TWENTY SECONDS TO COMPLY: AUTONOMOUS WEAPON SYSTEMS AND THE RECOGNITION OF SURRENDER.

Associate Professor Robert Sparrow, Department of Philosophy, Monash University

Forthcoming in *International Law Studies* (accepted 22 July 2015)

ABSTRACT

Autonomous weapon systems are widely predicted to be the future of war fighting, at least in the armed forces of highly industrialised nations. In this paper I consider a profound but so far little-studied problem relating to the ethics of the use of AWS. A fundamental requirement of the *jus in bello* principle of distinction in Just War Theory is that combatants should not attack enemy units that have clearly indicated their desire to surrender. This poses a serious challenge to the ethical use of AWS because both the difficulties of forming an accurate model of the world from the sensors available to robots, and the contextual nature of the signals used to indicate surrender, suggest that robots will struggle to recognise surrender for at least some years to come. After discussing the appropriate place at which to set the standard for "reliability" in surrender recognition by robots, I examine a number of possible ways around the problem and argue that none of them are likely to entirely avoid the dilemmas associated with the limited capacities of AWS to recognise surrender. Consequently, I discuss the ethics of the use of AWS that cannot reliably recognise surrender and argue that thinking about two different sorts of cases may usefully clarify the issues involved. Finally, I suggest that thinking about robots and the requirement to respect surrender may productively illuminate the larger debate about the ethics of AWS by shedding light on the vexed questions of whether it is more appropriate to think of these systems as weapons or as platforms and the locus of responsibility for civilian (and other) casualties caused by these systems.

Keywords: ethics; military ethics; robotics; robots; autonomous weapon systems; AWS; surrender; Just War Theory.

TWENTY SECONDS TO COMPLY: AWS AND THE RECOGNITION OF SURRENDER.

INTRODUCTION

The United States' Department of Defense defines autonomous weapons as "weapon systems that, once activated, can select and engage targets without further intervention by a human operator" (United States' Department of Defense 2012). Autonomous weapon systems are widely predicted to be the future of war fighting, at least in the armed forces of highly industrialised nations (Singer 2009). Consequently, there is now a vigorous debate going on about the ethics and policy of both the development and deployment of such weapons.[1]

A key controversy in this debate concerns the likelihood that AWS will be capable of reliably distinguishing between civilian and military objects (Human Rights Watch 2012; Krishnan 2009, pp. 98-99; Schmitt 2013; Sharkey 2012a). This is obviously a crucial question and one that I have discussed elsewhere (Sparrow 2009a; Sparrow 2011). However, for the sake of the current paper, I want to assume that — as seems likely — in at least some domains AWS will have the capacity to do this to a high degree of accuracy. While distinguishing insurgents from the civilian population in urban settings may be beyond the capacity of robots for many years yet, distinguishing between a tank and civilian cars or trucks is well within the capacity of existing systems, while in some domains, such as submarine warfare or missions directed at enemy air defences, there may be no potential targets other than military objects (Brutzman *et al.* 2013; Guarini and Bello 2012; Schmitt 2013).

A closely related but much more complex question is whether AWS are likely to be capable of complying with the *jus in bello* requirements of distinction and proportionality (Human Rights Watch 2012, pp. 3, 24-26, 30-34; Sharkey 2012b, pp. 788-790; Wagner 2011/12). *Jus in bello* is that portion of just war theory that is concerned with the legitimacy of the means used in fighting wars. (Very) roughly speaking, the principle of distinction requires that attacks only be directed at combatants, while proportionality requires that the military advantage aimed at by an attack justifies the foreseeable evils caused by the attack, especially — but not exclusively — any civilian casualties it might cause. As I have also discussed elsewhere, whether robots will ever be capable of making the required proportionality calculations remains highly controversial (Sparrow 2015. See also Sharkey 2012b, pp. 789-790; Wagner 2011/12).

However, in this paper, I wish to focus on a particular aspect of this larger problem, which has to date received almost no discussion in the published literature, being the question of the capacity of AWS to recognise surrender, and its implications for the ethical deployment of AWS.[2] As I outline in section 1, a fundamental requirement of the *jus in bello* principle of

[1] The literature on this topic is now too large to attempt to cite here in any detail. However, for some representative examples see: Anderson and Waxman 2013; Arkin 2010; Asaro 2012; Borenstein 2008; Krishnan 2009; Marchant, Allenby, Arkin *et al*. 2011; Schmitt and Thurnher 2013; Sharkey 2008; Singer 2009; Sparrow 2007 & 2015.

[2] Importantly, my concern here is solely with the *ethics* of the use of AWS: I leave the question of the legality of their use in various contexts to those better qualified to answer it. In conversation, George Lucas has suggested to me that the question of whether robots might be able to recognise surrender and, if so, how, was in fact one of

distinction in Just War Theory is that combatants should not attack enemy units that have clearly indicated their desire to surrender. By ceasing to participate in hostilities and signalling surrender, military units can acquire the moral status of non-combatants, such that deliberate attacks on them are no longer permissible. In the second section of the paper, I argue that even if robots can distinguish between military and civilian objects they may struggle to recognise surrender. [3] The perceptual task of recognising the actions that signal surrender is likely to be significantly harder than the task of identifying military objects, while the contextual nature of the signals used to indicate surrender implies that robots will need to be able to interpret and identify human intentions, which is a harder task again. In section 3, I suggest three possible standards of reliability in surrender recognition, which we might require of robots, and discuss their implications for the question of when robots will meet them. In Section 4, I examine a number of possible ways to try to avoid any problems that might arise as a result of the limited capacities of AWS to recognise surrender and argue that none of them are likely to entirely succeed. In section 5, I discuss the ethics of the use of AWS that cannot reliably recognise surrender and suggest that thinking about two different sorts of cases may usefully clarify the issues involved. We might think of AWS as being kept either on a "tight" or a "long" leash, depending upon how much opportunity they have for independent operations between release and impact. I also discuss several analogies that might assist us in thinking about these questions. In the final section, I draw out some of the implications of my discussion for the larger debate about the ethics of AWS and, in particular, for the vexed questions of whether it is more appropriate to think of these systems as weapons or as platforms and the locus of responsibility for civilian (and other) casualties caused by these systems.

§1. THE IMPORTANCE OF SURRENDER RECOGNITION

There is a powerful ethical case for conventions of war fighting that allow troops who wish to withdraw from participation in hostilities to surrender, which is that it greatly reduces the evils of war. Not only does it spare the lives of those who wish to surrender but it also saves the lives of those on the victorious side of the engagement who otherwise might have been killed had combat continued. These benefits — but especially the latter one — also establish a strong pragmatic grounds for troops to be willing to accept surrender and to support the development of expectations that will allow this practice.

The institution of surrender is so long established and fundamental to the ethics of war that it is actually under-represented in the law of war. In most statutes, surrender is mentioned only obliquely in connection with the offence of perfidy and in passing as one of the circumstances in which (those who were formally) combatants may become *hors de combat* and thus no longer legitimate targets of attack (Corn *et al.* 2012, 165-166). Nevertheless, Article 41 of *Additional Protocol I*, discussing "safeguard of an enemy *hors de combat*", clearly states that:

---

the first questions to arise in discussions regarding AWS in US military ethics and policy circles. Thus, for instance, Marchant, Allenby, Arkin *et al*. (2011) *mentions* the difficulties involved in recognising surrender. Similarly, Fielding (2006) flags the issues with which this paper is concerned. However, to my knowledge the current manuscript is the first full-length treatment of the ethical issues raised by the problem of surrender recognition.

[3] There is also a question about the capacity of AWS to *accept* surrender, which has also been neglected. However, for reasons of space I will not address this issue here except in passing where it is relevant to the purposes of the current manuscript.

> "1. A person who is recognized or who, in the circumstances, should be recognized to be 'hors de combat' shall not be made the object of attack."

