The Ethics of Acquiring Disruptive Technologies: Artificial Intelligence, Autonomous Weapon and Decisions Support Systems

C. Anthony Pfaff

Last spring, Google announced that it would not partner with the Department of Defense's "Project Maven," which sought to harness the power of artificial intelligence (AI) to improve intelligence collection and targeting. Google's corporate culture, which one employee characterized as "don't be evil," attracted a number of employees who were opposed to any arrangement where their research would be applied to military and surveillance applications. As a result, Google had to choose between keeping these talented and skilled employees and losing potentially hundreds of millions of dollars in defense contracts. Google chose the former.[1] In fact, a number of AI-related organizations and researchers have signed a "Lethal Autonomous Weapons Pledge" that expressly prohibits development of machines that can decide to take a human life.[2]

This kind of problem is not going to go away. Setting aside whether the kind of absolute pacifism exhibited by Google employees is morally preferable to its alternatives, it is worth taking these concerns seriously. Persons who enjoy the security a state provides are well within their rights to make personal commitments to avoiding violence of any kind. As Stanley Hauerwas puts it, only a complete commitment to non-violence, even if not entirely philosophically consistent, is often the only way to convince a society not just to consider, but privilege, non-violent approaches to conflict resolution over violent ones.[3]

Another way, however, of understanding the Google employees' objection is not that all military research is evil, but that development and use of AI weapon systems are *mala en se,* which means their use, in any context, constitutes a moral wrong. If true, then such weapons would fall into the same category as chemical and biological weapons, whose use is banned by international law. If these weapons do fit into that category then the U.S. government's only morally appropriate response would be to work to establish an international ban rather than develop them.

The difficulty here is that no one really knows if these weapons are inherently evil. Objections to their use tend to cluster around the themes that such weapons dehumanize warfare and introduce a "responsibility gap" that could undermine International Humanitarian Law. Moreover, even if one resolved these concerns, the application of such systems risks moral hazards associated with lowering the threshold to war, desensitizing soldiers to violence, and incentivizing a misguided trust in the machine that abdicates human responsibility. At the same time, however, proponents of such systems correctly point out that they are not only typically more precise than their human counterparts, they do not suffer from emotions such as anger, revenge, frustration, and others that give rise to war crimes.

Meanwhile, as the debate rages, adversaries of the United States who do not have these ethical concerns continue with their development. China, for example, has vowed to be the leader in AI by 2030.[4] No one should have any illusions that they will use this dominance for military as well as civilian purposes. So to maintain parity, if not advantage, the Department of Defense (DoD) has little choice but to proceed with development and employment of artificially intelligent systems. As it does so, ethical concerns will continue to arise, potentially excluding important expertise for their development. To include this expertise, the DoD needs to confront these ethical concerns upfront.

Part of the problem in addressing these concerns is that the technologies in question are "disruptive" in ways that more conventional technologies are not. The term "disruptive technology" typically applies to technologies that not only replace older ones but in doing so change how actors in any particular environment compete.[5] In this sense, these technologies change the "rules" that guide competitive behavior. For example, the development of the internet changed how people obtained news and information, forcing the closure of more traditional media, which struggled to find ways to generate revenue under these new conditions.

Changing the rules guiding competition, however, does not mean it is immediately obvious what the new rules are. This point holds true for both the practical norms that determine success as well as the ethical ones that ensure the legitimacy of that success. While practical norms will emerge through the application of these systems, the ethical norms need to be established beforehand to ensure the application a new technology conforms to one's prior moral commitments. It is this requirement that distinguishes acquiring technologies that rely on artificial intelligence from more conventional weapons. Next generation tanks, artillery, and aircraft, for example, not only have a technological and industrial base sufficient to keep up, if not surpass, adversaries, the rules associated with their use are, for the most part, well-understood. For example, norms prohibiting harms to noncombatants drive efforts to make these weapons more precise. That is not the case with the technologies under consideration here.

It is, of course, impossible to fully address the concerns of a committed pacifist when developing lethal systems. However, one can take on the question regarding whether military AI-systems are *mala en se* as well as determine measures for the moral hazards this new technology may present.

For the purposes of this discussion, the term AI-systems will refer to military AI-systems that may be involved in "life-and-death" decisions. These systems include both lethal autonomous weapons systems (LAWS) that can select and engage targets without human intervention and decision support systems (DSS) that facilitate complex decision-making processes, such as operational and logistics planning. After a brief discussion of military applications of AI, I will take on the question whether these systems are *mala en se* and argue that the objections described above are insufficient to establish that they are inherently evil. Still, this is new and disruptive technology that gives rise to moral hazards that are unique to its employment. I will take up that concern and discuss measures DoD can take to mitigate these hazards so that the employment of these systems conforms to our moral commitments.

