This space is bounded at the lower end by entities that have no, or at least little, autonomy. If they play a role in generating an outcome, it can only be a causal and not a moral one; if anyone is responsible for that outcome, it is the person who placed them in the position where they played that causal role. This space is bounded at the upper end by entities that are fully morally autonomous, are responsible for their own actions, are appropriate sites of punishment and also therefore make moral claims upon us. In between is a region, fuzzy at both the upper and lower ends, in which entities are sufficiently complex, and possess internal states that function as ends, such that their actions can no longer be attributed to those who set them in motion, but where they are not sufficiently well-formed moral agents to be fully morally responsible.

In practice, I think we try to close this space by stipulating that all entities should be treated as though they fit beneath the lower or above the upper bound. However, as the case of children shows, this is not always a convincing solution.

The problem that I have identified with robot weapons is that they seem likely to occupy this uneasy space for the foreseeable future. I have argued that it is easy to imagine autonomous machines entering this space from their current place well below its bottom limit. But is difficult to imagine them emerging from the upper limit of this space.

While they remain in this ambiguous space, the attribution of responsibility for their actions is deeply problematic. The only possible solution seems to be to *assign* responsibility to an appropriate individual — presumably the commanding officer who orders their use. However, as I have argued above, this solution holds the commanding officer responsible for things that they could not control, and therefore risks that they will be punished unfairly. The only way of meeting our obligation to enemy combatants to ensure that someone can be held responsible if they are killed unjustly, risks a grave injustice to our own military personnel who wield authority on the battlefield.

Conclusion

I have argued that it will be unethical to deploy autonomous systems involving sophisticated artificial intelligences in warfare unless someone can be held responsible for the decisions they make where these might threaten human life. While existing autonomous weapons systems remain analogous to other long-range weapons that may go awry, the more autonomous these systems become, the less it will be possible to properly hold those who designed them or ordered their use responsible for their actions. Yet the impossibility of punishing the machine means that we cannot hold the machine responsible. We can insist that the officer who orders their use be held responsible for their actions, but only at the cost of allowing that they should sometimes be held entirely responsible for actions over which they had no control. For the

foreseeable future then, the deployment of weapon systems controlled by artificial intelligences in warfare is therefore unfair either to potential casualties in the theatre of war, or to the officer who will be held responsible for their use.

Robert Sparrow, School of Philosophy and Bioethics, Faculty of Arts, Monash University, Victoria 3800, Australia. Robert.Sparrow@arts.monash.edu.au

