Journal of Military Ethics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/smil20

THE CASE AGAINST ROBOTIC WARFARE: A RESPONSE TO ARKIN

Ryan Tonkens [a]
[a] Novel Tech Ethics, Dalhousie University

Version of record first published: 10 Sep 2012.

To cite this article: Ryan Tonkens (2012): THE CASE AGAINST ROBOTIC WARFARE: A RESPONSE TO ARKIN, Journal of Military Ethics, 11:2, 149-168

To link to this article: http://dx.doi.org/10.1080/15027570.2012.708265

THE CASE AGAINST ROBOTIC WARFARE: A RESPONSE TO ARKIN

Ryan Tonkens

Novel Tech Ethics, Dalhousie University

Semi-autonomous robotic weapons are already carving out a role for themselves in modern warfare. Recently, Ronald Arkin has argued that autonomous lethal robotic systems could be more ethical than humans on the battlefield, and that this marks a significant reason in favour of their development and use. Here I offer a critical response to the position advanced by Arkin. Although I am sympathetic to the spirit of the motivation behind Arkin's project and agree that if *we decide to develop and use these machines they ought to be programmed to behave ethically, there are several major problems with his view as it stands. At present, it is not clear whether such machines would be capable of behaving more ethically than humans. More importantly, to the extent that humans would remain in the context of war, human moral transgressions will continue, especially in the face of complicated ethical challenges accompanying automated warfare. Moreover, even if machines could be more ethical than humans in certain ways and in certain situations, this says nothing about whether warfare that contains these machines would itself be overall more ethical than warfare that does not include them as participants, or whether the inclusion of lethal robots is the best way to guard against human moral transgressions in war.*

KEY WORDS: Automated warfare, autonomous lethal robotic systems, Ronald Arkin, robot ethics

Introduction

Semi-autonomous robotic weapons are already carving out a role for themselves in modern warfare. Given the billions of dollars being allocated for artificial intelligence and robotics research by military sources for military applications, if left unchallenged, there is a good chance that this trend will continue in the future, resulting in the increased development and use of more sophisticated military machines.[1] In light of this, it is important that we determine whether the development and use of such autonomous lethal robots is something that we are morally and legally permitted to do, before we (continue to) create and use these machines for this purpose.

One of the leading advocates of the development of such machines is the roboticist Ronald Arkin. In a series of recent works (2009a, 2009b, 2010) Arkin has argued that autonomous lethal robotic systems could be more ethical than humans on the battlefield, and that this marks a significant (indeed, a decisive) reason in favour of their development and use.[2] In this paper I offer a critical response to the position advanced by Arkin. Although I am largely sympathetic to the flavour of the motivation behind Arkin's project (i.e., to make war more ethical and to reduce human casualties), and agree that *if* we

Journal of Military Ethics, Vol. 11, No. 2, August 2012
ISSN 1502-7570 print/1502-7589 online/12/020149-20
 http://dx.doi.org/10.1080/15027570.2012.708265

Routledge
Taylor & Francis Group

decide to develop and use these machines then they ought to be programmed to behave ethically, there are nevertheless several problems with his view.

In what follows I examine the underlying assumptions of Arkin's view, and the case that he advances in favour of the development and use of autonomous lethal robotic systems for military purposes. I argue that, even if we allow Arkin's underlying assumptions (which there is good reason *not* to do), the case that he presents is not as strong as he suggests it to be. At present, it is not clear whether such machines would be capable of behaving more ethically than humans. More importantly, to the extent that humans would remain in the context of automated war, then human moral transgressions will continue to occur, albeit perhaps under different guises, especially in the face of other ethical challenges accompanying automated warfare (e.g., complications surrounding human use of autonomous lethal robotic systems in practice). Moreover, even if machines could be more ethical than humans in certain ways and in certain situations, this says nothing about whether warfare that contains these machines would be overall more ethical than warfare that does not include them as participants, or whether the inclusion of lethal autonomous robots is the best way to guard against human moral transgressions in war, or the inhumaneness of war in general.

By way of giving away the ending: we are not justified in *assuming* that the emergence of automated warfare is inevitable; any case for automated warfare that relies on this assumption is significantly weakened from the start. Moreover, the 'urgency' of lessening human unethical behaviour in war is exaggerated by Arkin, and there are many alternative ways to reduce such behaviour without requiring that we devote scare resources to developing lethal autonomous robotic systems. Even *if* we decide to go down the path of automating warfare, there is good reason to suggest that it will continue to be (just as) unethical, and thus would not alleviate the problem that it was intended to. The important upshot of all of this is that we need a better argument in favour of automating warfare, lest we pursue it in a morally dubious, unjustifiable manner.

Arkin's Case for Automated Warfare

The central thesis of Arkin's case in favour of automated warfare is that, the potential for improved military effectiveness aside, lethal robotic systems *will* be able to perform 'more ethically than human soldiers *are capable of* performing' (emphasis added). This is in part because they would not need to be equipped with the emotional and psychological underpinnings that lead to many of the war crimes and lesser moral violations committed by humans (2009a: 30–1). *If* robots could behave more ethically than human soldiers, argues Arkin, then we should bring such robots onto the battlefield, either *in lieu* of human soldiers (altogether), or as their partners.

At one point, Arkin makes a significantly weaker and more plausible claim: 'There is clear room for improvement, and autonomous systems may help' (2009a: 33). The important difference between this claim and the stronger claim above is the level of conviction – the stronger claim replacing 'may' with 'will'. It seems uncontentious that the current moral calibre of warfare leaves much room for improvement. And, if one goal of developing autonomous lethal robotic systems is to improve the moral calibre of warfare, then, *prima facie* at least, the inclusion of autonomous systems certainly *may* help in this regard. Yet there are many ways in which we could help to improve the current situation, and thus creating autonomous robots is but one of many potential ways to help.

Importantly, even Arkin's weaker, more passive claim depends on the strength of the reasons he advances for accepting his stronger claim (discussed below), including the empirical veracity of the claim that such machines will actually spawn an improvement in this regard. As Arkin repeatedly argues for his stronger claim, and it is the one that is more interesting and philosophically rich, it is that one which I focus on herein. I mention the weaker claim in passing in order to highlight the idea that rejecting *Arkin's* case in favour of automating warfare does not entail a rejection of *all reasons for* automating warfare. Nevertheless, as one of the leading proponents of automated warfare, dismantling his view puts the burden back on proponents to demonstrate anew that the development and use of such machines is a morally acceptable undertaking.