Military Applications of Artificial Intelligence

Military applications of artificial intelligence pose ethical challenges because, whatever the activity these systems partake in, the nature of warfare entails they will almost certainly play some role in making decisions that will affect the lives and well-being of humans.[6] Examples include the Phalanx ship-board air defense system, which performs functions associated with searching, detecting, evaluating, tracking, engaging, and destroying targets[7] or the Course of Action Display and Evaluation Tool, which produces detailed, actionable battle plans significantly faster than humans.[8]

While these are both simple examples of autonomous systems, the technology will continue to develop in both complexity and capability. As it does so, humans will likely become more dependent on these machines for making such life-and-death decisions. When they do so, humans will start to cede responsibility for those decisions in a way that can make accountability for failure impossible. Life-and-death decisions, precisely because they affect the well-being of other persons, are moral decisions. Thus, to maintain moral accountability, they should be made by moral agents capable of assuming responsibility for any failures. Otherwise, one risks setting conditions for war crimes or other moral failures for which no one is at fault.

Of course, different systems pose different concerns. The Phalanx system is more automated than autonomous and typically functions in environments where the likelihood of collateral harm is low. Armed Unmanned Aerial Vehicles such as the MQ-1 Predator or the more advanced MQ-9 Reaper are probably the best known examples of LAWs; however, these systems are not autonomous in the relevant sense. While they can stay aloft far beyond the scope of human endurance, the cognitive load on the humans who operate it from the ground is not much different than if they were on board. Humans still pilot the aircraft and operate the on-board sensors.[9]

Systems that raise even greater concerns would be the so far fictitious drone swarms portrayed in the anti-autonomous weapons video, *Slaughterbots*, where miniature drones are able to discriminate among human targets based on age, sex, fitness, uniform, and ethnicity. As a result, they are capable of hyper-precision, which arguably is a moral good: the ability to perfectly discriminate between legitimate and illegitimate targets is a moral as much as it is a practical pursuit and *should be* part of any weapon acquisition program. Many innocent lives would have been saved, for example, if the U.S. military could have deployed a swarm of drones into Mosul that would only attack persons who fit the profile of a Daesh member. In the film clip, however, the weapons fall into the wrong hands and are used by terrorists to assassinate individuals ranging from specific members of the U.S. congress to ideologically opposed political activists.[10] While the scenario is fictitious, swarm and sensor technology is not.[11]

While the threat in the video is certainly overstated,[12] it does portray the central moral concern driving opposition to the employment of autonomous weapons. Setting aside the possibility that terrorists will acquire such systems, which is a concern for any technology, the idea that we can deploy machines that can, on their own, discriminate between intended and unintended targets without further input from a human, is, or at least should be, concerning.

Mala En Se: Autonomous Systems and the Morality of War

It may not seem obvious at first why ceding some autonomy regarding life-and-death decisions should raise concerns. AI technology not only has the potential to decrease risk to combatants, but non-combatants as well. The result could be a less lethal form of warfare that allows for better conformity to the law of war and less killing overall. If that is the case, what does it matter if violations that do occur go unpunished? This point is, in fact, a main objection of proponents of AI-systems. If the employment of such systems entails more good than harm, it makes sense to put up with the harm. The difficulty with that response is that it reduces moral judgment to utilitarian concerns where the ends justify the means. Since the ends do not, in fact, always justify the means, this objection is insufficient to resolve the moral concerns associated with the employment of AI-systems.[13]

The War Convention

To answer that question and determine if such systems are *mala en se,* one must first account for the evil such systems will commit. Most discussions of ethics in war begin with Just War Theory as well the legal norms of International Humanitarian Law to which it gave rise. However, as Michael Walzer observes, the law does not exhaust the convention. Rather, he argues that the practice of war is accompanied by moral argument regarding specific acts of war and over time has resulted in a complex set of "norms, customs, professional codes, religious and philosophical principles, and reciprocal arrangements that shape our judgments of military conduct." He refers to that complex set as "the war convention."[14]

However, the War Convention, as Walzer conceives it, primarily governs relationships between enemies. In this context, norms associated with necessity, discrimination, proportionality, and avoidance of unnecessary suffering govern the use of force and any system the military employs should conform to them.[15] For example, any weapon that causes suffering or harm that is unnecessary to defeating the enemy, such as

bullets poisoned to create infections that linger after the wounded soldier has been taken off the battlefield, would be impermissible.[16] Similarly, weapons that are insufficiently discriminate or proportionate, such as biological weapons, are also banned.

Walzer's conception of the war convention says very little regarding obligations civilian leaders and military commanders have to their own troops. That is not to say those norms do not exist. Professional codes, as well as the leadership curriculum in every level of military education, specify the duties leaders have to those they lead. Space does not exist to fully articulate what those obligations are; however, Walzer recognizes that military commanders are expected to do what is necessary to achieve legitimate military objectives with the least cost in friendly lives and resources, while upholding International Humanitarian Law.[17] To the extent deployment of AI systems impact commanders' ability to fulfill those obligations, then those obligations should count as norms of warfighting as well and assimilated into the war convention. This broader notion of the war convention, which includes obligations to friends and enemies, will apply to rest of this discussion. What I will discuss next is what happens when autonomous systems share some of the responsibility of warfighting.