And clarifies that.

> "2. A person is 'hors de combat' if:
> ...
> (b) he clearly expresses an intention to surrender…
> …
> provided that … he abstains from any hostile act and does not attempt to escape."[4]

Correspondingly, that is profoundly morally wrong to attack an enemy who has surrendered is a fundamental tenet of international customary law of war and, correspondingly, is almost universally acknowledged in the military law, codes, and rules of operations promulgated by nation states (International Committee of the Red Cross 2015, Rule 47: Attacks against Persons Hors de Combat, Practice Relating to Rule 47: Attacks against Persons Hors de Combat). Moreover, Article 23 (d) of the *Hague Regulations* prohibits ordering that no quarter should be given, while Article 40 of *Additional Protocol I* forbids "conducting hostilities on the basis of a no survivors policy and threatening the enemy that there shall be no survivors" (Dinstein 2004, p. 145).[5]

Thus, that there is both a legal requirement, and an ethical obligation, to refrain from attacking targets that have indicated the desire to surrender is abundantly clear.

§2. WHY THE RECOGNITION OF SURRENDER IS A HARD PROBLEM FOR ROBOTS

If AWS are unable to recognise surrender, then, this suggests that their deployment may be ethically problematic. There is no reason, in principle, why recognition of surrender should be impossible for robots: if human beings can do it, then so too, theoretically, could an appropriately sophisticated machine. Nevertheless, there are two reasons why recognising surrender *is* likely to be difficult for robots for perhaps the next several decades. The first relates to the fact that perception is itself a notoriously hard task for computers. The second relates to the contextual nature of the means used to signal surrender in different circumstances.

§2.1 THE PROBLEM OF PERCEPTION

When robots were first being developed it was widely believed that the main challenge would be to get them to solve meaningful problems, to reason and to "plan". However, it turned out that it was actually *perception* – the ability to form a model of the world and to locate themselves within it based upon information from their sensors — that robots struggled with

[4] *1977 Geneva Protocol I Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts*, Article 41, in Roberts and Guelff (2010), p. 443.
[5] *Annex to the Convention: Regulations Respecting the Laws and Customs of War on Land*, Article 23, in Roberts and Guelff (2010), p. 77; *1977 Geneva Protocol I Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts*, Article 40, in Roberts and Guelff (2010), p. 443.

and that has constituted the main obstacle to their use in more than a handful of roles (Brooks 2002, pp. 36-37). Despite significant progress in addressing this problem in recent years, robust real-time object recognition across a range of environments by systems in motion, in natural lighting conditions, remains beyond the capacity of even the most sophisticated computer vision systems. Object recognition and classification will be particularly difficult in military applications given that wars often take place in complex and chaotic environments, in various lighting conditions, and with smoke and fog obstructing the views of combatants (Krishnan 2009, p. 98).

The difficulty robots have with building up an accurate picture of the world is one of the reasons that some critics have been cynical about the capacity of AWS to reliably distinguish between military and civilian targets (Human Rights Watch 2012, pp. 31-33; Sharkey 2012b, p. 788; Krishnan 2009, pp. 98-99). However, as I suggested above, this problem may be soluble in some contexts. While perception in general remains a hard problem, recognition of specific items of interest in a scene (say, for instance, an enemy battle tank) is much more manageable. Indeed, some plausible targets for AWS, including submarines, naval vessels, fighter aircraft, and radar installations, have distinguishing features that make recognising them comparatively easy. Even the task of identifying enemy troops may be amenable to solution, for instance, by identifying every human-body-sized infrared heat-signature within half a metre of a short-wave-radar reflection characteristic of a firearm as an enemy soldier. Of course, this is only half of what is necessary in order for AWS to be able to identify the presence of objects and persons relevant to the requirements of *jus in bello*. In order to be able to refrain from attacks that would cause disproportionate civilian casualties, a robot must also be capable of identifying the presence of civilians and non-military objects in the battlespace. Again, however, I suspect that this problem may not be beyond the capacities of robots in some contexts. In antisubmarine warfare, for instance, there are highly unlikely to be any civilian targets that might be mistaken for an enemy submarine. Indeed, war at sea more generally often — if not always — occurs far from civilian shipping, while the existence of unique acoustic profiles for every ship provides an obvious mechanism whereby (some) robots might recognise both civilian and military shipping. Similarly, air-to-air combat may often proceed without concerns about causing collateral damage.

In any case, my aim here is not to settle the question of whether computers are likely to be able to meet these challenges but to point out that, even if they can, a further and significant challenge remains. Given the prohibition on attacking those who have surrendered, robots must also be capable of perceiving the changes in orientation and force posture of combatants that are conventionally associated with the indication of surrender. This is a much more difficult task. It is one thing, for instance, to be able to pick out human beings in a scene and identify them as enemy soldiers, it is another — and much more difficult — to tell when they have dropped their weapons, left cover, and put their hands up. In order to be able to recognise surrender then, robots will not just need to be able to recognise possible targets but also to recognise what they are doing.

§2.2 THE SIGNIFICANCE OF CONTEXT

In fact, recognising surrender is more difficult than even this description suggests. The actions that indicate surrender vary with context, both internationally, and also amongst different types of military units (Coleman 2013, p. 229). For this reason — and given the possibility that the forces involved in a conflict may be operating with different understandings as to the relevant conventions – recognising surrender is fundamentally a

question of recognising an intention. Indeed, recognising surrender requires the capacity to identify the presence or absence of a number of intentions. First, surrendering involves the intent to cease to participate in hostilities and to place oneself under the control of enemy troops. Second, it involves the intention to signal this so that others perceive the intention. However, there is a further intention that is necessary to surrender, which is the intention to make it such that a failure to perceive the intent to cease to participate in hostilities would be negligent or unreasonable. This third intention is necessary because there is a "performative" aspect to surrender (Austin 1962). That is, as long as certain "felicity conditions" are met — most obviously that one has the genuine intent to cease fighting — the indication of surrender just *is* surrender and, as if by magic, transforms one from being a legitimate target of attack to an illegitimate target. In order that such a miraculous transformation be achieved, the indication of surrender must have the declarative force aimed at by this third intention.

Human beings have a tremendously sophisticated and powerful capacity to interpret the actions of other human beings and to identify their intentions – to "read minds" — which has been honed by millennia of primate evolution wherein the ability to know what other individuals were thinking and were about to do provided a crucial selective advantage (Frith 2007; Premack and Woodruff 1978). It will be extremely challenging indeed for any machine to come close to replicating this.

The problem of surrender recognition is especially hard — and the role of context especially important — because of the relation between surrender and "perfidy". Essentially, perfidy is the attempt to manipulate the laws of war to one's own military advantage in a fashion that would, were it to become widely practised, undercut respect for those laws and, in particular, mean that it would be unreasonable to expect one's enemy to be bound by them in the future (Dinstein 2004, pp. 198-208). Feigned surrender is a paradigmatic case of perfidy (Dinstein 2004, p. 200). If troops who have indicated the desire to surrender recommence hostilities once the enemy has rendered themselves vulnerable by ceasing firing or moving so as to accept surrender then the opposing forces are unlikely to respect indications of surrender in the future. Because the institution of surrender is so valuable but also so vulnerable to being undercut in this fashion all parties to a conflict also have very strong reasons to punish instances of perfidy where they occur.

The possibility of perfidious indications of surrender means that not only must robots be capable of recognising the conventional indicators of surrender but they must be capable of distinguishing between real and feigned intentions. There are circumstances in which it *is* legitimate to attack enemy forces who are acting in a way that would ordinarily clearly indicate a desire to surrender — when it is reasonable to conclude that — in fact — their intent is perfidious (Coleman 2013, p. 235). However, assessing whether a signalled intention is likely to be perfidious or not is a significantly more difficult task than recognising the signal in the first place.