One important point is worth making at the outset. Although it is true that human soldiers are capable of moral transgressions, they are also capable of behaving in line with law and morality, and the great majority of human soldiers that are *capable of* immoral action do not actually behave immorally. Arkin himself (2009a: 35) notes that humans are often reluctant to kill in war, and experience psychological malaise after witnessing others behaving unethically. One implication of this is that increased attention could be placed on strengthening human moral convictions, which is a promising avenue for improving the overall ethical calibre of warfare. Thus, the psychological and dispositional foundations for improving the moral calibre of human warfare are already in place, and may be further strengthened and reinforced through deeper ethical training, for example.[3]

Moreover, humans are also capable of morally praiseworthy and supererogatory behaviour, exemplified by (for example) heroism in battle, something that machines may not be capable of. Something that typically gets overlooked in the debate surrounding automated warfare is the idea that not only will it prove to be quite challenging to actually create robots that are thoroughly sensitive to morality in their programming and behaviour, but that replacing humans with such machines may also eliminate the occurrence of soldiers 'going beyond the call of duty'. Indeed, one may challenge the pursuit of automated warfare on the grounds that it unduly threatens the ability of human soldiers to exhibit morally exceptional behaviour, and undermines important aspects of the military profession. Thus, even if machines could behave more ethically than some humans do, their ceiling for morally good behaviour may not be as high as that of human soldiers; although robots could act more ethically than some human soldiers do, they may also not act as morally excellently as certain human soldiers do. If the inclusion of autonomous robots made it so that humans have less opportunity to demonstrate bravery, or 'selfless sacrifice' in the name of the citizenry that they represent and protect, and if robots were incapable of behaving in such ways themselves, then certain morally praiseworthy aspects of *human* warfare might be lost. Thus making warfare less inhumane by removing humans from the battlefield may also take away certain desirable human-based qualities as well. One upshot of this is that, if there are ways available to improve the moral calibre of warfare that do not mitigate the occurrence of morally praiseworthy and supererogatory action (by humans), then (*ceteris paribus*) this would be good reason to pursue those avenues instead of automating warfare.

Contra Arkin, then, the important issue is not really whether robots could behave better than humans 'are capable of' behaving, but rather whether robots could perform equally as (or more) ethically than humans *actually do* on the battlefield, given that *some* humans commit war crimes and atrocities, under *some* conditions and in *certain* scenarios, *and* given that certain human soldiers perform actions of the highest moral calibre, honour, and bravery. In the face of certain moral transgressions committed by individual human

soldiers (and perhaps even robotic soldiers as well), it is prophesized that robots *as a species* will outperform humans as a species, ethically speaking.

According to Arkin, a major aspect of the increased ethicality of robotic systems would stem from their better treatment of non-combatants (e.g., innocent civilians), and their being directly programmed to adhere to the established international laws of war. Although Arkin (2009a: 30–1) admits that autonomous machines may not always behave morally flawlessly, and hence that they will still be imperfect moral entities (i.e., *capable of* immoral actions), rendering the battlefield more moral in this way is thought to be a welcome outcome of developing autonomous military machines; although war may continue to be hellish,[4] making it a more moral context is not a futile enterprise, but rather a welcome one.[5] Arkin takes his project to be a humanitarian one; if the use of lethal robotic systems in war can reduce human casualties, and collateral damage, and reduce the risk of (*certain kinds of*) human moral transgressions, then it would render the context of warfare less inhumane. In essence (and somewhat ironically), the goal is to make war less inhumane by lessening or removing the human element from warfare.[6]

According to Arkin, human dispositions like fear, greed, rage, revenge, stress, and fatigue act as catalysts for misbehaviour in battle. Since we could neglect to program machines with such emotions and psychophysical capacities without thereby sacrificing their military performance, they would not have the same predispositions towards unethical behaviour that human soldiers sometimes fall prey to. In addition to this, as noted above, their ability to behave more ethically would also stem from their programmed (inherent) adherence to the laws of war, which may prove to be more effective than the often limited indoctrination of morality experienced by human soldiers. Even when human combatants are aware of these regulations on their conduct, they have the will to disregard them if they so desire (or in case their emotions get the better of them), something that robotic systems need not be equipped with, despite their otherwise being autonomous.[7] To this extent, the *jus in bello* arm of the international laws of war may be more easily respected by robotic combatants than has traditionally been the case with (some) human combatants.

Arkin (2009a) proposes that autonomous lethal robotic systems could be programmed to possess what he terms an 'ethical governor' and a 'responsibility advisor' that would help ensure to a significant extent that robots behave ethically. Not only would the laws of war be programmed into the very makeup of the robot,[8] but it could constantly survey its own intended behaviour, and advise any human operators or commanders of the ethical consequences of its being used for certain purposes in the given context, prior to its engagement with targets. Although there may still be *some* room for error (due to malfunctioning perhaps, or unforeseen behaviours arising in novel situations), these mechanisms would go towards ensuring that the robotic system performs only those actions that are deemed to be morally permissible.

It is important to note that Arkin is concerned primarily with *autonomous* robots, rather than non-autonomous machines or the sort of semi-autonomous weapons currently in use on the battlefield. Although he recognizes that keeping a human in the operational loop may be beneficial in some cases, he also recognizes several downsides to this more limited autonomy (such as detachment and the dehumanizing effect of long distance and 'remote-controlled' military initiatives [2009a: 35]), and his language suggests that more autonomy in military robots is both desirable (ibid.: 36) and inevitable (ibid.: 29).

With this sketch of Arkin's view in place, I move now to a critical examination of it.

The Inevitability of Automated Warfare

It is important to note that Arkin is not primarily concerned with the larger context of just wars or with human behaviour outside of the context of war. Implicitly at least, he seems to assume an otherwise justifiable context for the use of autonomous lethal robotic systems. Although this starting point is understandable, given that he is not an ethicist, and that his primary concern is with making these military robots safe and moral, the soundness of his case *depends on* the veracity and tenability of these and other implicit assumptions.[9] To use a somewhat inflated example: that concentration camps are used for unjust and inhumane purposes is a crucial fact that needs to be considered prior to construction of such camps. *If* wars that involve autonomous lethal robotic systems are unjust, then ensuring that these robots are safe and effective is morally futile, and unacceptably retrospective.

In this section, however, I would like to highlight and challenge two other assumptions underlying Arkin's view. Arkin's view depends on the veracity and tenability of these assumptions; if there is good reason to reject them, then there is *a fortiori* good reason to reject the underlying reasons for Arkin's project as well.

Despite Arkin's (2009a: 1) explicit recognition that peace is preferable to war, he nevertheless offers the following:

> The trend is clear: Warfare will continue and autonomous robots will ultimately be deployed in its conduct. Given this, questions then arise regarding if and how these systems can conform as well or better than our [human] soldiers with respect to adherence to the existing Laws of War (2009a: 29).

Elsewhere, Arkin concludes:

> My personal hope would be that they will never be needed in the present or the future. But mankind's tendency toward war seems overwhelming and inevitable. At the very least, if we can reduce civilian casualties in compliance with applicable protocols of the Geneva Conventions and the ideals enshrined within the Just War tradition, the result will have constituted a significant humanitarian achievement, even while staring directly at the face of war (2010: 339).

There are two assumptions that Arkin is making here that deserve to be scrutinized:[10] (1) that (automated) warfare will *inevitably* continue in the future, and (2) that the continued development of (more) sophisticated and autonomous lethal robotic systems is itself *inevitable*.[11] For one thing, the language here is much too strong; it is not 'inevitable' that war will continue or that such robots will be developed, especially if this is meant to suggest that it is impossible that things go otherwise. The obvious counter-example here is that we simply do not devote the time and resources to developing such machines (or to waging wars), which are *active* human undertakings. To be sure, pacifists are often dismissed as being naïve and unrealistic, and Arkin's prophecy of the occurrence of (automated) warfare *may* prove to be accurate in the end. However, at this point in time, this outcome is certainly not unavoidable. Moreover, even if our drive towards automated warfare continues, this alone says nothing about whether doing so is morally acceptable. Neither the development nor the use of autonomous lethal robotic systems is inevitable. It may be difficult to curtail the development of these machines, since early versions are already in use and many more are on the drawing board, and it may seem as though these technologies are coming, whether we like it or not. Nevertheless, I am inclined to believe

that we still, at present, have a real say in whether this actually occurs. Indeed, even if we do develop autonomous lethal robotic systems, we would still have the unfettered choice *not* to use them in practice. And, we would need a very good reason to both develop and use these robotic systems, beyond the putative inevitability of their development, and the colossal waste of resources if we do not use them once they are available.