Responsibility Gap

Advocates of AI-systems often make the point that these system's capabilities enable better ethical behavior than human beings, especially in combat. Ronald G. Arkin, in his book *Governing Lethal Behavior in Autonomous Robots*, argues that not only do such systems often have greater situational awareness, they are also not motivated by self-preservation, fear, anger, revenge, or misplaced loyalty suggesting that they will be less likely to violate the war convention than their human counterparts.[18] Improving ethical outcomes, however, is only part of the problem. While machines may perform ethically better—in certain conditions at least—than humans, they still make mistakes. Separately, depending on the complexity of the system, it may not always be possible to tell why it behaved the way it did.

For example, in June 2007, the first three "Talon" Special Weapons Observation Reconnaissance Detection Systems, which are remotely controlled robots that can aim at targets automatically, were deployed to Iraq but reportedly never used because their guns moved when they should not have.[19] In October 2007, a semi-autonomous robotic cannon employed by the South African military killed nine soldiers and wounded 14 others, though no one is certain whether it was a hardware or software malfunction.[20] As Paul Scharre and Michael Horowitz observe, at least some of the information AI systems use to determine their responses is often encoded in the strength of connections of their neural networks and not as code that human operators can analyze. As a result, machine thinking can be something of a "black box."[21]

So, even if one is able to design machines that *can* function ethically in battle, they can still make mistakes. Moreover, those mistakes may occur even when the machine is operating properly. Whom, then, should be held responsible for such mistakes? The soldier who operated the machine, the manufacturer who designed the machine, or the acquisition officer who set the system requirements?

Wendell Wallach and Colin Allen in their book, *Moral Machines, Teaching Robots Right from Wrong,* echo this concern. As they point out, "As either the environment becomes more complex or the internal processing of the computational system requires the management of a wide array of variables, the designers and engineers who built the system may no longer be able to predict the many circumstances the system will encounter or the manner in which it will process new information."[22] If it is not possible to fully account for machine behavior in terms of decisions by human beings, then it is possible to have an ethical violation for which no one is responsible. Thus the inability to fully account for machine behavior introduces a "responsibility gap," which threatens to undermine the application of the war convention and dehumanize warfighting.[23]

This concern applies to DSS as much as it does LAWs. The more commanders rely on and come to trust the output of DSS, the more difficult it can be to question that output. Eventually, DSS's could come to exert more control over the decision-making process and thus increase the responsibility gap beyond what is morally acceptable.[24]

In Iraq, for example, some units employed a DSS to select the safest route for convoys to travel. It made this recommendation based on attack and other incident data it received. In one instance, the machine recommended a route in which a convoy suffered an attack where U.S. soldiers' lives were lost. As it turned out, the recommended route had previously been categorized as one of the most dangerous routes in the area of operations. Because it was one of the more dangerous routes, convoys quit travelling on it so over time it appeared from the perspective of the DSS as one of the safer routes since there had not been any recorded attacks.[25] The question then is, who or what is responsible for choosing the riskier route: the machine or the human who approved the route based on the machine's input? This is, of course, a simple example from a very rudimentary DSS. **The point here is as machines get more complex, determining how to trust the information they provide will be increasingly problematic.**

To the extent AI-systems cannot be morally responsible for the harms they cause, Hin-Yan Liu views the application of these systems as the moral equivalent of employing child soldiers who are also not morally responsible for the harms they commit. In this case, Liu is specifically addressing LAWS though the same concerns could apply to DSS. Rather, he argues, since international law criminalizes the introduction of children on the battlefield, regardless of how they behave, the crime for those who employ them is not simply that they are victimizing children, but that they are also introducing "irresponsible entities" on the battlefield. Since AI-systems are also "irresponsible" in the relevant sense, they too should be banned and those who do introduce them subject to criminal penalty.[26]

Since AI-systems cannot be coerced this objection does not seem to apply to them. That point shifts the focus of concern onto AI-system behavior. In this context, it is worth asking, if children did not enjoy this special status, would their behavior then matter? The answer is arguably yes. Moreover, if children turned out to be more humane than adults, would that be a reason to employ them in place of adults? Assuming children have no special rights, again, the answer is arguably yes. The point here, obviously, is not that children should be recruited as soldiers. Forcing children to fight is exploitive. However, the fact that AI-systems are not coerced and often more humane than human counterparts suggest the analogy to child soldiers is inadequate to justify a ban on AI-systems. If the use of AI-systems diminishes the overall horrors of war, it makes sense to tolerate the reduced number of violations they commit.

This last point, however, ignores how norms function in governing human behavior. Norms, whatever form they come in, moral, legal, or practical, are the means by which we communicate to others what we hold them, and ourselves, accountable. However, when norms are not upheld, they die.[27] Consider a work-place environment where there is a norm, for example, to show up on time. If workers instead habitually show up late and are not held accountable, then they will likely continue to do so and others are likely to follow. Eventually, the norm to show up on time will cease to be a norm.