Of course if a unit indicates "surrender" to an AWS and then begins firing immediately after the AWS has aborted its attack, it is plausible to think that the weapon *would* be capable of recognising this as perfidy. However, the point is that in order to launch an attack on the same unit when it "surrendered" again one would have to be extremely confident indeed of the original identification of a (feigned) surrender *and* also that the target's intentions were the same in this case. That is, one would have to be confident that the actions that one originally took to indicate surrender were indeed intended to convey that intention duplicitously and that this subsequent indication of surrender was motivated by the same

intentions. Yet distinguishing real from feigned intentions where a target has indicated surrender is a very difficult task indeed.

If a weapon wasn't capable of recognising perfidy then it would be extremely vulnerable to being spoofed in this manner: potential targets would simply indicate surrender the moment they were vulnerable to being attacked in order to be spared and then recommence hostilities as soon as the AWS had moved away.[6] Such a response would be unethical and illegal but might go unremarked and unpunished unless the AWS was capable of recording/and or transmitting footage of their activities so it would be seen by human beings who might recognise it as perfidy and prosecute it as such. If the AWS were unable to transmit data to another location at the time then such instances of perfidious surrender might even go unpunished if the AWS was destroyed during the course of the engagement.

The contextual nature of the signals used to indicate surrender and the conceptual connection between surrender and perfidy mean that it will be extremely difficult indeed for robots to recognise surrender in many contexts.

§3. How reliable must a robot be?

Of course, human beings often fail to recognise surrender in war, with tragic results. It might therefore be argued that all that is required of robots in order to avoid any ethical problems arising out of the difficulties of recognising surrender is that they should be capable of doing so at least as well as human warfighters. Once a robot can meet this standard, it will be no more likely to attack a surrendered target than would be a human warfighters; once it exceeds it, replacing human warfighters with robots in the same role will save the lives of (some) surrendered troops (Arkin 2013).

However, as I will argue further below, this claim is properly controversial: we might well expect more of robots than this. In order to avoid prejudging this controversy, then, let me call the standard required in order for the specific set of ethical issues that might arise as a result of any inability of AWS to detect surrender to lapse "reliable" surrender recognition. With this stipulation in hand we can then proceed to examine the question of precisely where this standard should be set, which in turn will determine the likelihood that robots will be capable of achieving it within any given time frame.

There are, I think, at least two — but arguably three – places at which we might fix the standard for "reliability" in surrender recognition.

[6] If — as seems likely — AWS are unable to accept surrender by taking surrendered forces into custody this may further complicate the proper interpretation of intentions. If the enemy knows that AWS have no capacity to take them prisoner they might well surrender to an AWS, knowing when they do so that this will serve to protect them from attack while having no implication for their long-term capacity to participate in the armed conflict. Moreover, if there is no manned unit nearby capable of taking them prisoner they would indeed be within their rights to take up arms to rejoin the conflict after a reasonable period of time has elapsed. However, awareness of this possibility might in turn lead the parties to a conflict to suspect all indications of surrender to AWS as perfidious, which would be disastrous. My thanks to Dr Shane Dunn of the Australian Defence Science and Technology Organisation (DSTO) for drawing my attention to this issue. Note that if one believes that it is not possible to surrender unless there is a unit capable of accepting the surrender and taking effective control of the surrendered forces (Coleman 2013, p. 233) then in all likelihood it would be impossible to surrender to an AWS, which would absolve these systems of any requirement of being capable of recognising surrender, but arguably at the cost of rendering them unethical in a wide range of roles. This will be discussed further below.

As suggested above, we might judge robots reliable when they achieve or approach the performance of actual human warfighters in the field when it comes to the recognition of surrender: call this first standard, the "empirical standard of human warfighters".[7]

Importantly though, we already expect more than this of *human* warfighters. The actual performance of human beings in wartime is inevitably significantly less than our moral expectations with regard to surrender recognition because in reality human beings are sometimes negligent in their efforts to determine whether an enemy has surrendered or perhaps do not even bother to try to do so. At the very least, what we expect of human warfighters is that they make every reasonable effort to determine whether or not enemy troops have surrendered before they attacked them. Thus, we might judge robots reliable at surrender recognition when they reach or approach the standard of performance of human beings who are meeting their moral obligations in this regard: call this second possible place at which to fix the standard for reliability the "reasonable expectation standard".

However, it might be argued that the reasonable expectation standard mistakes an account of how we should evaluate agents for an account of their obligations. While we may not wish to blame or condemn warfighters who fail to meet this standard, what warfighters are actually required to do is to *never* attack a surrendered target. We might therefore only judge robots reliable at recognising surrender when they approach 100% accuracy in the task of recognising which enemy forces have surrendered and which have not: call this third — most demanding — standard, the standard of "perfection".[8]

How hard it will be for robots to detect surrender "reliably" — and consequently when they are likely to be able to do so — will depend upon which of these standards we believe is appropriate. Those who proffer a consequentialist account of the justification of the principles of *jus in bello* should clearly favour the first of these standards; once robots can meet this standard their use will reduce the evils of war. However, those who are inclined to understand the principle of distinction, in particular, as justified by a Kantian ethics or by deontological concerns more generally (Nagel 1972), should favour at least the "reasonable expectation standard" and *might* be tempted to insist on (near) "perfect" recognition. To insist on the "reasonable expectation standard" is just to insist that we do not owe surrendered combatants any less when we send a robot rather than a human being into combat. The case for expecting (near) "perfect" recognition from robots is more tendentious but still, I think, arguable. Despite their extraordinary capabilities, human beings are cognitively limited systems with an even more limited set of perceptual powers. In contrast, there is no obvious upper limit on the performance of a machine at recognising objects or intentions. When it comes to the appropriate ethical standards to impose on the performance of robots, then, there is no reason why we should take the performance of human beings as definitive. Given that surrendered combatants have a right to be protected *entirely* against attack, it might be argued that this is the appropriate standard to demand of a machine.[9]

[7] Arkin (2010 & 2013) suggests and defends this standard eloquently — albeit in the context of a discussion of the general requirement to target only combatants rather than the question of surrender recognition in particular.
[8] As the concept of reliability allows for the possibility of occasional failures, it is not plausible to demand perfect accuracy in order to judge a system reliable. However, *near* perfect accuracy is a plausible — if demanding — standard of reliability in some contexts.
[9] Again, given that it is impossible to demonstrate that a system is 100% reliable (even if a system has never failed, it remains possible that it will do so in the future) it seems that in practice we must settle for near-perfect performance even where morality demands perfection.

I am not going to attempt to settle here which of these is the appropriate standard to expect of robots when it comes to reliability at surrender recognition, which is a matter best settled in the context of a larger debate about the standards of ethical performance we should expect of robots. However, given the challenges involved in recognising surrender, discussed above, even if we settle for the lowest of these, the "empirical standard of human warfighters", it may be some time yet before robots are capable of reliably recognising surrender.

§4. Four Possible Solutions?

Thus far I have been concerned to establish only that detecting surrender is likely to be difficult for AWS and consequently that they may be unable to do so reliably for some years or perhaps even decades. It does not (yet) follow from this that their use would be unethical. In particular, if it is plausible to assign responsibility for the task of determining whether a target has surrendered or not to the person who launches the weapon, the inability of the weapon itself to detect surrender may pose no barrier to its ethical use. On the other hand, as I will discuss further below, for weapons with a long loiter time or a large degree of independence when it comes to determining which target they will attack, we might wonder whether this solution were available.

However, before turning to examine these questions, I want to first consider four different sets of policy responses, which, if successful, might avoid the need to confront them. Requiring AWS to seek permission from human controller before initiating an attack would avoid the need for the weapon itself to be able to recognise surrender. Radically constraining the nature of possible targets of AWS or munitions they could deploy, revising conventions regarding surrender so that all military units carried "surrender beacons", or confining the use of AWS to particular domains might either make it much easier for robots to recognise surrender or mitigate the need for them to do so.