NOTES

1 Indeed, the military is one of the major sources of funding for research into robotics and artificial intelligence. In a recent survey of cutting edge research in robotics, there is scarcely a laboratory in Europe or the US that is not receiving funding from the military. See P. Menzel, and F. D'Aluisio, *Robo Sapiens: Evolution of a New Species* (Cambridge, MA: MIT Press, 2000).
2 C. Beal, 'Briefing Autonomous Weapons Systems: Brave New World', *Jane's Defence Weekly*, 33, 6 (2000): 22–26.
3 'Low Cost Autonomous Attack System (LOCAAS)', http://www.fas.org.man/dod-101/sys/smart/locaas.htm, at 14.8.02.
4 T. Barela, 'Anti-Armor Einstein' *Airman Magazine* September (1996); Air Armament Center Public Affairs Report 'This bomb can think before it acts' *Leading Edge Magazine*, 42, 2 (2000): 12.
5 P. Meilinger, 'Precision Aerospace Power, Discrimination, and the Future of War', *Aerospace Power Journal* 15, 3 (2001): 12–20.
6 Information on US Navy UUV programs is available via, 'Autonomous Operations', http://www.onr.navy.mil/auto-ops/introduction.asp, at 5.8.02. See also 'Ocean systems — Surveillance & Reconnaissance Systems', http://www.boeing.com/defense-space/infoelect/lmrs/sld001.htm, at 5.8.02.
7 Office of the Secretary of Defense, United States Government (2001) *Unmanned Aerial Vehicles Roadmap 2000–2025* (Department of Defense, United States Government, Washington D.C. Available at www.acq.osd.mil/usd/uav_roadmap.pdf), p. i.
8 R. Chapman II, 'Unmanned Combat Aerial Vehicles: Dawn of a New Age?' *Aerospace Power Journal* 16, 2 (2002): 60–73.
9 S. Shactman, 'Attack of the Drones' *Wired Magazine* 13, 6 (2005) available at http://www.wired.com/wired/archive//13.06/drones_pr.htmal, at 25.08.05.
10 S. Kainikara, 'UCAVs probable lynchpins of future air warfare', *Asia-Pacific Defence Reporter*, 28, 6 (2002): 42–45.
11 Kainikara op. cit. See also, J. Kerr, 'Boeing looks to the future' *Asia-Pacific Defence Reporter*, May 28, 4 (2002): 24–25. The cost savings offered by UCAVs are subject to some controversy but are expected to flow from the relative expense of pilot training, support and training mission related maintenance over the operational life of manned systems. Office of the Secretary of Defense, United States Government (2001) *Unmanned Aerial Vehicles Roadmap 2000–2025*.
12 M. Behar, 'The New Mobile Infantry' *Wired* May, 10, 5 (2002): 110–114; B. Bender 'USMC to buy robots for urban warfare operations', *Jane's Defence Weekly*, 7 June 33, 23 (2000): 2; D. Lewin, 'Smart cars in peace and war' *IEEE Intelligent Systems* May–June, (2001): 4–8.
13 'US plans "robot troops" for Iraq', BBC News World Edition Website, 23 January (2005), available at http://news.bbc.co.uk/2/hi/americas/4199935.stm, at 25.9.05.
14 'Marine Corps Warfighting Laboratory', http://www.mcwl.quantico.usmc.mil/tech.html, at 2.8.02.
15 Descriptions of the FCS program may be found at; 'TTO Programs — Future Combat System (FCS)', http://www.darpa.mil/tto/programs/fcs.html, at 5.8.02; http://www.darpa.mil/darpatech2000/speeches/TTOspeeches/TTOFutureCombat(Van%20Fosson).doc, at 5.8.02; 'Future Combat Sytems/Future Combat System', http://www.fas.org/man/dod-101/sys/land/fcs.htm, at 5.8.02. See also, J. Kerr, 'Boeing looks to the future' (2002).
16 D. Wastell, 'Robot soldiers herald the era of cheaper, "safer" wars', *The Age*, Melbourne, Australia, March 4, (2002).
17 In particular, existing UAVs and UCAVs rely, for at least some of their operations, on a human operator piloting them remotely via satellite data link.
18 Beal op. cit. Note that LOCAAS is being designed to make these decisions, for instance.

19 R. A. Brooks, *Robot: The Future of Flesh and Machines* (London; Penguin, 2003); G. Dyson, *Darwin Amongst the Machines: The Evolution of Global Intelligence* (Reading, MA: Addison-Wesley Pub. Co., 1997); R. Kurzweil, *The Age of Spiritual Machines: When Computers Exceed Human Intelligence* (Sydney: Allen & Unwin, 1999); H. Moravec, *Robot: Mere Machine to Transcendent Mind* (Oxford: Oxford University Press, 1998).

20 Brooks op. cit.; Dyson op. cit.; Kurzweil op. cit.; Moravec op. cit.

21 If I am guilty of exaggerating the capabilities of these machines, I am at least not alone in this. See for instance, Barela, 'Anti-Armor Einstein' op. cit.; Air Armament Center Public Affairs Report (2000) 'This bomb can think before it acts'.

22 M. De Landa, *War in the Age of Intelligent Machines* (New York: Swerve Editions, 1991), pp. 164–170.

23 Brooks op. cit.; Dyson op. cit.; Kurzweil op. cit.; Moravec op. cit.

24 It might even be argued that possession of autonomy is a *necessary* condition of genuine intelligence. See, M. Midgley, 'Artificial Intelligence and Creativity', in her *Utopias, Dolphins, and Computers: Problems of Philosophical Plumbing* (London and New York: Routledge, 1996) pp. 151–173.