The employment of AI-systems risks the same fate to International Humanitarian Law. The fact that LAWS and DSS can absolve humans of accountability for at least some violations will establish an incentive to employ the machines more often and find ways to blame them when something goes wrong, even *when* a human is actually responsible. It is not hard to imagine that over time there would be sufficient unaccountable violations that the rules themselves are rarely applied, even to humans. This point suggests that it will be insufficient to defend the use of AI-systems simply because they can be necessary, proportionate, discriminate, and avoid unnecessary suffering if their use threatens to undermine the rules themselves.[28]

Banning AI systems for this reason, however, raises its own concerns. As Geoffrey S. Corn argues, while humanitarian constraints on the conduct of war are a "noble goal," they do not exhaust the war convention, which permits states to defend themselves and others from aggression. As he states, "when these constraints are perceived as prohibiting operationally or tactically logical methods or means of warfare, it creates risk that the profession of arms—the very constituents who must embrace the law—will see it as a fiction at best or, at worst, that will feign commitment to the law while pursuing *sub rosa* agendas to sidestep obligations."[29]

Of course, the concern here is not simply that soldiers will sidestep legal or moral obligations because upholding them represents excessive risk. The fact that humans are motivated by self-preservation to do so only underscores the positive ethical role AI-systems can play. There are, however, deeper obligations at play here. States not only have an obligation to defend their citizens, they have an obligation to ensure that those citizens who come to that defense have every advantage to do so successfully at the least risk possible.[30] Thus any ban on LAWS or DSS, to the extent it limits chances for success or puts soldiers at greater risk, represents its own kind of moral failure.

This point, however, is insufficient to fully establish the permissibility of AI-systems. It is not enough to point out that AI-systems can lead to better ethical outcomes without fully accounting for the unethical ones. It may be permissible to accept some ethical "risk" regarding human incentives as these can be compensated for by additional rules and oversight. When those are inadequate, as I will discuss later, there are still other ways to address the responsibility gap given any particular human-machine relationship. This point suggests that where humans can establish sufficient control over these systems to be responsible for their behavior, their use would be permissible. The difficulty with this approach, however, is that such control usually comes at the expense of using this technology to its full capability. I will take up closing this "capability gap" later. The point here is that the responsibility gap—while real—is insufficient to establish AI-systems as *mala en se.*

Dehumanizing War and Respect for Persons

Another major critique of AI-systems is that they will dehumanize warfighting. On the surface this seems like an odd case to make. War may be a human activity, but rarely does it feel to those involved as a particularly humane activity, often bringing out the worst in humans more often than it brings out the best. If LAWS and DSS can shed some of the cruelty and pain war inevitably brings then it is reasonable to question whether dehumanizing war is really a bad thing. As Scharre notes, the complaint that respecting human dignity requires that only humans make decisions about killing "is an unusual, almost bizarre critique of autonomous weapons" and adds, "[t]here is no legal, ethical, or historical tradition of combatants affording their enemies the right to die a dignified death in war."[31]

Scharre is correct that AI-systems do not represent a fundamentally different way for enemy soldiers and civilians to die than those human soldiers are permitted to employ. The concern here, however, is not that death or bad planning by robot represents a more horrible outcome than when a human pulls the trigger or publishes an operations order. Rather it has to do with the nature of morality itself and the central role respect, where respect is understood in the Kantian sense as something moral agents owe each other, plays in forming our moral judgments.

Drawing on Kant, Robert Sparrow argues that respect for persons entails that even in war, one must acknowledge the personhood of those one interacts with, including the enemy. Acknowledging that personhood requires that whatever one does to another, it is done intentionally with the knowledge that whatever the act is, it is affecting another person.[32] Actors must see affected persons as subjects of that action and affected persons must see themselves as the object of that act is to be subject to moral judgment. The point here is that moral acts arise in the interaction of moral agents and that interaction requires an

interpersonal relationship.[33] That relationship does not require communication or even the awareness by one actor that he or she may be acted upon by another. It just requires that the reasons actors give for any act that affects another human being take into account the respect owed that particular human being. To make life-and-death decisions absent that relationship subjects human beings to an impersonal and pre-determined process and subjecting human beings to such a process is *disrespectful* of their status as human beings.

Thus a concern arises when *non-moral* agents impose moral consequences on *moral* agents. Consider, for example, an artificially intelligent legal process that imposes penalties on human violators. It is certainly conceivable that engineers could design a machine that could take into account a large quantity and variety of data and analyze it relative to a certain law or other legal standard. The difficulty with the judgment it renders, however, is its impersonal, pre-determined nature.[34] Absent an interpersonal relationship between judge and defendant, defendants have little ability to appeal to judges who may be able to get beyond the letter of the law and decide in their favor. In fact, the European Union has enshrined the right of persons not to be subject to decisions based solely on automated data processing. In the United States, a number of states limit the applicability to computer-generated decisions and typically ensure an appeals process where a human makes any final decisions.[35]

This ability to interact with other moral agents is thus central to treating others morally. Being in an interpersonal relationship allows all sides to give and take reasons regarding how they are to be treated by the other and to take up relevant factors that they may not have considered before-hand.[36] In fact, what might distinguish machine judgments from human ones is the human ability to establish what is relevant as part of the judicial process rather than before-hand. That ability is, in fact, what creates space for sentiments such as mercy and compassion to arise. This point is why only persons—so far at least—can show respect for other persons.