Given the rapidly changing nature of modern warfare and the complicated issues that are emerging in the face of increased technological sophistication (Singer 2010), we are presented with an opportunity for critical reflection on whether or not we want to continue to promote the occurrence of (automated) warfare at all. As the nature of warfare continues to morph, there is renewed reason to question our motivations and reasons for allowing war to occur, and to determine whether or not (automated) warfare is something that we want to continue to see happen; as Arkin himself acknowledges (2009a: 5), we have an obligation to determine whether or not we are morally permitted to pursue this project. To this I would add: *prior to actually doing so*. One hopes that if automated warfare is found to be unjust or immoral (as some thinkers have already suggested it to be[12]), outside of whether such robots could behave more ethically than some humans do, then this is something that we would not allow to come to fruition. The only way to ensure this is to ask whether warfare should continue and, if so, whether autonomous lethal robots should be allowed to be deployed in its conduct, rather than to assume that these outcomes are inevitable, no matter what we do. Although this assumption underlying Arkin's view is a common one, it does not speak to any inherent aspect of the world, humanity, or the nature of warfare, and thus needs to be *argued for* insofar as it is to serve as a foundational tenet of our assessment of the ethics of automated warfare. (One way to view my charge here is that, although Arkin recognizes the importance of this issue, he himself does not attend to it, favouring instead issues surrounding the actual development and use of such machines. Yet, by doing this, his view puts the cart before the horse, ethically speaking.) Put differently, it is not enough to recognize, in passing, that we are making the assumptions noted above; the important issue is whether we are *justified* in making these assumptions. Otherwise, we risk begging the question from the very beginning: the development and use of autonomous lethal robotic systems is something that we ought to do because it is something that we assume to be inevitable. If, however, automated warfare turns out to be avoidable, then we need to find other reasons for working towards bringing it about.[13]

I suspect that, to *everyone's detriment*, these assumptions will continue to be made, regardless of whether we are justified in doing so. I discuss them here since their being taken for granted as truthful fuels their putative 'inevitability' – serving as a self-fulfilling prophecy of sorts – and neglects to take seriously the important ethical and legal issues that may otherwise stand in their way. And, to repeat, in order to pursue automated warfare (or not) in a morally justifiable manner, we need good reasons to do so. Removing these assumptions (*pseudo*-reasons) from the foundation of the case in favour of automated warfare significantly weakens its strength and justifiability from the very beginning.

Ethical, Lethal, Autonomous Robots

Arkin (2009a, 2009b, 2010) offers six main reasons why robotic systems will be ethically better than human soldiers are capable of being: (1) with the help of (for example) the

Global Information Grid, robots could be much better than humans at processing relevant information in real time; (2) it is prophesized by some that they will be equipped with superior sensory apparatus for making battlefield observations; (3) robots would not have emotions such as anger, fear, and frustration, which can lead (some) humans to behave unethically; (4) robots need not have faulty psychological dispositions which, again, can lead (some) humans to commit immoral actions; (5) robots will act conservatively and 'self-sacrificially'; and (6) robots could actively monitor the behaviour of (human) soldiers, and report any unethical behaviour they witness. Taken together, these reasons lead Arkin to conclude that robots could behave more ethically than human soldiers are capable of performing in the context of war, at least given certain crucial technological advances in the future (e.g., the establishment of a sophisticated ability for discrimination).

Arkin's language is somewhat misleading here. As it turns out, several of these aspects of heightened robotic ability are not *ethical* in nature; not all of these reasons actually support the idea that robots could be more *ethical* than humans, as Arkin intends them to. For example, claim (2) does not obviously support this thesis, at least not directly; as discussed below, the problems associated with human behaviour cited by Arkin do not stem from a lack of the sensory abilities in human soldiers to make accurate battlefield observations. There is a moral difference between intentional and unintentional wrongdoing, and between culpable and non-culpable ignorance, which Arkin seems to conflate here. Given the conditions of modern warfare, there are cases where human soldiers act responsibly given the information they have, and yet where doing so sometimes leads to *unintentional* wrongdoing (e.g., unforeseen non-combatant collateral damage). It is a stretch to say that human soldiers behave unethically in such situations, even if having more and better information would have lessened the risk of unforeseen, unintentional wrongdoing. To be sure, such collateral damage is undesirable, and if machines have access to more accurate information, then they may be better than humans in many contexts. Yet, it is misleading to suggest that this lack of information automatically renders the human soldier's actions *unethical*. If a lack of sensory abilities results in *unintentional* 'unethical' behaviour by humans, then this is hardly demonstrative of *intentional* unethical activity (however undesirable it may be), and thus is not necessarily something that we *can* guard against (whether in humans or machines). Moreover, it is not clear whether autonomous robots could avoid such problems anytime in the foreseeable future, and thus near-term future robots may be susceptible to the same sort of deficiencies in this regard as human soldiers presumably are. Receiving a sufficient amount of information and processing it properly so as to *eliminate* the prospect of unintentional wrongdoing in its entirety is a tremendous feat indeed.

Perhaps Arkin means to suggest that robots could be better at discriminating than human soldiers are. But, the majority of human moral transgressions are not *unintentional* (e.g., representative of attempted *yet failed* discrimination, or unintentional torture, etc.) and hence are not typically problems of limited sensory abilities leading to poor discrimination. Put differently, misbehaviour stemming from *incorrect* discrimination does not typically represent *immoral* conduct by those human soldiers, although the results are undesirable; the problems that arise here typically stem from a failure or reluctance to discriminate in the first place or, at worst, the misguided belief that targeting, engaging, and killing non-combatants is morally acceptable.[14] Moreover, as Arkin himself repeatedly notes, at present robots are nowhere near as effective at discriminating as humans are. Even if robots could have superior sensory abilities, this does not *necessarily*

mean that they would be ethically superior to humans. For one thing, an argument needs to be made to support the claim that this heightened sensory ability, *if* it can be established, would be used ethically; an increased ability to discriminate effectively between combatants and civilians would only be useful to the extent that the robotic (or human) soldier used it to avoid targeting civilians in practice.