25 R. Brooks 'Designed for Life', *New Scientist*, May 1, (2002).

26 This is of course an oversimplification. Others may be responsible for the circumstances in which an actor acts, or for assisting them in their actions, or in various other ways can come to share some responsibility for the outcome of an actor's actions. Nonetheless the basic point remains that autonomous agents are in some sense the authors of their own actions and that this serves to block the attribution of responsibility for those actions to others.

27 Beal op. cit.; A. Baird, 'The Ethics of Autonomous Weapons Systems', paper presented to the Royal College of Defence Studies, London (2000); A. Lazarski, 'Legal Implications of the Uninhabited Combat Aerial Vehicle' *Aerospace Power Journal* 16, 2 (2002): 74–83.

28 Though to be sure, if nations do deploy AWS, despite the fact that it is unethical to do so, the last two of these questions will remain important.

29 Note that these reasons would not necessarily be internally represented in the robot in this form. They might instead take the form of the outcome of a complex calculation of, or algorithm based upon, the relative weights of the battlefield factors the robot was programmed to consider. The account I provide here is intended to reflect the way an observer would interpret its reasons, using the "intentional stance". See D. C. Dennett, *The Intentional Stance* (Cambridge, MA: MIT Press, 1987). Nor is it that far fetched to imagine that AWS will be programmed with objectives that might lead to such a result (for example, to preserve friendly military capacity, attack enemy morale, gather information about the effectiveness of new ordnance, take account of friendly casualties, etc.).

30 T. Nagel, 'War and massacre', *Philosophy and Public Affairs* 1 (1972): 123–44 at pp. 133–142.

31 N. Fotion, *Military Ethics* (Stanford: Hoover Institution Press, 1990), Chapter 3.

32 We do not usually think of the problem with weapons of mass destruction this way, because we are quick to *assign* responsibility for all these deaths. We hold the person who used or ordered the use of the weapon responsible for all the deaths it causes. It is natural to extend their responsibility because it is reasonable to expect them to be aware of the likelihood that non-combatants would be killed. Furthermore, no alternative locus of responsibility is plausible. However, as we shall see below, neither of these things may be true in the case of an AWS.

33 M. Walzer, *Just and Unjust Wars: A Moral Argument with Historical Illustrations* 3rd edn. (New York: Basic Books, 2000), pp. 134–136 & 138–175; T. Nagel op. cit. pp. 133–142.

34 Beal op. cit.; Lazarski op. cit.

35 Personal communication from US Army Colonel William R. Johnson, via email, 15.4.02.

36 Beal op. cit.

37 Kainikara op. cit.

38 S. Mulgund, G. Rinkus, C. Illgen, G. Zacharias, and J. Friskie, 'OLIPSA: On-Line Intelligent Processor for Situation Assessment', Presented at the *Second Annual Symposium and Exhibition on Situational Awareness in the Tactical Air Environment* (Patuxent River, MD, 1997).

39 Perhaps one of the most disturbing applications of AI in war fighting is the likely role of intelligent systems in the US's ballistic missile defence program. Any system capable of defending mainland US from even a limited ballistic missile attack launched from a 'rogue' state, must be capable of detecting and tracking dozens, if not hundreds, of real and dummy warheads, using multiple sources of information, and of co-ordinating their interception and destruction by orbiting satellites or ground based missile systems. The complexity of this task virtually guarantees that AI technology will play a vital role in any missile defence

system. Hopefully, a human finger will be 'on the button' when it comes to any decision to launch a counter strike. However, as long as they are relying on information provided by an AI, the role of the AI remains critically important. For an early discussion of the ethical issues raised by the use of AI in ballistic missile defence, see T. Forrester, *Computer Ethics* (Cambridge, MA: MIT Press, 1990), pp. 173–181.

40 This is the solution adopted by A. Kuflik, 'Computers in Control: Rational transfer of authority or irresponsible abdication of autonomy', *Ethics and Information Technology*, 1, 3 (1999): 173–184.