So if it seems wrong to subject persons to legal penalties based on machine judgment, it seems even more wrong to subject them to life-and-death decisions based on machine judgment. A machine might be able to enforce the law, but it is less clear if it can provide justice. Sparrow further observes that what distinguishes murder from justified killing cannot be expressed by a "set of rules that distinguish murder from other forms of killing, but only by its place within a wider network of moral and emotional responses."[37] Rather, combatants must "acknowledge the morally relevant features" that render another person a legitimate target for killing. In doing so, they must also grant the possibility that the other person may have the right not to be attacked by virtue of their noncombatant status or other morally relevant feature.[38]

The concern here is not whether using robots obscures moral responsibility; rather the concern here is that the employment of AI-systems obscures the good humans *can* do, even in war. Because humans can experience mercy and compassion they can choose not to kill, even when, all things being equal, it may be permissible. To illustrate this important point, at the end of *Slaughterbots*, some unknown actor releases a swarm of slaughterbots into a crowded classroom. As one of the bots moves towards a young student, whom the viewer knows engaged in a minor act of protest earlier, he begs for his life. The bot, of course, kills him. A human, who can take more into account than the limited targeting criteria of the bot may have been moved by his youth and pleas and chosen not to kill him.

The fact that AI-driven systems cannot have the kind interpersonal relationships necessary for moral behavior accounts, in part, for much of the opposition to their use.[39] If it is wrong to treat persons as mere means, then it seems wrong to have a "mere means" be in a position to decide how to treat persons. One problem with this line of argument, which Sparrow recognizes, is that not all employment of autonomous systems breaks the relevant interpersonal relationship. To the extent humans still make the decision to kill or act on the output of a DSS, they maintain respect for the persons affected by those decisions.

However, even with semi-autonomous weapons, some decision-making is taken on by the machine, mediating, if not breaking, the interpersonal relationship. Here Scharre's point is somewhat relevant. Morality may demand an interpersonal relationship between killer and killed but as a matter of practice, few persons in those roles directly encounter the other. An ISIS fighter would have no idea whether the bomb that struck him was the result of a human or machine process; therefore, it does not seem to matter much which one it was. A problem remains, however, regarding harms to noncombatants. While, as a practical matter, they have no more experience of an interpersonal relationship than a combatant in most cases, it still seems wrong to subject decisions about *their* lives and deaths to a lethal AI-system just as it would seem wrong to subject decisions about ones liberty to a legal AI-system. Moreover, as the legal analogy suggests, it seems wrong even if the machine judgment were the correct one.

This legal analogy, of course, has its limits. States do not have the same obligations to enemy civilians that they do towards their own. States may be obligated to ensure justice for their citizens and not be so obligated to citizens of other states. There is a difference, however, between promoting justice and avoiding injustice. States may not be obligated to ensure the justice of another state; however, they must still avoid acting unjustly toward that other state's citizens. In the context of warfighting, it is a generally held principle that one should not put enemy noncombatant civilians at any greater risk than one would one's own. So if states would not employ autonomous weapons on their own territory, then they should not employ them in enemy territory.[40]

What matters here, however, is context. States may choose not to apply LAWS in their own territory in conditions of peace; however, the technology could get to the point where they would employ such systems under conditions of war precisely because they are less lethal. If that were to be the case, then the concern regarding the inherent injustice of AI-driven systems could be resolved. So, this concern, while important, is not sufficient to justify a ban on autonomous weapons, as many writers suggest.[41]

Of course, the fact that such systems can be used in ways that satisfy the responsibility and respect requirements, does not entail that they will be. Thus there is much more to say about the conditions such systems may be employed as well as how to ensure human sentiments such as compassion and mercy are fully accounted for on the battlefield. To address these concerns, I will next discuss the kinds of moral hazards to which these systems give rise.

Moral Hazard and AI-Systems

While the use of AI-systems may not be inherently evil, the dual concerns of responsibility and dehumanization suggest their use will give rise to a number of moral hazards that need to be addressed if states are to ethically use these systems to their full capability. Moral hazards arise when one person assumes greater risk because they know some other person will bear the burden of that risk.[42] Given the reduction in risk—both for political leaders who decide to use force as well as the combatants who employ it—the employment of AI-systems will establish an incentive structure to ignore the moral risks described above. AI-systems may not be *mala en se*; however, that point does not entail they cannot or will not be used in an evil manner. To address this concern, we need to have a better account of how moral autonomy relates to machine autonomy and how humans, who have moral autonomy, can relate to machines.

Autonomy and Moral Agency

To mitigate the hazards of unaccountable moral failure and dehumanizing war, which are essentially two sides to the same coin, one either has to ensure a human makes life-and-death decisions or that machines develop the capability to act as an "autonomous moral agents." Both give rise to moral concerns unique to AI-systems. Keeping a human in the loop often means that one cannot take full advantage of the capabilities of LAWS and DSS systems. When that decision is to fire at a rapidly approaching missile, any delay

can make the difference between life and death. Additionally, employing such systems risk desensitizing soldiers, lowering the threshold for violence, and a bias favoring machine judgments over human ones.