§4.1 Requiring permission to attack from a human controller

The desire to placate widespread public unease about the prospect of robots being granted the power to kill, plus the widely acknowledged difficulties involved in producing "ethical" robots, have led a number of commentators to suggest that autonomous weapon systems could be required to seek permission from a human controller or supervisor before deploying lethal munitions or at least before deploying them in cases where the ethical deliberation required to determine whether or not an attack is permissible is beyond the capacities of the robot (Asaro 2012, p. 702; Brutzman *et al.* 2013; Gulam and Lee 2006, p. 132; Fitzsimonds and Mahnken 2007, pp. 101, 103; Kenyon 2006, p. 43; Leveringhaus and de Greef 2014).

However, there are two obvious difficulties with this proposal as a solution to the problems involved in surrender recognition. First, unless the robot was required to seek permission before each and every attack, this would not avoid the problems associated with the possibility of false negatives (that is, where the robot wrongly decides that a target is not surrendering — and is therefore a legitimate target). Second, and more importantly, requiring robots to seek permission from human beings before attacking targets would render them unable to carry out attacks in circumstances where communications were denied or otherwise unavailable or where the tempo of battle means that human beings can't make good decisions in the time available to them. As the capacity to operate in environments where communications are unavailable or unreliable is perhaps the main advantage of AWS over

tele-operated weapons (Anderson and Waxman 2013, p. 7; Schmitt and Thurnher 2013, p. 238) and the capacity to make decisions faster than human beings one of the main advantages of AWS over manned and tele-operated systems (Adams 2001), sacrificing these would mean sacrificing many of the benefits of AWS. Indeed, a system that required permission from human operator before it could attack a target would not be able to "select and engage targets without a human operator" and would therefore not constitute an autonomous weapon system according to the definition I cited above when operated in this fashion.

§4.2 Radically constrain the nature of possible targets or the munitions deployed by AWS?

Another workaround, which would allow AWS to operate across a wide range of domains but that might avoid any ethical problems associated with any inability of AWS to recognise surrender, would be to radically constrain the targets AWS were tasked to attack or the sorts of weapons with which they were armed. John Canning, who was (and is) an influential figure in the initial debate about the ethics of AWS in US policy circles was an early and strong advocate of this option (Canning 2006 & 2009). In order to minimise the chance of attacking a surrendered force, AWS might be programmed to attack only: (1) other unmanned systems; (2) forces that are actively firing their weapons at the time; or, (3) the enemy's weapons rather than persons or (whole) systems. Alternatively, we might only arm autonomous weapon systems with non-lethal weapons so as to avoid the risk of killing surrendered troops. While each of these possible solutions has its merits, unfortunately, they either do not succeed entirely in removing the need for AWS to be able to recognise surrender or they very seriously restrict the nature of the operations that AWS could carry out and therefore the military utility of these systems — or both.

The idea that wars of the future might be confined to battles between robots often comes up in discussions of the ethics of AWS, although in my experience it often reflects the desire to avoid engaging seriously with these issues rather than any real faith that such a circumstance will ever come about. Nevertheless, if AWS were tasked only with attacking other unmanned systems, it is true that this would avoid any need to be able to recognise surrender, as presumably there is no moral requirement not to attack "surrendered" targets except in the case where human lives are directly at stake.[10] The obvious cost of this solution, however, would be to massively restrict the military utility of AWS. Indeed, if AWS could only ethically be deployed against other AWS, there would be little incentive to develop such weapons in the first place.

Similarly, enemy units that are actively firing their weapons have clearly not surrendered and so concerns about the risks of attacking surrendered troops pose no barrier to attacking them. Again, however, restricting operations of AWS in this way would sacrifice a good portion of their military potential. One imagines that enemy troops would quickly learn to stop firing when AWS were within striking range — and it would be entirely ethical for them to do so; it is not perfidious to cease firing when confronted by superior enemy force in the hope of escaping their attention or of not being judged worthy of attack in the circumstances. While a weapon that could effectively suppress enemy fire merely by loitering in the area would have

[10] Of course, this approach substitutes the problem of distinguishing between manned and unmanned systems for the problem of recognising surrender. In some contexts, the former task may not be straightforward. For instance, robotic tanks may look very much like manned tanks. For that matter, in the future many military systems may have the capacity to function either as manned or unmanned systems. However, if it *were* possible to be confident that a potential target was unmanned then there need be little hesitation in attacking it

significant military utility, such a weapon could still not play a number of other militarily valuable roles, including attacking units that were strategically emplaced or manoeuvring.

Targeting only enemy weapons or deploying only non-lethal munitions from AWS would not reduce the difficulty of determining whether or not an enemy unit wishes to surrender but *would* reduce the risk of killing surrendered troops (Canning 2006 & 2009). Unfortunately, because neither of these policies would reduce this risk to zero, they do not mitigate the requirement not to attack surrendered units. Any projectile or energy emission powerful enough to disable an enemy's weapons will usually impose some risk to human life. For instance, even low-energy kinetic attacks on artillery pieces, the canons of tanks, or the missile rails of aircraft risks killing their crew, while attacking naval guns may hole the ship and endanger the lives of everyone on board. "Non"-lethal weapons are better described as "sub"-lethal, as almost all carry some risk of killing people in particular circumstances.[11] Rubber bullets may strike people in the temples or eye sockets, gases may cause asthmatics to asphyxiate, microwave based area denial weapons may cause heart attacks or burns on those who are unable to leave the area of effect for some reason, *et cetera*. For these reasons, forces that have surrendered have the right not to be attacked even with weapons of this sort.[12]

§4.3 SURRENDER BEACONS?

To this point I have been discussing how an AWS might recognise surrender and presuming, if implicitly, that, if they cannot do so reliably, this may exclude their use in some circumstances. However, the history of warfare consist in repeated episodes of weapons that were morally controversial when first invented being deployed for the sake of military advantage, whereupon the way in which wars are fought *and* the legal and ethical conventions governing military conduct each evolved to take account of the new weapons. With this history in mind, then, one might instead frame the ethical question as "what are the obligations on military forces as a result of the need to clearly convey surrender to AWS with the limitations I have described above, given that in all probability these weapons will be deployed in a wide range of roles in future conflicts?"

One option would be to insist that troops should be capable of communicating surrender to AWS. For instance, were all military units to carry electronic devices capable of emitting an internationally agreed upon "surrender signal" on an agreed-upon frequency, then activating this beacon would serve to protect them from attack by AWS should they wish to surrender. It is perhaps plausible to imagine such a system being fitted to aircraft, naval systems, and armoured units, where it would also have the advantage of facilitating surrender to human forces. Unfortunately, however, it is a bit more of a stretch to imagine every infantry unit in the world carrying such a device, let alone every soldier — and it is wildly implausible to imagine that irregular militias and insurgents will have the resources to equip their members with an electronic beacon in order to facilitate surrender to AWS. While there are, of course, questions about the proprietary of the participation of such forces in armed conflict, where they are involved it is both morally incumbent and politically advantageous to be able to recognise and accept their surrender. At most then, such a convention would only be a partial

[11] In fairness to Canning, in his publications he imagines AWS being armed with weapons that *are* capable of disarming soldiers without harming them (he mentions, for instance, the use of diamond tip saws to destroy rifles): my point here is effectively to dispute the likelihood and practicality of this scenario in practice.
[12] Presuming that they are complying with the reasonable directives of the forces to whom they have surrendered.

solution to the problem of surrender recognition and might succeed only in conflicts involving regular forces on both sides.[13]

Moreover, the introduction of such beacons would only solve the problem posed by the introduction of AWS if it was accompanied by a further radical modification of existing conventions regarding surrender, which we should countenance only after careful deliberation, if at all. That is, it would have to become an established understanding that military units remain legitimate targets *unless* they have activated their beacons, with all other means of surrender ruled out. Otherwise, AWS would still need to be capable of recognising surrender by (all the various) conventional means. Adopting such a policy would obviously be disastrous for any troops who had been separated from their surrender beacons or whose surrender beacons had become inoperable for some reason, as they would then be unable to surrender to an AWS.