Claim (1) also seems misplaced here. It is not clear that human soldiers that violate the laws of war typically do so because of a *lack of* information (about those laws, or the nature of the context, or the status of the target they are engaging, etc.). Perhaps there is a case to be made to suggest that increased information helps soldiers to make better, ethical decisions. And yet, humans (and robots) could use their increased information for ill. Moreover, the information that a soldier needs *in order to be able to behave morally* is quite narrow (e.g., what rules are governing one's behaviour and whether or not the potential target is a legitimate combatant), and thus it seems that even those humans that behave unethically would still have access to this information, even if they choose to ignore it – lacking this information is not the source of their unethical behaviour. (That robots would not ignore this information is certainly an asset, but the point is that increased information does not automatically result in being more ethical, and much of the unethical behaviour of humans does not stem from a culpable lack of information.) Furthermore, any recognized *lack of* relevant information is typically accompanied with a default position of not engaging under such uncertainty, and thus would likely lead to conservatism rather than rash actions, and could be solved by making more information available to human soldiers. At any rate, this potential asset of autonomous robotic systems for enhanced informational capabilities is not obviously *ethically* beneficial, without significant qualification.

Lastly, claim (5) may not support Arkin's thesis either. Arkin (2009a: 31) argues that:

> Autonomous armed robotic vehicles [for example] do not need to have self-preservation as a foremost drive, if at all. They can be used in a self-sacrificing manner if needed and appropriate, without reservation by a [presumably human] commanding officer.

It is not clear how this self-sacrificial disposition would make robots more ethical than humans. It is not typically an ethical breach when a human refuses to sacrifice herself in battle or turns down a suicide mission, for example. Indeed, it is usually morally justifiable for her to do so, although it may often be a sign of heroism or bravery when she accepts such a mission, something lost on the robot that was forced to do so through its very programming.[15]

Perhaps Arkin's point here is that the drive for self-preservation often puts pressure on human soldiers in terms of information processing and decision-making, so that certain sorts of ethical transgressions may emerge when the soldier believes she is acting in ways that are necessary for her self-preservation, ways that are (or turn out to be) unethical, as in cases where she mistakes a civilian for a legitimate threat and erroneously believes her life to be in danger when it is not. The stress, fear, and anxiety that accompany the desire not to be killed in battle (the drive to survive) certainly influence the actions of human combatants, and need not similarly influence robots. Although the sort of scenario described above is unfortunate and undesirable, it is unclear whether the soldier has behaved unethically when she defends herself in that way, and thus whether *this* source of behavioural motivation is ethically problematic, at least in the general case.[16] To be sure, she has made a mistake and harmed someone that she should not have, in the name of her

self-preservation. But, at the same time, she was acting out of what she reasonably believed to be self-defence, something that is almost always morally justifiable under (what seems to be) a legitimate threat, and a routine occurrence in the context of warfare. It seems incorrect to suggest that robots could be 'more ethical' than humans in this way – as the human soldier that defends herself has not clearly acted unethically, strictly speaking – but rather that they would not be prone to the same sorts of mistakes, in the rare and unfortunate cases that self-preservation leads to such mistakes (McMahan 1994).[17] Indeed, *unintentional* breaches of the laws of war are not usually taken to be morally problematic and thus she is not found legally culpable (as in cases where a soldier unintentionally and unforeseeably harms a non-combatant), and she all along retains the right to self-defence.[18]

As discussed next, some of Arkin's other points are more plausible, and speak more directly to the ethical aspects of combatant behaviour.

A. *Emotionless drones with minds of steel*: It is true that robots would not need to have the emotions that lead (some) humans to act unethically in battle. At the same time, autonomous robots could be programmed to have certain affective abilities associated with (for example) guilt and remorse, which may make them more effective at learning from their mistakes, and empathizing with illegitimate targets, than they would otherwise be, making them moral from the other direction as well. Worth noting, however, is that this positive affective component only really bears on the morality of such robots in a retroactive manner, after an ethical breach has already occurred; the robot would have nothing to feel guilty about if it had acted morally flawlessly in the first place. Although this affective component may help to prevent further breaches on their behalf in the future, its use is futile unless the robot behaves unethically, *just like some human soldiers do*.

Human users and commanders of autonomous lethal robotic systems would still be susceptible to *their* emotions, which may affect how *they* use robots for their desired ends. Insofar as humans often commit injustices or atrocities even in the face of the threat of being punished (which is a central motivational tenet of Arkin's project as a whole), then the override procedure and ethical advisory role that Arkin suggests should be part of the robot's programming could all along be used for ill, or overruled by its (ethically fallible) human users and commanders. Thus, an important issue that emerges here is *the ethical use of robots by humans*, who would still be susceptible to the emotional, psychological, and physiological catalysts for unethical behaviour. Although the robots themselves may be inclined to behave morally, their human users and commanders would all along be at risk of misusing them; the human element that leads to inhumane behaviour may not have been removed enough. (I return to this line of thought later on.)

To be sure, robots may not commit unethical acts out of emotional shortcomings, since they would not have the emotions necessary for this to be possible. And, as noted above, Arkin is primarily concerned with *autonomous* robots, and hence the extent to which the emotions of human users or commanders could affect the behaviour of these robots may be minimal (although I take this to be an empirical issue). Indeed, he suggests the need for a built-in responsibility advisor that would need to be overruled in order for the human user/commander to succeed in their unsavoury application of the robot's arsenal. Yet, if we assume that humans are as prone to unethical behaviour as Arkin suggests, then this may be little solace and a weak obstacle deterring humans from overruling the ethical prescriptions of the machine. (If humans are not as susceptible to

unethical behaviour as Arkin suggests, then *the need for* autonomous lethal robotic systems becomes far less pronounced.)

Other psychological predispositions that Arkin identifies as being correlated with the unethical activity of (some) human soldiers can also be left out of the robot's programming (e.g., limited attention and focus, being prone to mental fatigue and boredom, etc.). Again, Arkin is correct to argue that robots would not need to possess the physical and psychological triggers for wartime atrocities and immoral behaviour. At the same time, we may want to press Arkin a little on this point. For one thing, there are different sources of unethical behaviour other than the flawed emotional or psychological dispositions of human soldiers, including system malfunctioning, vague mission goals or rules of engagement, corrupt command structures, overly strict guiding principles that prevent the soldier (human or robot) from adapting to novel situations in real time, the misuse of robots by humans, etc. Just as human soldiers are prone to certain sorts of imperfections, so too may autonomous robots be similarly flawed (at least in the early stages of their development and use). And, the degree to which human emotion and psychology is the source of the moral transgressions that occur in war is difficult to determine, given that other sources exist. Thus, even if robots could be better than humans in certain ways, the extent to which this would lessen the overall immorality of warfare is one that is open and empirical.

B. *Ethical police*: Arkin is correct to suggest that robots could serve as ethical advisors and 'moral police' on the battlefield. Indeed, one of the most appealing roles for autonomous robots in war would be to evaluate the conduct of human soldiers and the overall justness of wars in general. Interestingly, this role would only be necessary in cases where humans still participated in warfare to a significant extent; in the absence of humans in combat, or in the absence of humans being able to override the recommendations of the robot, then this role for robots may become somewhat superfluous. Thus, the move away from human warfare, as it proceeds closer and closer to human*less* warfare, would eliminate this need and consequently this reason in favour of developing and using these machines in the first place. More importantly, there is no obvious reason why such robotic moral advisors or ethical police would need to have *lethal* capabilities, or actively engage in real military initiatives *as soldiers*. The laudable end of monitoring, improving, and reprimanding unethical activity on the battlefield would not require lethal capabilities in the robot. Thus, an independent reason for making these robotic advisors and moral police capable of autonomous action *and lethal force* is required, something which Arkin does not provide.