To understand the difficulty in resolving these concerns it will help to understand what it would take to make an AI-system what Wendell Wallach and Colin Allen call "autonomous moral agents." They argue in their book, *Moral Machines: Teaching Robots Right from Wrong*, that moral agents require the ability to "monitor and regulate their behavior in light of the harms their actions may cause or the duties they may neglect."[43] Moreover, they further require the ability to "detect the possibility of harm or neglect of duty" as well as take steps to minimize or avoid any undesirable outcomes.[44] There are two routes to achieving this level of agency. First, is for designers and programmers to anticipate all possible courses of action and determine rules that result in desired outcomes in the situations the autonomous system will be employed. Second, is to build "learning" systems that can gather information, attempt to predict consequences of its actions based on that information, and determine an appropriate response.[45]

The former requires either a great deal of knowledge on the part of the programmer or a very limited application for the machine. It would be, of course, daunting—though not conceptually impossible, for a programmer to describe and then encode the different situations one may find in combat. However, what makes this approach especially difficult are the related problems of *ascription* and *isotropy*.

Ascription refers to the difficulties associated with inferring other persons' intent from their actions and isotropy refers to determining what elements of the environment or background knowledge are relevant to that ascription. Consider, for example, a group of soldiers who burst into a room looking for an insurgent and see a couple of young children wearing a knives and running in their direction. On what basis do they consider the children a threat? If the operation were conducted in a Sikh village, these soldiers might know that Sikhs often wear a traditional dagger, which is a religious symbol and never used as a weapon. Moreover, they might be able to discern from the way the children were running and other relevant environmental cues that children were, in fact, at play, and not to be taken as a threat.[46]

To program that capability in a robot would be a daunting, perhaps impossible, challenge. Ascribing mental states to others requires humans to see in others the beliefs, desires, hopes, fears, intentions, and so on, that they see in themselves. As Marcello Guarini and Paul Bello observe, the relevant information associated with such ascriptions is extensive and include such diverse things as "[f]acial expressions, gaze orientation, body language, attire, information about the agents movement through an environment, information about the agent's sensory apparatus, information about the agents background beliefs, desires, hopes, fears, and other mental states."[47]

This difficulty is frequently referred to as the *frame* problem. Human knowledge about the world is holistic where changes in one bit of knowledge can affect others. For example, learning that one is out of milk may require one to schedule a grocery shopping trip, which in turn might cause on to reschedule a meeting as well as determine to buy more milk than one had previously. While humans do this easily, computational AI-systems must go through all of its stored information to test how running out of milk affects it.[48]

Even when an AI-system can reasonably handle sorting through alternatives, their output—whether it is behavior or a course of action—lacks intention, an important component to moral analysis. In his famous "Chinese Room" thought experiment, philosopher John Searle described a man who sits in a room. His job is to take input, in the form of Chinese characters, and then consult a rule book that tells him what the output should be. He then takes the appropriate characters and provides them to whomever is outside the room. He does not understand the meaning of the characters, only how they relate to the rules. Thus, given a sufficiently complex rule book, he conceptually can mimic a fluent Chinese speaker without understanding anything being said.[49] Moreover, while he may be causally responsible for the output, since he does not

understand it, he cannot be said to intend its content. If he does not intend the content of the response, it can further be concluded that he is indifferent to it.

The point here is that one thing at least that differentiates humans from machines is that humans, in the words of AI theorist John Haugeland, "give a damn." He argues that understanding language depends not only on caring about one's self, but also the world in which one lives.[50] The human in the Chinese room may care (or at least can care) that he gets the rules right, but the machine itself does not have the capacity to care what the output actually means. There is a difference between something being manipulated according to a set of rules and someone acting, on one's volition, according to rules.[51] The output of the Chinese Room is clearly an example of the former.

At this point it is worth considering what structurally differentiates machines from humans. As Philosopher Harry Frankfurt puts it, what separates humans—or at least human agency—from that of animals is the ability to form effective second-order desires regarding first-order desires.[52] For example, as a matter of their physiology, both humans and animals will experience a desire for food. Only humans, however, can form a second-order desire regarding whether they *want* to desire food. Persons who are on a diet because they care about their weight, for instance, may not be able to choose whether they desire food, but they can choose whether they want to desire food. Moreover, they can choose which desire to act on, the first-order desire for food or the second-order desire not to eat.

To the extent the second-order desire is effective, they can be said to exercise their will in a way that is uniquely human. As Frankfurt says it, "a person that [sic] enjoys freedom of the will means … that he is free to want what he wants to want." Put another way, it means persons are free to have the will they want and are not bound by particular instincts or programming in how they respond to situations in which they find themselves.[53] It is this structure of the will that allows persons to set their own goals and direct their actions towards those goals in ways which they can be held morally responsible. This is what it means to be an autonomous agent in a way in which one can be held morally responsible for one's actions.

This concept of autonomy is, of course, different than that associated with autonomous systems. What makes an AI-system autonomous is that it is "self-directed towards a goal" and able to make decisions about how to achieve that goal without external interference.[54] As DoD Directive 3000.09 (Autonomy in Weapon Systems) defines it, a LAW is "a weapon system that, once activated, can select and engage targets without further intervention by a human operator."[55] What it cannot do—yet at least—is make decisions about the desirability of that goal. If it cannot do that, then, as described above, it cannot be considered morally autonomous in the human sense. It could be said that we will know when we have achieved true robot autonomy not when it perfectly and ethically employs lethal force; rather, we will know it when it refuses to use that force because, perhaps based on humanitarian grounds, it rejects the purpose for which it was designed.