Whether such a circumstance — and the policy that produced it — would be acceptable or not is, I think, a contestable matter. On the one hand, as noted above, it is explicitly forbidden to order that there should be "no quarter given", which suggest that it would be problematic to deploy a weapon that might attack troops who were clearly — if not to the machine — indicating the desire to surrender. On the other hand, no person or policy can guarantee that every attempt to surrender will be successful: there will always be — tragic — situations where signals are missed or intentions misunderstood. It might therefore be argued that providing AWS with the capacity to recognise surrender beacons in the context of a convention which requires military units to carry such beacons exhausted the obligations of the designers of these weapons.

§4.4 Confining the use of AWS to particular domains

Like excluding the presence of civilian objects, surrender is a more tractable problem in some sorts of warfare than others. It is extremely difficult when targeting infantry or irregular forces in urban or jungle environments because both the difficulties posed by perception and the importance of context are at their most acute in this setting. Yet it is effectively non-existent in air-to-air combat, where there are currently no widely-agreed upon mechanisms to allow aircraft to surrender. In several other types of warfare, the conventions regarding surrender make it much more plausible to think that robots could recognise surrender with a high degree of accuracy. In combat between armoured vehicles for instance, units that wish to indicate surrender typically reverse the turret of the tank, open the crew hatches, and place themselves on top of — or alongside of — the vehicle. It is plausible to think that robots might be capable of recognising this.[14] In operations conducted against submarines, submarines surrender by communicating their intentions via "underwater telephone" or by surfacing and flying a white flag while the crew takes to the deck or the boats, which again robots might be capable of recognising.[15] The flags and lights that naval vessels use to indicate surrender are also the sorts of signals that machines are already reasonably

[13] Radio beacons — and a convention regarding their use — *would* solve another problem that is likely to beset (some) autonomous systems, which is the need to be able to reliably identify military ships that have taken on the role of hospital ships: it is much more plausible to expect that any ship that wishes to cease to participate in hostilities in order to take on board and care for the wounded should begin broadcasting a signal indicating that it is doing so. Similarly, beacons might also help to protect other civilian installations such as churches, hospitals, et cetera, from inadvertent attack by AWS. My thanks to Tim McCormack for drawing my attention to this possibility.

[14] Admittedly, this doesn't resolve the problems arising out of the possibility of perfidious surrender.

[15] My thanks to Rob McLaughlin for clarifying these conventions for me.

competent at detecting and interpreting. Thus, by confining AWS to roles in which they are tasked with attacking only targets of these sorts, it would be possible either to avoid the problems associated with recognising surrender (in the case of air-to-air combat) or to make it much more likely that robots could recognise surrender.

Of the possible policy solutions I have considered here, this is the one that I believe offers the best prospects of allowing AWS to play a valuable military role without courting ethical controversy due to an inability to recognise surrender.[16] However, even this approach is likely to significantly restrict the sorts of warfare in which AWS could be used.

§5. IF ROBOTS CANNOT RELIABLY RECOGNISE SURRENDER…

None of the approaches I have surveyed thus far succeed, therefore, in avoiding the ethical issues associated with an incapacity to reliably recognise surrender without imposing severe restrictions on the roles in which they may be used. Of course, it might also be possible to combine the various approaches I have discussed here in different ways in order to allow AWS to operate in more domains or attack a wider variety of targets whilst still (mostly) being able to recognise surrender. For reasons of space I cannot consider all the possible options here. However, I hope my discussion of the promise and limitations of each approach will prove useful to such a project.

For the remainder of this paper, I wish to consider the ethics of deploying AWS that cannot reliably recognise surrender. I want to suggest that two different ways of framing this issue produce very different results. In cases where there is a limited window of opportunity for targets to surrender between an AWS being released and its impact, the necessity for the AWS to be able to recognise surrender is greatly diminished: I will refer to this as the AWS being "on a tight leash".[17] However, in cases where the AWS may travel or loiter for a significant period between release and destruction of its target or (perhaps) where the AWS has a high degree of autonomy in which particular target it will attack, the requirement that the AWS be capable of reliably detecting surrender is much more pressing: I will refer to this as the AWS being "on a long leash". Moreover, I suggest, there are existing or historical analogies that are foregrounded by adopting each approach. Thus, we may be able to make significant progress on resolving the ethics of the use of AWS if we can decide which of these ways of framing is more appropriate in any particular case.

§5.1 AWS ON A TIGHT LEASH?

Where AWS travel at very high speeds (or where they are used at short range) and where they are tasked with selecting amongst a small number of targets that are already under direct observation when the AWS is launched, it might be argued that there is no need for the AWS

[16] Note that this was effectively the approach adopted with the United States' MK 60 CAPTOR (encapsulated torpedo) antisubmarine mine, which arguably should be classified as an autonomous weapon system: as this weapon was designed to target only enemy submarines underwater, there was very little risk that it would attack a surrendered target.

[17] I have struggled to find a form of words to describe the ultimate destructive effects of the operations of AWS on its target that does not prejudice the question of whether the AWS is a weapon with which the operator attacks a target or a system or platform which itself launches an attack on the target, which is a question I wish to leave open in this paper. I have therefore settled for speaking of the "impact" or "strike" of an AWS but it is important to emphasise that I understand these expressions to include cases where the AWS itself launches a sub-munition to destroy a target.

to be capable of recognising surrender. Instead, it would be up to the human being who launches the AWS to ensure that any of the potential targets were not signalling surrender before they release the AWS.

The plausibility of this approach is suggested by consideration of a hypothetical scenario involving weapon that has been in operation for nearly a century — a "dumb" torpedo – or, perhaps more plausibly, another contemporary weapon — a "fire-and-forget" torpedo with an active sonar homing system. Because such weapons may travel in the water for a number of minutes between firing and reaching their target, it is theoretically possible that the enemy unit being targeted might indicate surrender between the launch of the weapon and its impact.[18] Were such a thing to happen, it would be a tragedy, but not a war crime. As long as the target was a legitimate target when the weapon was launched and had made no indication that they wished to surrender or (perhaps) were about to surrender, then the person who authorised the release of the weapon bears no moral responsibility for the tragic outcome. Moreover, while we might wish that it were possible to abort the torpedo's run once it were realised that the target had surrendered, it does not seem as though there is any ethical problem arising from the fact that the torpedo itself isn't capable of recognising surrender and aborting its attack. If we believe that this is an appropriate analogy than it may seem that an inability to detect surrender need not prohibit the ethical use of AWS.

§5.2 AWS ON A LONG LEASH?

While weapons systems capable of choosing between a limited number of targets and operating over a narrow timeframe will qualify as AWS, there is a sense in which they are not "very" autonomous. Much of the military potential of AWS consists in their (theoretical) capacity to operate with very long loiter times and to engage targets of opportunity that may not have been explicitly singled out for attack when the system was launched. We might therefore think of such systems as operating on a "long leash". A paradigmatic example of an AWS operated on a long leash, for instance, might involve an autonomous hunter-killer Unmanned Undersea Vehicle (UUV) tasked with attacking all military shipping within some geographically defined (wide) area.[19]

Whether operating on a long leash is any different to operating on a tight leash, morally speaking, is, I think, the central question when it comes to the ethics of the use of AWS that are unable to reliably recognise surrender (Fielding 2006, p. 102).