Although Arkin provides an important insight here, namely that the inclusion of robots alongside human combatants may change the dynamics of military group behaviour for the better, owing to their role as observers and recorders of what transpires on the battlefield, the role of robotic ethical monitors should not be exaggerated. It seems likely that most instances of human unethical activity would occur without the knowledge of the machine (behind its back or in its absence). In cases where the robot is aware of the immoral activity as it is occurring, the most the robot could accomplish is to arrest the soldier(s) and aid in passing retrospective judgment. But, in such cases, *the moral transgression would already have occurred*. Not even Arkin wants to argue that robots could routinely *prevent* humans from committing war crimes and atrocities, although their presence may act as a deterrent if human soldiers are more aware of the likelihood of their getting caught and being punished thereafter. Thus, whether *human* warfare would

necessarily come to be more ethical in practice in the presence of such robots is an important, yet open, empirical issue.

The six reasons that Arkin offers to support the claim that autonomous lethal robotic systems would be more ethical than humans are capable of being are far less definitive than Arkin suggests and, because of this, his case in favour of automated warfare is significantly weakened. Many of them are not ethical in nature (strictly speaking), and thus do not speak directly to improving the moral calibre of human warfare. The remaining reasons are certainly important to consider, and yet overlook the idea that the humans that remained on the battlefield would continue to be able to behave unethically, albeit perhaps in different ways. Although these issues remain inescapably empirical, there is good reason to suggest that Arkin's case is already significantly weaker that it first appeared.

Military Effectiveness and Ethical Superiority

A related part of Arkin's argument in favour of automating warfare is that autonomous robots could be superior to humans in their military effectiveness. As discussed in this section, this relates to the ethical nature of these machines in several ways.

1. *Ethics versus military effectiveness*: Not all weapons or military techniques ought to be used just in case they are more effective than other weapons or techniques. Nuclear bombs and torture are paradigmatic examples here, as their moral repulsiveness (i.e., indiscriminate nature and inhumaneness, respectively) trumps their military effectiveness. Even if we could build autonomous robotic torturers that could render the practice of torture less unethical and inhumane, this is not a good enough reason to build machines for this unethical purpose.[19] Thus, even if it turns out that autonomous robotic systems are militarily effective, i.e., 'faster, cheaper, have better mission success, longer range, greater persistence and endurance, higher precision, faster target engagement, and are immune to chemical and biological weapons' (Arkin 2009a: 30), this does not necessarily entail that they *ought to be* developed or used. To be sure, military effectiveness matters, and being better in this regard *may* bring with it a higher moral acceptability as well (for example, the use of weapons that are more precise may be more ethical than the use of those that are less precise). And yet, if competing moral issues go unattended and unsatisfied, then military effectiveness risks being accomplished at the expense of ethical effectiveness.

2. *Discrimination and proportionality*: Just because robots could, theoretically, behave more ethically than some human soldiers do, does not mean that they will do so in practice; the technological proof will be in the pudding.[20] At the very least, achieving such ethical superiority will be a long process of trial and error in the laboratory and in the field, and will require the realization of presently only prophetical technological advancements, such as the ability for effective real-time discrimination.

Whether or not autonomous military machines could be designed to discriminate effectively and exert proper proportionality in real-world military contexts is a crucial yet open question, although designing a robot with these abilities may not be impossible in principle. Regardless, it seems uncontentious to suggest that the level of autonomy and the ability of machines to act in real-world contexts will increase sooner than our perfecting their ability to exert proper discrimination and proportionality, especially if

these latter tasks involve possessing rich mind-reading and folk psychological capabilities (among other things), which are currently only poorly understood.[21]

This is important to recognize since, until autonomous robots can accurately and reliably discriminate between legitimate and illegitimate targets (i.e., to discriminate between combatants and non-combatants, between surrendering combatants and aggressive combatants, between allies and enemies, between weapons and non-weapons, and be able to detect obvious perfidy, and [in the extreme] 'refrain from zapping a little girl who is trying to share her ice cream with the robot' [Sharkey 2008]), then they do not meet a central tenet of the international laws of war. In this case, using such machines in warfare would be legally and morally problematic, and the developers and users of indiscriminate autonomous machines are thereby morally suspect. As mentioned earlier, Arkin is not primarily concerned with this aspect of the ethics of automated warfare. My point here is that he needs to be; even *if* autonomous lethal robotic systems could be more ethical than human soldiers, if their use in warfare is unjust, then improving the moral calibre of war in this way is morally bankrupt.

Once autonomous military robots no longer depend on human operators to discriminate on their behalf and decide for them what constitutes proportionate response, they will need to be effective at doing so for themselves. If they cannot do so, then their participation in warfare is unjust. On the other hand, insofar as humans would continue to be the ones doing the discriminating and assessment of proportional force (on the robot's behalf), then *their* human fallibilities (emotions, psychophysical dispositions, and putative limited sensory abilities) in this regard remain in play, and hence this aspect of 'superior robot ethicality' becomes moot to that extent.

3. *The moral calibre of human soldiers*: As already stressed above, not *all* human soldiers behave unethically, and human soldiers *can* behave more ethically than they currently do on the battlefield. Arkin argues that it is 'unrealistic to expect normal human beings by their very nature to adhere to the Laws of Warfare when confronted with the horror of the battlefield, even when trained' (2010: 338). And yet, (a) we *do* expect this, and justifiably so – why else would we have established the international laws of war and punishments for intentional deviations from them?;[22] (b) there are other ways to work towards changing the conditions of the battlefield so that it could be more conducive to moral behaviour by human soldiers, other than introducing autonomous lethal robotic systems (e.g., lessening fatigue by having shorter tours of duty, improving therapeutic aid available on military bases, tightening up psychophysical screening of soldiers ahead of time, holding more rigorous training on the ethics of warfare and international laws of war, devising clearer rules of engagement, etc.); and (c) despite the existence of certain ethical deviants, *the great majority* of human soldiers do not behave immorally, and hence this problem that Arkin focuses on may be relatively mundane.

Indeed, the statistics that Arkin (2009a) cites about the moral transgressions of humans during warfare are not as overwhelming as he makes them out to be. If ten percent of soldiers report mistreating non-combatants (2009a: 31), then this means that ninety per cent (the vast majority) do not.[23] While it may be argued that ten percent is still a number that we should be uncomfortable with, it is unclear how much the introduction of autonomous lethal robotic systems could reduce this relatively small percentage. The inclusion of drones in certain military contexts has been shown to increase the moral calibre of those contexts (Strawser 2010). And yet, it is way too early to tell if this can be

generalized to other contexts, with other (more sophisticated and autonomous) robots, and whether the *overall* immorality of war has decreased (or increased).

Although there is certainly room here for soldiers not to report mistreatment, or to provide faulty reports (in either direction), the above suggests that the great majority of human soldiers, despite their flaws and limitations, do not intentionally mistreat non-combatants. Moreover, although *some* soldiers surveyed believe that torture should be allowed under certain circumstances (2009a: 32), it is not clear that these soldiers actually do participate in torture, even under those circumstances. (Even if some humans do resort to torture in certain contexts, it is unclear how the inclusion of autonomous lethal robotic systems would lessen the occurrence of torture; those humans that desire to torture could continue to do so in the absence of the 'moral machines'.) Thus, although some soldiers may believe that unethical behaviour is sometimes called for, it does not follow that these soldiers have actually dirtied their hands in this way. To be sure, robots would *not* be programmed to be able to torture or rape, and presumably they would not have genocidal tendencies or fascist ideologies. And yet, insofar as humans still take part in warfare, then these moral deviations would continue to be possible *by humans*, despite the presence of ethical robots.