There are, of course, different approaches to AI that do not operate according to a strict computational approach. Connectionist approaches, like Artificial Neural Networks or Parallel Distribution Processing, rely on networks of "neural like" processing units, similar to the way the human brain works. Machines based on this technology can learn from examples they encounter and determine responses without additional programming. While these systems seem more capable than symbolic ones at certain tasks, there are things symbolic ones can do better, such as handle sequential, complex tasks. The future of the technology will likely employ hybrid strategies to increase machine intelligence.[56] To the extent future technologies rely on such models, these concerns will endure.

Currently, however, AI-systems lack the ability to take into context or adapt to new situations outside their design parameters. What this limitation means is that while AI-systems may be able to accurately identify human faces, emotions and body movements, they cannot determine a plausible story explaining a person's

behavior. Without that capability, one will not likely be able to ascribe intent or motivation to a person's actions, which is typically necessary to determining if that person is a threat.[57] Moreover, even if a machine-learning process could improve the machine's ability in this regard, doing so would require interacting with the environment and in doing so these machines would be bound to make mistakes.[58] When those mistakes entail harm to human life, it is reasonable to ask where the responsibility lies, with the machine or the humans who employed it?

It is this capacity for self-direction that gives rise to moral concern. If the goal involves killing a human being, or in the case of DSS's, potentially placing soldiers at risk, then decisions associated with achieving those goals are inherently moral. As discussed above, it seems axiomatic then that moral decisions should be taken by moral agents. As Christopher Heyns asks the question, "is it right for machines to have the power of life and death over humans or the ability to inflict serious injury?"[59] The answer, as discussed previously, is no. Accountability requires that moral acts are committed by moral agents.

Moral Responsibility and Meaningful Human Control

However, there is space short of full moral agency to address these concerns. As Wallach and Allen argue, machines can exhibit operational and functional morality. Operational morality means humans design the machine in a way that reflects a moral norm. For example, placing a safety device on a rifle reflects a concern for individual well-being. Functional morality means the machine has the ability to assess morally significant aspects of how they operate. Moreover, machines do not have to be artificially intelligent to have this capability. For example, autopilots take passenger comfort into account when they make course corrections and limit the kinds of maneuvers they can make, not because these maneuvers might prevent the plane from reaching its destination, but because the designer cared about passenger comfort.[60]

In this way moral agents can impart moral values into the function of the machine. Doing so will not make the machine moral, but it is a step toward ensuring that the machine "behaves" in a moral fashion as well as closing the responsibility gap autonomous machines introduce. Moreover, the fact that designers and manufacturers can take into account operational and functional morality of the machine entails that they should.

Paying attention to the demands of operational and functional morality may not get you an autonomous moral agent, but it can set conditions for *trust.* Even "dumb" machines can be causally responsible for harms through no fault of the human operator. For example, a bullet fired at a legitimate target can ricochet and kill an innocent person. However, that does not mean that we cannot trust machines to function reliably enough within a certain context that their use is morally permissible. As Wallach and Allen point out, all technology fits on the dual spectrums of autonomy and ethical sensitivity. Some tools, like a hammer, have neither. The rifle has no autonomy but can have some ethical sensitivity reflected in its design. The autopilot has more of both autonomy and ethical sensitivity.[61] Within their respective functions, moral agents trust them to operate in accordance with norms for which moral agents may be held accountable. That trust does not mean there will not be accidents, but even then we typically understand the relationship between the actions of the moral agent and the system to know where accountability lies.

However, as noted above, that often may not be possible with artificially intelligent systems. So the question is, can one develop AI-systems to the point they have sufficient ethical sensitivity that moral agents can trust them to operate in ways those same moral agents can be held accountable. The answer depends on the relationship between the moral agent and machine. Before we can discuss the nature of those relationships, we need first to establish a standard of accountability so that we can understand what counts as meaningful human control.

Moral Responsibility

Frankfurt argued that the source of moral responsibility in humans is the ability to form second-order desires. When a human acts on that second-order desire, they are morally responsible for the resulting behavior. This suggests moral responsibility has two parts: 1) intent as reflected in the second order desire and 2) action, which results from the agent making that desire effective. This concept of moral responsibility maps well onto the legal standards regarding individual and command responsibility.

Criminal responsibility, in both civil and military contexts rests on an individual's intent to violate a law (*mens rea*) and his or her act of violating that law (*actus reus*).[62] This is the standard employed by the post-World War II Nuremburg trials as well as the Rome Statute that governs the International Criminal Court. As the Nuremburg trial document states, for an accused to be responsible for a war-crime, "there must be a breach of some moral obligation fixed by international law, a personal act voluntarily done with the knowledge of its inherent criminality under international law."[63] This criminal responsibility can also extend to anyone, such as commanders, who orders the commission of any crime.

It is clear from the previous discussion that an autonomous machine will not likely, in the near future at least, be capable of meeting such a standard as it cannot possess the relevant kind of intent. Therefore, responsibility for any violation committed by a machine will have to find a human host.