At the very least, the chances of a target surrendering between the launch of the AWS and its impact are larger when AWS are operating on a long leash. This increase is not proportional with time, as one presumes that surrendered targets would, after some discrete period of time, be taken into custody and exit the battlespace. Nevertheless, it is clear that AWS operating on a long leash have a significantly greater chance of encountering a surrendered target.

---

[18] Unclassified sources suggest that the US's Mk 48 ADCAP heavyweight torpedo might travel for as long as 40 minutes between launch and impact when used to attack a target at maximum range (see range and speed figures given at http://web.archive.org/web/20010401035621/http://www.janes.com/defence/naval_forces/news/juws/juws010202_1_n.shtml), while several of the heavyweight torpedoes fielded by other nations might travel for approximately 30 minutes (Fuller 2010).

[19] The US Office of Naval Research (ONR) has announced its interest in providing the U.S. Navy's Large Displacement Unmanned Undersea Vehicle (LDUUV) with an Anti-Submarine Warfare capability (Scott 2014), which suggests that this prospect is more than hypothetical.

Moreover, the person who authorises the release of the AWS has little sense of whether particular targets have surrendered or are about to surrender.

However, it might be argued that this fact is not distinguish the ethics of operating AWS on a long leash from the ethics of operating them on a tight leash, as in each case responsibility for the consequences of the attack rest with the person who authorises the launch of the AWS (Schmitt 2013). According to this way of thinking, as long as the chance of striking a surrendered target does not exceed some reasonable threshold, there will be nothing ethically problematic about using the AWS. This calculation will need to take into account both the capacities of the AWS to recognise surrender (in what percentage of cases does it fail to do so?) and the chance that a target might have surrendered between launch and impact, which in turn will depend on the AWS' role, area of operations, and targeting criteria.

Yet, this sanguine attitude might be challenged in two ways.

First, even where the risk of attacking a surrendered target is judged acceptable, it might be argued that making this calculation is not sufficient to count as taking "reasonable precautions" to avoid attacking surrendered targets in the circumstances.[20] When AWS are being used on a long leash, when it comes to any particular target engaged by the AWS, no human being has assessed whether that target has surrendered or not, while (*ex hypothesi*) the AWS itself is not capable of reliably determining this. If we understand the requirement to take reasonable precautions as being founded in an obligation to the particular person whose life is on the line when the attack is being contemplated, rather than as a product of a generalised obligation to avoid non-combatant casualties then it is arguable that the calculation before launch that the AWS is unlikely to attack a surrendered target is not sufficient to exhaust this obligation.[21]

Second — and relatedly — the use of AWS on a long leash might be thought to run afoul of the prohibition on issuing orders that there should be "no quarter" given. This objection seems especially compelling if one believes that enemy forces *cannot* surrender to an AWS because the AWS has no means of "accepting" surrender (Coleman 2013, p. 233). If this is true then although it removes the necessity for AWS to be capable of recognising surrender it also would, I believe, prohibit using them on a long leash. Even if one denies — as I believe we should — that the lack of the capacity of AWS to render enemy combatants who wish to surrender prisoners of war excuses them from the requirement to be able to recognise surrender, the prohibition on ordering that there shall be no quarter given might be thought to render the use of AWS on a long leash morally problematic. In such a circumstance, enemy forces who wish to surrender may have no opportunity to do so because the AWS fails to recognise their attempt; moreover, the person who authorised the release of the AWS was (or at least should have been) aware of this when the system was deployed. Of course, strictly speaking, the intention of those deploying an AWS need not be that units should have *no* opportunity to surrender (they might, for example, plausibly wish that the system they were using was more capable of recognising surrender); rather, they are guilty of employing a means of warfare that fails to safeguard the opportunity to surrender. Whether this objection will have force or not in such cases, will therefore depend on whether the prohibition on

[20] *1977 Geneva Protocol I Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts*, Article 57 (2) (a), in Roberts and Guelff (2010), pp. 452-453. For discussion of the nature and significance of the obligation on warfighters to take reasonable precautions in attack, see: Dinstein 2004, pp. 125-128; Rogers 2012, pp. 125- 159, 160-174.

[21] Just how plausible it would be to attempt to ground the obligation to take reasonable precautions in attack along these lines is a further question, which reasons of space prevent me from attempting to answer here.

ordering that no quarter should be given is understood as requiring combatants to safeguard the opportunity to surrender or merely not to intentionally deny it absolutely.[22]

According to each of these objections, then, the period of time between the release and the impact of the AWS *is* morally significant by virtue of the extent to which it moves the burden of recognising surrender from the person authorising the release of the AWS to the system itself.

In some ways, the issues in the debate ethics of operating AWS on a "long leash" are similar to those in the historical (and ongoing) debate about the ethics of mine warfare. Mines may detonate long after they emplaced and without regard to whether or not their targets have surrendered (or, indeed, are combatants at all). The lack of control that those who emplace or lay mines have over the nature of the targets the mines attack has led to mines becoming controversial. Anti-personnel mines are banned by the *Ottawa Convention*, while the use of free floating contact mines that do not become harmless one hour after they are deployed in naval warfare is explicitly prohibited by Article 1 (1) of *Hague Convention VIII*.[23] Critics of AWS might push this analogy in order to insist that the use of AWS on a long leash should similarly be prohibited on the grounds that those who deploy them cannot adequately control where they strike. Enthusiasts for AWS are likely to reply that contemporary anti-tank and tethered naval influence mines are capable of a high degree of discrimination between civilian and military targets and are not prohibited by IHL despite the fact that they cannot recognise surrender. While evocative, then, the analogy with mine warfare seems unlikely to settle the question of the ethics of the use of AWS which cannot recognise surrender on a long leash.

Ultimately, I remain myself conflicted about the ethics of the use of AWS that cannot recognise surrender on a long leash. If the chance of them striking a surrendered target is low enough — taking into account both their capacity to recognise a surrendered target and the chance that they will encounter a surrendered target given their role, area of operations, and targeting criteria — then perhaps it would be ethical for the person authorising the release of the AWS to accept responsibility for the consequences of its deployment, including the possibility that the system will attack a surrendered target. Yet the intuition that that it would be wrong to launch a system that might strike a target months later, regardless of whether or not it had surrendered, remains. As my own thinking on this topic remains unsettled, I can only hope that my treatment here will help others in their thinking about these issues.

§6. PLATFORMS, WEAPONS, AND RESPONSIBILITY

The preceding discussion intersects at a number of points with discussions of two key controversies in the larger debate about the ethics of AWS: the appropriate locus of moral responsibility for casualties produced by AWS; and, whether AWS should be thought of as weapons or weapon platforms. A brief consideration of these points of intersection may, I

[22] As Coleman (2013, p. 237) notes, combatants are typically *not* held to be under an obligation to provide enemy forces with an explicit opportunity to surrender before attacking.

[23] *1997 Ottawa Convention on the Prohibition of the Use, Stockpiling, Production and Transfer of Anti-Personnel Mines and on their Destruction*, in Roberts and Guelff (2010), pp. 645-666; *1907 Hague Convention VIII Relative to the Laying of Automatic Submarine Contact Mines*, in Roberts and Guelff (2010), pp. 103-110. For discussion see Doswald-Beck (1995), p. 171.

hope, cast some light on these questions as well as the ethics of the use of AWS that are unable to reliably recognise surrender.

One of the first controversies to erupt as the prospect of AWS emerged concerned the question of the appropriate locus of responsibility for deaths caused by these systems. Critics alleged that the development of AWS risked no one being responsible for the consequences of their use (Sparrow 2007). The person who releases the weapon cannot be held responsible for the choices and decisions of the robot, while the robot itself is not the sort of thing that can be held morally responsible; thus, a "responsibility gap" emerges (Matthias 2004; Roff 2013). This claim remains contested and a number of authorities have argued that the attribution of responsibility to the person who authorises the release of the weapon is, in fact, straightforward, with talk of a "responsibility gap" obfuscating this by mis-attributing a mysterious quasi-moral agency to robots (Lokhorst and van den Hoven 2012, pp. 150-151; Schmitt 2013, p. 33).