Furthermore, Arkin cites reports that found that 83 percent of soldiers did not agree that civilians should be treated as insurgents (ibid.: 32). Even though this number is promising, it is also unclear what such treatment entailed; acting *as if* an individual is a potential threat until given a reason to believe otherwise does not in itself indicate any *unethical* behaviour; it is the *actual* harm to non-combatants without good reason that would be unethical, not the disposition towards reasonable suspicion. Even though many soldiers reported not being directly informed not to mistreat non-combatants by their superiors (ibid.), this does not mean that those soldiers believed it permissible to, *or actually did*, mistreat non-combatants. *Much more detailed empirical research is required on this important issue* (Sullins 2010). At most, these data may support Arkin's weaker claim noted in Section 1, i.e., that the inclusion of military machines *may* improve things, in certain ways. However, there is not enough here to support Arkin's stronger claim, i.e., that automated warfare would be less immoral than human warfare.

All of this suggests that this putative problem that robots are designed to remedy (i.e., the pervasive unethicality of human soldiers) may not be the calibre of problem that Arkin makes it out to be. The question remains whether or not it is of the level of urgency and significance so as to *require* that we create autonomous lethal robotic systems with the goal of lessening it. The empirical findings that Arkin cites to support his claim of the pervasiveness of the immorality of human soldiers, and to motivate the need for the development and use of autonomous *and* lethal robotic systems, do not seem as worrisome as Arkin suggests them to be. At the very least, the empirical data that Arkin draws upon are open to different and competing interpretations, casting some doubt on their definitiveness. More importantly, the putative abundance of unethical behaviour by human soldiers *alone* does not entail that bringing machines into the battlefield is *the* (or the most appropriate) answer to this problem, even if it is a problem that needs to be attended to, since human unethical behaviour would continue even if robots were present on the battlefield.

Arkin (2009a: 32) also cites findings that suggest that human soldiers sometimes feel unprepared for the situations that they are confronted with in war. Morally salient situations that soldiers are not adequately prepared for is certainly a problem, although it

will always be a problem, for both humans and machines, especially as novel and complicated situations arise and as modern warfare changes, and (if) autonomous robots and other futuristic technologies emerge on the battlefield.

Moreover, it is not clear whether machines could perform better in such novel, often morally indeterminate situations. The machine could be programmed to follow the laws of war, at least in principle. However, insofar as novel situations will continue to arise, it is unclear at present whether the robot itself could be adequately prepared (not to suggest that it necessarily would not be); it may be just as (un)prepared as human soldiers report that they are. Presumably, the robot would choose the action that its programming dictates as being the most moral, given the possible alternatives available, in some cases leading to inaction. And yet there would still be room for error here. Moreover, this strategy seems dangerously similar to the fallible 'do the best you can' strategy that human soldiers would predictably adopt in similar cases, something that Arkin wants to guard against. Thus, the introduction of autonomous lethal robotic systems may not remedy this problem that it is intended to remedy.

On the one hand, the unethical activity of human soldiers in the context of war – itself often acknowledged as being an inherently unethical realm of human activity (Walzer 2000, Hawk 2009, McMahan 2009) – seems exaggerated by Arkin. On the other hand, even if the unethical behaviour of some human soldiers is a serious enough problem to deserve sustained and widespread attention, it is not clear that the introduction of autonomous lethal robotic systems would remedy it, or is the most appropriate (ethical, effective, cost efficient, just) way to do so, even if it could. As noted above, alternative strategies are certainly available, including improved moral and legal education of military personnel, the establishment of clearer rules of engagement and international laws of war, more effective deterrents from devious behaviour, increased monitoring of the activities of soldiers (by robots, perhaps), more severe punishments for moral transgressions, and a reduction of the triggers of immoral human behaviour (i.e., stress, fatigue, boredom, confusion, de-sensitization, detachment, anxiety), etc. All of these are reasonable and open options, and may be more cost-efficient than creating and maintaining autonomous lethal robotic systems. Thus, even granting the humanitarian orientation of Arkin's project, he needs to offer independent reasons to justify why he is proposing to develop autonomous lethal robotic systems for the purpose of killing humans, instead of devoting the necessary energy and resources to making human soldiers themselves more ethical. The latter (parsimonious) option would also serve to make warfare less unethical and less inhumane.

A Paradox of Automated Warfare

In light of what has been said above, a paradox of automated warfare begins to emerge. Following Arkin, the goal is to create and use autonomous lethal robotic systems for military purposes. The motivation for this is that robots are expected to be ethically (and militarily) superior to (some) human soldiers, and that automated warfare could presumably lessen the number of human casualties in war.

There are a number of different ways in which automated warfare could manifest itself. Three general manifestations include: (1) (*predominantly*) *human warfare*, which involves the minimal (or negligible) use of autonomous and semi-autonomous lethal robotic systems by humans in war – the great majority of contemporary military conflicts

fall into this category; (2) *human/automated warfare*, which involves a substantial use of autonomous and semi-autonomous lethal robotic systems by humans in war, with a significant number of lethal robotic systems working with or alongside human combatants; and, at the other end of the extreme from human warfare *tout court*, (3) *fully automated (humanless) warfare*, where, outside of the presence of human non-combatants, direct human interaction in combat is negligible, on either or both sides of the conflict.

With respect to (1), it is difficult to see how the inclusion of machines to this minimal extent could have a noticeable and significant impact on improving the moral calibre of warfare overall. In this context, the great majority of combatants are human, and thus come complete with their human fallibilities. Although there has been some research demonstrating that certain kinds of machines already in use have had some impact in this vein, it is unclear at present whether or not such wars are more or less ethical in other ways as well. For Arkin's predictions to hold, and for his project to be worthwhile and well-founded, we would likely need to move to the next stage of automated warfare.

And yet, it is with respect to (2) that *all* of the worries mounted earlier, and other worries surrounding the justness and legality of automated warfare, become most salient. Putting aside the questionable assumptions that I have argued underlie Arkin's rationale, human/automated warfare, of the sort considered throughout this paper, would not clearly be any less unethical (or inhumane) than predominantly human warfare. Even if certain sorts of machines could increase the morality of war in certain ways, other machines may equally detract from this, all in the face of humans continuing to behave immorally on the battlefield. If the worries mounted earlier cannot be overcome, then there is very good reason to resist getting to the stage where this manifestation of automated warfare becomes commonplace. Perhaps, however, removing humans from the context of war *entirely* would be a viable solution.