41 See Lazarski op. cit. This is also the solution advocated in Beal op. cit.

42 Meilinger op. cit. For further sources and discussion, see C. J. Dunlap, Jr, 'Technology: Recomplicating Moral Life for the Nation's Defenders' *Parameters: US Army War College Quarterly* Autumn (1999): 24–53.

43 Discussion of responsibility for events involving computers typically do not even consider the possibility that the machine itself might be responsible. See, for example, Forrester op. cit., p. 122; Kuflik op. cit.; S. I. Edgar, *Morality and Machines* (Sudbury, MA: Jones and Bartlett Publishers, 1997), pp. 368–69.

44 L. L. Floridi and J. W. Sanders, 'Artificial Evil and the Foundation of Computer Ethics' in D. G. Johnson, J. H. Moor and H. Tavani (eds.) *Proceedings for Computer Ethics: Philosophical Enquiry 2000, Dartmouth College, July 14–16* (Hanover, New Hampshire, 2000), pp. 142–156; L. L. Floridi and J. W. Sanders 'On the Morality of Artificial Agents' in R. F. Chadwick, L. Introna, and A. Marturano (eds.), *Proceedings of the Computer Ethics: Philosophical Enquiry 2001 Conference: IT and the Body, Lancaster University, 14–16 December*, (Lancaster, U.K., 2001), pp. 84–107.

45 The most serious defence that I am aware of by a philosopher, of the idea that machines could be morally responsible for their actions is D. Dennett, 'When HAL Kills, Who's to Blame? Computer Ethics' in D. G. Stork (ed.) *HAL's Legacy: 2001's Computer as Dream and Reality* (Cambridge, MA: MIT Press, 1997), pp. 351–365.

46 There are many cases where we acknowledge that we cannot, as a contingent fact, punish the person responsible for some crime. However, for them to be guilty of a crime, it must be the case that it would at least be appropriate to punish them. If we cannot for some reason, even imagine punishing them for what they have done, then it makes no sense to say that they are guilty of a crime, as guilt implies the appropriateness of punishment.

47 Dennett, (1987) op. cit.; R. W. Picard, *Affective Computing* (Cambridge, MA: MIT Press, 1997).

48 The retributivist core at the heart of our notion of punishment, which explains why we must punish always and only those who deserve punishment — as opposed to others whose punishment might have more beneficial deterrent or rehabilitative effects — seems also to require that those punished suffer thereby. If punishment is justified by its deterrent effect then this also suggests a requirement that punishment causes those punished to suffer. If punishment aims at the education and rehabilitation of offenders then it may or may not require that they suffer by being punished, depending on our account of the motivations necessary to achieve these results. See C. L. Ten, 'Crime and Punishment' in Peter Singer (ed.) *A Companion to Ethics* (Oxford: Blackwell Reference, 1991), pp. 366–72.

49 R. Gaita, *A Common Humanity: Thinking About Love & Truth & Justice* (Melbourne: Text Publishing, 1999), esp. pp. 262–274; R. Gaita, *Good and Evil: An Absolute Conception* (London: MacMillan, 1991), Chapter 9, pp. 144–165; R. Gaita, 'The Personal in Ethics' in D. Z. Phillips and P. Winch (eds.). *Wittgenstein: Attention to Particulars* (London: MacMillan, 1989), pp. 124–50.

50 Research on facial expression for the purpose of conveying emotions is being undertaken in a number of laboratories around the world. For accounts of several of these projects, see Menzel, and D'Aluisio op. cit.

51 For argument to this affect see, R. Sparrow, 'The Turing Triage Test' *Ethics and Information Technology* 6, 4 (2004): 203–213.

52 I would like to thank Jessica Wolfendale for discussions that helped me to see the merits of this comparison.

53 Of course, this is far from being the only, or even the main, thing wrong with the participation of children in war. However it is the aspect that I wish to concentrate on here, in the hope that it can shed light on the use of robot weapons.