Whether AWS should be thought of as weapons or platforms is controversial because of various proposals to prohibit AWS by means of international law (Altmann 2013; Campaign to Stop Killer Robots 2015; Human Rights Watch 2012; O'Connell 2014; Wallach and Allen 2013). If there is, as some have argued, something especially wrong about killing people with robots (Asaro 2012; Gubrud 2014, p. 40; Sharkey 2012a) — and robots are weapons – then it is possible (although obviously controversial) that they should be considered *mala in se* and prohibited as such (Wallach 2013). If, on the other hand, AWS are better thought of as platforms (which might be used to deliver different sorts of weapons) then it would be difficult indeed to explain how the mere fact that a weapon was mounted on a AWS should make it an "evil means" for killing; moreover, there is little historical precedent for banning a platform.

These two controversies are already intertwined: if we assign responsibility to the person who uses the AWS to kill, then the robot is clearly the means by which they kill — and thus a weapon; moreover, the possibility opens up that this means itself might be morally problematic. If AWS are platforms, then *they* attack targets with weapons and it is most natural to look to assign responsibility for targeting decisions to the controller (the computer) on the platform. However, this dialectic become still clearer in the light of the preceding discussion of the ethics of surrender recognition.

Notice, for instance, how my treatment of the ethics of use of AWS on a tight leash assigns responsibility to the person who launches the AWS and treats AWS as analogous to other *weapons*, which might also occasionally strike surrendered targets. Yet this has a number of challenging implications for the ethics of the use of AWS that are unable to reliably recognise surrender. If AWS are weapons then launching an AWS is launching an attack. Moreover, it seems most natural to think of this as launching an attack against *all* of the targets that the AWS might in fact strike.[24] If including a military unit within the targeting criteria of an AWS counts as attacking that unit, though, then the chance of attacking a surrendered target increases with the size of the weapon's target set *regardless of the capacity of the AWS to detect surrender*.[25]

[24] One suspects, for instance, that the launch of an AWS into a position from which it might attack will be perceived as a hostile act by any military unit within its range.

[25] Note that by target set here I mean the number of enemy forces or units the AWS might strike as a result of its targeting criteria rather than the number of targets the individual who launched it intended it to attack: the former figure may be larger than the latter where new targets fulfilling its targeting criteria enter the battlespace.

This implication in turn suggests that the use of AWS on a long leash will be problematic while they are unable to reliably recognise surrender: AWS on a long leash will tend to have larger target sets, both because having more autonomy to select targets is one way in which an AWS may *count* as being on a long leash and because the longer the period between the release of an AWS and its impact the more opportunity there is for unanticipated targets to come to fulfil its targeting criteria. Interestingly, then, if AWS that are unable to reliably recognise surrender are to be used on a long leash a "responsibility gap" is actually *required*. Conceiving of these systems as platforms which themselves launch attacks is one way to open up this gap. Of course, if we do have the intuition that it is important that someone should be held morally responsible for each and every use of lethal force in the course of war (Sparrow 2007), then the use of AWS on a long leash may be problematic for this reason. Thinking about the ethics of surrender recognition highlights the persistence and significance of intuitions about the attribution of responsibility even where the nature of the AWS is not such as to raise questions about its moral agency (Roff 2013).

I do not pretend to have attempted to settle here either the appropriate locus for the attribution of responsibility for casualties produced by AWS or the question of whether (or, better, perhaps, which) AWS should be thought of as weapons or platforms: these are matters for a much larger — and longer — debate. Again, my hope is merely that these reflections on the ethics of the use of AWS that are unable to reliably recognise surrender might cast some light on these larger questions.

CONCLUSION

I have argued that the difficulties involved in accurately identifying the nature of the actions of potential targets and the role played by context in determining surrender mean that the recognition of surrender will be a profound challenge for autonomous weapon systems. Even if we ask only that robots be capable of recognising surrender at close to the level achieved by human beings in wartime, reliable surrender recognition may be beyond the capacity of machines, in some contexts at least, for some years to come. A lack of the capacity to reliably recognise surrender would not rule out the ethical use of AWS in some roles where the question of surrender recognition seldom, if ever, arises, such as attacks on aircraft in flight or submarines while submerged. Moreover, various policies, discussed in Section II, or combinations thereof, might mitigate the danger of attacking surrendered targets in some (other) contexts. Nevertheless, the lack of the capacity to reliably distinguish surrender would problematise the use of AWS in a wide range of militarily valuable roles. I have suggested that the lack of the capacity to reliably recognise surrender need not rule out their ethical use where AWS could plausibly be described as operating on a "tight leash", such that it was appropriate to assign responsibility for surrender detection to the person who authorises the release of the weapon. However, it is possible that some AWS, with more choice about which targets to engage and/or long periods of time between release and impact, should be better thought of as operating on a "long leash". I have suggested that such applications are likely to be controversial and thus a crucial test for the moral permissibility of the use of AWS that are unable to reliably recognise surrender. While I have been unable to settle the question of the ethics of the use of AWS on a long leash, I have tried to clarify the arguments that might plausibly be made for or against them. I have also highlighted a number of historical analogies which are helpful for thinking through these questions. Finally, I have explored the connections between the issues discussed in this paper and two important controversies in the larger debate concerning the ethics of AWS.

It is possible that progress in the science and technology of artificial intelligence will eventually allow robots to achieve whatever standard of surrender recognition we believe to be required of them. Until that day, the questions I have raised and tried — if not entirely successfully — to answer here will remain crucial to the ethics of the design and use of AWS. Given that the anticipated military value of AWS will establish a strong dynamic driving towards their deployment and use it is vital that philosophers and ethicists consider these matters further.

ACKNOWLEDGEMENTS

I would like to thank Professor Tim McCormack, Dr Stephen Coleman, Professor Philip H. S. Torr, Dr BJ Strawser, Professor George Lucas, Dr Shane Dunn, and Associate Professor Rob McLaughlin, for conversations and correspondence about these topics early in the process of the development of this manuscript. Mark Howard ably assisted me with sources and with preparing the paper for publication.

REFERENCES

Adams, T. K. 2001. Future warfare and the decline of human decision making. *Parameters: US Army War College Quarterly* (Winter): 57-71.

Altmann, J. 2013. Arms control for armed uninhabited vehicles: An ethical issue. *Ethics and Information Technology* 15(2): 137-152.

Anderson, K., and M. Waxman. 2013. Law and ethics for autonomous weapon systems: Why a ban won't work and how the laws of war can. *Jean Perkins Task Force on National Security and Law Essay Series*, The Hoover Institution, Stanford University; American University, WCL Research Paper 2013-11; Columbia Public Law Research Paper 13-351. Available at: http://ssrn.com/abstract=2250126 or http://dx.doi.org/10.2139/ssrn.2250126

Arkin, R. C. 2010. The case for ethical autonomy in unmanned systems. *Journal of Military Ethics* 9(4): 332-341.

Arkin, R. 2013. Lethal autonomous systems and the plight of the non-combatant. *AISB Quarterly* 137: 1-9.

Asaro, P. 2012. On banning autonomous weapon systems: Human rights, automation, and the dehumanization of lethal decision-making. *International Review of the Red Cross* 94(886): 687-709.

Austin, J. L. 1962. *How to Do Things with Words*. Oxford: Clarendon Press.

Borenstein, J. 2008. The ethics of autonomous military robots. *Studies in Ethics, Law, and Technology* 2(1): Article 2. DOI: 10.2202/1941-6008.1036.

Brooks, R. A. 2002. *Robot: The future of flesh and machines*. London: Penguin Books.