I see no reason to support this claim, however. Fully automated warfare (i.e., humanless warfare) is undesirable and unproductive, as it would eliminate 'the severity of war', in the sense that human soldiers and the states they represent would no longer have something inherently valuable at stake (i.e., their lives). Indeed, in order for war to make sense, humans will need to remain on the battlefield and be at risk of enemy fire or capture; automated warfare *tout court* is absurd, and hence in order for automated warfare to work, it requires the involvement of humans and human risk in some direct capacity (Krishnan 2009). To be sure, fully automated warfare may be more ethical and humane than the other two manifestations of automated warfare. And yet, machines fighting machines would systematically exclude those things that we are fighting for, and could continue *indefinitely*, given the will and available resources to do so. If such a war came to an end, what would have been settled thereby, or gained by the victor and lost by the defeated? It seems that, *assuming that wars will continue to be waged and fought*, humans will need to continue to be involved in war, whether we like it or not, either alongside robotic systems or as their users and commanders (or both), since wars that do not involve humans and (legitimate) human casualties, and where neither side has anything substantial at stake, cannot be won or lost. In the advent of fully automated warfare, once the machines had defeated the machines of the opposing force, humans would then need to move into the battlefield to settle the dispute themselves.

Yet, insofar as humans are still directly involved in the context of automated warfare (options 1 and 2 above), then these humans would still be at risk of exhibiting the sorts of immoral behaviour identified by Arkin – especially if we devote our attention exclusively

to developing autonomous military machines, rather than improving the situation for human soldiers – and hence introducing robots to fight alongside them may not ultimately reduce *human* moral transgressions, or the *overall* immorality of war.[24] Thus, even *if* Arkin is correct to suggest that robots could be more ethical than (some) human soldiers, the presence of autonomous lethal robotic systems in battle may not necessarily reduce the *overall* unethical behaviour performed during war.

It seems that there will need to be a cap on the extent to which autonomous lethal robotic systems should be used, especially if not doing so would mean taking humans out of the context of war altogether. But this means that automated warfare will continue to include human soldiers (if it does continue), complete with all of their emotional, psychophysical, and ethical flaws going more or less uncontested. Doing so entails continuing to sacrifice *many* human lives in battle (both combatants and, collaterally, non-combatants alike), and continuing to endure human moral transgressions, all *in addition to* the ethical challenges that accompany automated warfare (e.g., responsibility for the behaviour of robots, the fairness of remote engagement, etc.). Not only would automated warfare likely continue to witness the sort of moral transgressions already at play in contemporary human warfare, but it also opens up other avenues for unethical behaviour, by both humans and robots, which may actually exacerbate the problem their development and use was intended to remedy.

Conclusion

There may be good reasons to create and use autonomous lethal robotic systems for military purposes. But, the elimination of human unethical behaviour and the overall moral betterment of war are not them, since, as I have argued, introducing these machines into the context of war would in all likelihood not reap these tidings.

In the foregoing I have challenged the assumptions and arguments that underlie Ronald Arkin's view that the development and use of autonomous lethal robotic systems for military purposes is something that we ought to do. Despite certain merits of his view, there are a number of problems with his account that need to be fixed if it is to continue to be accepted. The burden of proof remains on thinkers like Arkin to provide sound argument to show that automated warfare (in any of its three general manifestations) is inevitable. Moreover, even *if* autonomous lethal robotic systems could be more ethical than some human soldiers in certain ways and certain situations, this is insufficient reason for developing and using these machines. And, even if these robots could be more militarily effective, this alone does not entail that they ought to be developed and used for military purposes. Thus, independent motivation is needed, which Arkin does not currently have on offer. In light of this, I take Arkin's case in favour of automating warfare to be a weak one, and, by association, any practice of developing and using autonomous lethal robotic systems that relies on it unconvincing.

NOTES

1. As argued for below, the likelihood of this happening should not be mistaken for the *inevitability* of this occurring.
2. Arkin is not alone in holding this belief. See for example Lin et al. (2008).

3. Something like this is already taking place in Canada. See Canadian Government, Department of National Defence (2010).
4. Walzer (2000).
5. As discussed below, there may be other ways to make war more moral, even aside from the obvious strategy of eliminating war altogether, which would effectively eliminate the potential for *war* crimes in its entirety.
6. *One* way to challenge Arkin's admission here is to note that Arkin is 'criticizing' human soldiers for possessing a similar fallibility that he concedes would be present in machines as well. This raises the question of why we should be including morally flawed (robotic) soldiers as a replacement for other morally flawed (human) soldiers. Indeed, if there are ways to hone in on what makes the *morally solid* human soldiers morally solid, then we may be better positioned to replace morally flawed human soldiers with morally solid *human* soldiers, instead of with morally fallible robotic soldiers.
7. Of course, this will depend on the degree of autonomy that the robot has reached. At a certain level of sophistication, it may turn out that robots would need to be able to disregard orders, in order to meet the requirements for full moral agency. This is a complicated issue, discussion of which is beyond the scope of this paper. See Allen et al. (2000) and Tonkens (2009).
8. Interestingly, because such robots would be programmed to follow the international laws of war, we need to ask – as Sparrow (2007) does – whether their development and use is consistent with those laws in the first place. Indeed, to the best of my knowledge, Sparrow's charges that the use of autonomous 'killer robots' would violate international humanitarian law, and create a morally dubious vacuum of responsibility, remain unanswered by proponents of automated warfare. Without doing so, however, we risk building machines whose very existence goes against the moral and legal codes they are designed to abide by.
9. For a nice discussion on whether automated warfare is consistent with the tenets of received just war theory, see Asaro (2008). See also Sparrow (2007) for a closely related discussion.
10. Arkin is not alone in making these assumptions. See for example Krishnan (2009: 117) and Sullins (2010: 264, 274).
11. Elsewhere, Arkin (2009a: 56) is ambiguous on this point: 'In any case, there is no doubt at the very least that we should *proceed with caution* as these systems *inevitably* move forward toward military use. We must ensure that lethal autonomous systems, *when and if* they are introduced into routine battlefield operations, behave in a manner consistent with international law' (emphasis added). Proceeding with caution suggests that we are (someone is) in control of this enterprise, and thus we can decide the extent to which we pursue this particular path. Yet, if moving in this direction is 'inevitable', then this control may be somewhat illusory. From Arkin's writings as a whole, I believe he is working from the stronger assumptions already noted. If he is *not*, however, then the strength of his underlying motivation comes even further into question; it is only because many people agree that (automated) warfare is inevitable that proponents of automated warfare have been able to get away with *not asking whether we should pursue automated warfare in the first place*.
12. See for example Asaro (2008), Sharkey (2008), and Sparrow (2007).
13. Even if war is avoidable, making it less immoral when it does occur would *prima facie* be a morally desirable task. And yet, without independent reasons for devoting our time

and resources to making *avoidable* wars less immoral (rather than, say, putting those resources into public health and education sectors), we cloak many ethical issues that need to be attended to. See Tonkens (2012).

14. See Slim (2008), especially Chapters 4 and 5.
15. Other issues arise here as well. For example, it may not make much sense to say that a machine could act in a self-sacrificial manner, since it may have no real self to speak of. Moreover, to say that it was acting out of an interest in protecting its fellow (human) soldiers or 'its country' seems to over-extend the meaning of self-sacrifice, as the robot was *merely* following orders. See Vallor (forthcoming) for an interesting discussion in this vein.
16. This example is one of non-culpable ignorance, based on objectively false but subjectively valid belief (McMahan 2009). Similar, less unproblematic cases could also emerge, for example cases of mistaken belief accompanying acts of indiscriminate recklessness. 'Firing at anything with a heartbeat' under some conditions would constitute reckless and negligent behaviour, although in others it may be a reasonable (excusable) response to highly uncertain levels of threat. Thank you to an anonymous reviewer for this point.
17. Elsewhere I have argued that, if machines were to gain a sufficient degree of autonomy (and moral agency in general) then it may be problematic to treat them as if they did not deserve certain rights (e.g., the right to self-preservation). See Tonkens (2011).
18. We can think of cases where, in the heat of battle, human soldiers may be prone to engaging with 'anything that moves'. As robots could be programmed to have a 'higher' standard of assessment and threshold for engagement, in essence championing accuracy and legitimacy over its own personal safety, then such robots would likely not 'fire at anything with a heartbeat'. And yet, it is difficult to imagine a case where a human soldier that opens fire in this way would be behaving *unethically*, assuming that she is being engaged with, and thus has reason to believe that her life is being threatened.
19. For those who disagree that torture is unethical, please see Bufacchi and Arrigo (2006). See also Shunzo Majima's article in this issue of *Journal of Military Ethics*.
20. See for example Sullins (2010: 269), who argues that 'it is unlikely that semiautonomous tele-robotically operated weapons [for example] are enhancing battlefield ethics'. For an opposing view, see Strawser (2010).
21. See Guarini and Bello (2011).
22. One alternative here would be to ground the international laws of war on foundations that humans are capable of adhering to; 'ought implies can', and hence if human soldiers *cannot* abide by the international laws of war (as Arkin suggests), then we may be demanding too much of them, morally speaking. In other words, if acting consistently with the laws of war is practically unrealistic, then (this author is afraid to say) perhaps they ought to be modified accordingly, rather than replacing those who cannot abide by them with autonomous lethal robotic systems.
23. I am somewhat inclined to agree with Arkin that ten percent is a large number when we are talking about the number of human soldiers that behave unethically in the context of war. Two places where we disagree, however, are (1) that this number is high enough to demand serious attention and allocation of an abundance of scarce resources, especially given the fact that most wars are inherently unjust (McMahan 2009), and other sectors of human activity could greatly benefit from such attention and resources (e.g., healthcare, education, welfare); and (2) even *if* this problem deserves our attention and resources,

that creating and using autonomous lethal robotic systems is the *best* way to lower said percentage, given due consideration to *all* the available alternatives.

24. Worth noting is that it may become increasingly unfair to include humans in automated warfare, where their opponents are robots, as it would become exponentially more difficult for them to reign victorious over such opponents, and the robots are not putting similar things on the line (e.g., their lives).

REFERENCES

Allen, Colin, Gary Varner & Jason Zinser (2000) Prolegomena to any Future Artificial Moral Agent, *Journal of Experimental and Theoretical Artificial Intelligence*, 12(3), pp. 251–261.

Arkin, Ronald (2009a) *Governing Lethal Behavior in Autonomous Robots* (Dordrecht: Chapman & Hall).

Arkin, Ronald (2009b) Ethical Robots in Warfare, *IEEE Technology and Society Magazine*, Spring, pp. 30–33.

Arkin, Ronald (2010) The Case for Ethical Autonomy in Unmanned Systems, *Journal of Military Ethics*, 9(4), pp. 332–341.

Asaro, Peter (2008) How Just Could a Robot War Be? in: P. Brey, A. Briggle & K. Waelbers (Eds), *Current Issues in Computing and Philosophy*, pp. 50–64 (Amsterdam: IOS Press).

Bufacchi, Vittorio & Jean Maria Arrigo (2006) Torture, Terrorism, and the State: A Refutation of the Ticking-bomb Argument, *Journal of Applied Philosophy*, 23(3), pp. 355–373.

Hawk, W. J. (2009) Pacifism: Reclaiming the moral presumption, in: H. LaFollette (Ed), *Ethics in practice*, (3rd edition), pp. 735–745 (Oxford: Blackwell).

Government of Canada, Department of National Defence, *Defence Ethics Programme* (2010). http://www.dep.forces.gc.ca/dep-ped/index-eng.aspx. Accessed September 10, 2011.

Guarini, Marcello & Paul Bello (2011) Robotic Warfare: Some Challenges in Moving from Non-civilian to Civilian Theaters, in: Patrick Lin, George Bekey & Keith Abney (Eds), *Robot Ethics: The Ethical and Social Implications of Robotics*, pp. 129–144 (Cambridge: MIT Press).

Krishnan, Armin (2009) *Killer Robots* (Dordrecht: Ashgate).

Lin, Patrick, George Bekey & Keith Abney (2008) Autonomous Military Robotics: Risk, Ethics, and Design. Funded by US Department of Defense/Office of Naval Research; accessed, available at: http://ethics.calpoly.edu/ONR_report.pdf; Internet. Accessed November 11, 2011.

McMahan, Jeff (2009) *Killing in War* (Oxford: Oxford University Press).

Sharkey, Noel (2008) Cassandra or False Prophet of Doom: AI Robots and War, *IEEE Intelligent Systems*, 23(4), pp. 14–17.

Singer, Peter W. (2010) *Wired for War: The Robotics Revolution and Conflict in the 21st Century* (New York: Penguin).

Slim, Hugo (2008) *Killing Civilians: Method, Madness, and Morality in War* (New York: Columbia University Press).

Sparrow, Robert (2007) Killer Robots, *Journal of Applied Philosophy*, 24(1), pp. 62–77.

Strawser, Bradley J. (2010) Moral predators: The Duty to Employ Uninhabited Aerial Vehicles, *Journal of Military Ethics*, 9(4), pp. 342–268.

Sullins, John (2010) RoboWarfare: Can Robots Be More Ethical than Humans on the Battlefield? *Ethics and Information Technology*, 12, pp. 263–275.

Tonkens, Ryan (2009) A Challenge for Machine Ethics, *Minds and Machines*, 19(3), pp. 421–438.

Tonkens, Ryan (2011) Out of Character: On the Creation of Virtuous Machines, *Ethics and Information Technology*, 14(2), pp. 137–149.

Tonkens, Ryan (2012) Should Autonomous Robots be Pacifists? *Ethics and Information Technology*, Online first May 16, 2012. DOI: 10.1007/s10676-012-9292-z

Walzer, Michael (2000) *Just and Unjust War: A Moral Argument with Historical Illustrations*, 3rd edition. (New York: Basic Books).

Ryan Tonkens earned his MA in Philosophy from the University of Windsor in 2006, and a PhD in Philosophy from York University (Toronto) in 2012. His PhD dissertation is an analysis of the ethics of human prenatal genetic alteration, from a character-based perspective. He is currently a postdoctoral research fellow at Novel Tech Ethics, Dalhousie University, where he is conducting research in reproductive ethics and the ethics of parenthood. His primary research interests are in applied ethics, especially ethical issues that accompany advances in biotechnology and artificial intelligence. He has published his work in the *Journal of Medical Ethics*, the *Journal of Medicine and Philosophy*, the *Journal of Applied Philosophy*, *Minds and Machines*, and *Ethics and Information Technology*. *Correspondence address*: Novel Tech Ethics, University of Dalhousie, Halifax, NS B3H 4R2, Canada. E-mail: tonkens@yorku.ca