Brutzman, D. P., D. T. Davis, G. R. Lucas, and R. B. McGhee. 2013. Run-time ethics checking for autonomous unmanned vehicles: Developing a practical approach. *Proceedings of the 18th International Symposium on Unmanned Untethered Submersible Technology (UUST),* Portsmouth, New Hampshire, August 2013. Available at:

https://savage.nps.edu/AuvWorkbench/website/documentation/papers/UUST2013PracticalRuntimeAUVEthics.pdf (accessed October 24, 2013).

Campaign to Stop Killer Robots. 2015. Campaign to stop killer robots: About us. Available at: http://www.stopkillerrobots.org/about-us/ (accessed April 25, 2015).

Canning, J. S. 2006. *A concept of operations for armed autonomous systems*. Paper presented at the 3rd Annual Disruptive Technology Conference, Washington, DC. Available at: http://www.dtic.mil/ndia/2006disruptive_tech/canning.pdf

Canning, J. S. 2009. You've just been disarmed. Have a nice day! *IEEE Technology and Society Magazine* 28(1): 12-15.

Coleman, S. 2013. *Military ethics: An introduction with case studies*. New York: Oxford University Press.

Corn, G. S., V. Hansen, R. B. Jackson, C. Jenks, E. T Jensen, and J. A. Schoettler. 2012. *The law of armed conflict: an operational approach*. New York: Wolters Kluwer Law & Business.

Dinstein, Y. 2004. *The conduct of hostilities under the law of international armed conflict*. Cambridge: Cambridge University Press.

Doswald-Beck, L. ed. 1995. *San Remo manual on international law applicable to armed conflicts at sea*. Cambridge; New York: Cambridge University Press.

Fielding, M. 2006. Robotics in future land warfare. *Australian Army Journal* 3(2): 1-10.

Fitzsimonds, J. R., and T. G. Mahnken. 2007. Military officer attitudes toward UAV adoption: Exploring institutional impediments to innovation. *JFQ: Joint Force Quarterly* 46(3rd Quarter): 96-103.

Frith, C. D. 2007. The social brain? *Philosophical Transactions of the Royal Society B: Biological Sciences* 362(1480): 671-678.

Fuller, M. 2010. Silent might: heavyweight torpedoes still pack a punch. *Janes Navy International* 115(5): 24-30.

Guarini, M., and P. Bello. 2012. Robotic warfare: Some challenges in moving from noncivilian to civilian theaters, in *Robot ethics: The ethical and social implications of robotics*, eds. P. Lin, K. Abney, and G. A. Bekey. Cambridge, MA; London, England: MIT Press, 129-144.

Gubrud, M. 2014. Stopping killer robots. *Bulletin of the Atomic Scientists* 70(1): 32-42.

Gulam, H., and S. W. Lee. 2006. Uninhabited combat aerial vehicles and the law of armed conflict. *Australian Army Journal* 3(2): 1-14.

Human Rights Watch. 2012. *Losing humanity: The case against killer robots*. Available at: http://www.hrw.org/reports/2012/11/19/losing-humanity-0 (accessed December 1, 2014).

International Committee of the Red Cross. 2015. *Customary IHL database*. Available at: https://www.icrc.org/customary-ihl/eng/docs/home (accessed July 7, 2015).

Kenyon, H. S. 2006. Israel deploys robot guardians. *Signal* 60(7): 41-44.

Krishnan, A. 2009. *Killer robots: Legality and ethicality of autonomous weapons*. Farnham, England: Ashgate Publishing.

Leveringhaus, A., and T. de Greef. 2014. Autonomous weapons: A qualified defence, in *Precision-strike technology and international intervention: Strategic, legal and moral*

implications, eds. T. Dyson, W. Aslam, R. Rauxloh, and M. Aaronson. Abingdon and New York: Routledge.

Lokhorst, G.-J., and van den Hoven, J. 2012. Responsibility for military robots, in *Robot ethics: the ethical and social implications of robotics*, eds. P. Lin, K. Abney, and G. A. Bekey. Cambridge, MA; London: MIT Press, 145-156.

Marchant, G. E., B. Allenby, R. Arkin, E. T. Barrett, J. Borenstein, L. Gaudet, O. Kittrie, P. Lin, G.R. Lucas, and R. O'Meara. 2011. International governance of autonomous military robots. *Columbia University Review of Science & Technology Law* 12: 272-315.

Matthias, A. 2004. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology* 6: 175-183.

Nagel, T. 1972. War and massacre. *Philosophy and Public Affairs* 1: 123-144.

O'Connell, M. E. 2014. Banning autonomous killing, in *The American way of bombing: How legal and ethical norms change*, eds. M. Evangelista, and H. Shue. Ithaca, NY: Cornell University Press.

Premack, D., and G. Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences* 1(4): 515-526.

Roberts, A., and R. Guelff. eds. 2010. *Documents on the laws of war, Third Edition*. Oxford: Oxford University Press.

Roff, H. M. 2013. Responsibility, liability, and lethal autonomous robots, in *Routledge handbook of ethics and war: Just war theory in the 21st Century*, eds. F. Allhoff, N. Evans and A. Henschke. Milton Park, Oxon: Routledge.

Rogers, A.P.V. 2012. *Law on the battlefield, Third Edition*. Manchester: Manchester University Press

Schmitt, M. 2013. Autonomous weapon systems and international humanitarian law: A reply to the critics. *Harvard National Security Journal*. Available at: http://harvardnsj.org/2013/02/autonomous-weapon-systems-and-international-humanitarian-law-a-reply-to-the-critics/.

Schmitt, M. N., and J. S. Thurnher. 2013. 'Out of the loop': Autonomous weapon systems and the law of armed conflict. *Harvard National Security Journal* 4(2): 231-281.

Scott, R. 2014. ONR to swim ahead on ASW package for large UUV. *Jane's Navy International*. 20 November, 2014.

Sharkey, N. 2008. Cassandra or false prophet of doom: AI robots and war. *IEEE Intelligent Systems* 23(4): 14-17.

Sharkey, N. 2012a. Autonomous robots and the automation of warfare. *International Humanitarian Law Magazine* 2: 18-19.

Sharkey, N. 2012b. The evitability of autonomous robot warfare. *International Review of the Red Cross* 94(886): 787-799.

Singer, P. W. 2009. *Wired for war: The robotics revolution and 21st century conflict*. New York: The Penguin Press.

Sparrow, R. 2007. Killer robots. *Journal of Applied Philosophy* 24(1): 62-77.

Sparrow, R. 2009a. Building a better warbot: Ethical issues in the design of unmanned systems for military applications. *Science and Engineering Ethics* 15(2):169–187.

Sparrow, R. 2009b. Predators or plowshares? Arms control of robotic weapons. *IEEE Technology and Society* 28(1): 25-29.

Sparrow, R. 2011. Robotic weapons and the future of war, in *New wars and new soldiers: Military ethics in the contemporary world*, eds. J. Wolfendale, and P. Tripodi . Surrey, UK; Burlington, VA: Ashgate, 117-133.

Sparrow, R. 2015. *Robots and respect: Assessing the case against autonomous weapon systems* (Manuscript under consideration).

United States' Department of Defense. 2012. *Directive 3000.09: Autonomy in Weapon Systems, Nov. 21, 2012*. Available via: http://www.dtic.mil/whs/directives/index.html.

Wallach, W. 2013. Terminating the Terminator: What to do about autonomous weapons. *Science Progress*, January 29. Available at: http://scienceprogress.org/2013/01/terminating-the-terminator-what-to-do-about-autonomous-weapons/.

Wallach, W., and C. Allen. 2013. Framing robot arms control. *Ethics and Information Technology* 15(2): 125-136.

Wagner, M. 2011/12. Taking humans out of the loop: Implications for international humanitarian law. *Journal of Law Information and Scienc*e 21(2): 155-165.