One implication of this point is that it may expand the kinds of persons we hold accountable for specific violations. Because soldiers possess the relevant autonomy, responsibility for any violations they commit rests, for the most part, with them. Because it rests with them, it makes little sense to charge the recruiters who bring them into the military or the drill sergeants who train them. One might, upon investigating, determine that there were flaws in the recruiting and training processes that contributed to the crime. In that case, it may make sense to hold individuals responsible for those flaws, but not for the specific violation itself.

In the case of AI-systems, however, that responsibility is diffused. Since the behavior of the machine has been wholly determined in advance by individuals related to its design, manufacture, and employment, it makes sense to hold them—to some degree at least—culpable of the actions of the machine.[64] The connection between their actions and the moral harm is not mediated by another moral agent who can exercise his or her own second-order desires and thus choose whether or at least how to respond to their influence.

Thus, more than operators and, in some circumstances at least, commanders, it may make sense to hold acquisition officials, programmers, and manufacturers responsible for crimes their machines commit as well. Of course, such responsibility will be limited to what they actually intended as they fulfilled their various roles in the acquisition and employment of the system. Just as is the case in every human endeavor, there is room for honest mistakes. To the extent one of these persons intended a violation of the war convention or other relevant norms or were grossly negligent or indifferent to those norms as they fulfilled their role, they can be held responsible not just for malicious intent, negligence, or indifference but for the specific violation the machine committed.

This responsibility, of course, has its limits. As noted earlier, AI-systems thinking can be something of a "black box," which could make it difficult to figure out who played what role in any particular harm. As Liu notes, as AI programming interacts with other machine systems and those systems interact with the operational environment machine behavior will become increasingly difficult to predict, much less account for. Moreover, that difficulty is further increased when these machines are employed in swarms. So even if individual machines were entirely controllable, when employed in large numbers, predicting and accounting for their aggregate behavior will be much more difficult.[65] Thus, as Liu points out, violations can arise not just from bad intent and negligence, but from everyone doing a "job well done."[66]

This point suggests that while diffusing responsibility will be a necessary feature in any ethic associated with the employment of AI-systems, it will not fully close the responsibility gap. It is possible, however, to have moral responsibility for outcomes one may not have intended based on the role one plays. Role responsibility attaches to an individual by virtue of the position they hold and the functions they are supposed to fulfil.[67] I have already argued that acquisition officials, designers, programmers, and manufacturers may have responsibility for specific violations by virtue of the role they played. However, that responsibility depends largely on what they intended or failed to intend relative to the violation in question. Commanders, on the other hand, can be held responsible even when they did not have a particular intent.

Of course, commanders are certainly responsible for any crimes they do intend, even if they do not commit them themselves; however, our understanding of command responsibility also entails that they are also responsible for crimes that they should have prevented, whether they intended them or not. For a commander to be held accountable for a war crime he or she neither intended nor committed, two conditions must hold: 1) the commander must have knowledge of the crime and 2) have responsibility for those perpetrating the crime.[68]

Regarding the first condition, it does not matter whether a commander did have knowledge of a particular crime, but whether he or she should have known. As the Nuremberg trial documents state, "an army commander will not ordinarily be permitted to deny knowledge of reports received at his headquarters, they being sent there for his special benefit."[69] Moreover, to the extent a commander has responsibility over an organization, he or she also takes on an affirmative duty to become cognizant of the actions of his or her subordinates. Also as the Nuremberg Trial documents state, "If he (a commander) fails to require and obtain complete information, the dereliction of duty rests upon him and he is in no position to plead his own dereliction as defense."[70] These points suggest that commanders must, in addition to not ordering illegal or immoral acts, must also take steps to ensure they are knowledgeable of their subordinate activities, limit unintended harm, and if those fail, take steps to hold violators accountable.[71]

This kind of command responsibility extends what it means to have meaningful human control. To the extent humans are in a position to give the machine instructions, understand the system well-enough that they can take steps to ensure it behaves appropriately, and act when it appears the machine has not, then humans have meaningful control.

This discussion so far suggests that establishing meaningful human control entails the following conditions: 1) acquisition officials, designers, programmers, manufacturers, as well as commanders and operators must fulfill their roles with the war convention in mind; 2) commanders and operators must not only be knowledgeable regarding what the machine is doing, they must be sufficiently knowledgeable regarding how the machine works so they better understand how it will interpret and act on instructions as well as provide output; 3) to the extent possible commanders and operators must be in position to prevent machine violations, either by ensuring they authorize all potentially harmful actions by the machine or that they are able to monitor the operations of the machine and prevent them from happening; and 4) systems where operator intervention is not possible should only be employed in situations where commanders and operators can trust them to perform at least as well as human soldiers.

With this concept of meaningful control in mind, it remains to be discussed how to manage the additional moral hazards associated with AI-systems. Obviously, the closer the relationship, the less the responsibility gap or dehumanizing warfare are. However, the more dependent humans become on machines, one runs the risk of moral violations that humans may be responsible for, but which they are more likely to make.

Meaningful Human Control: Autonomy, Responsibility, and Human-Machine Relationships

Humans interact with autonomous and semi-autonomous machines generally in three ways: "human in the loop," "human on the loop," and "human off the loop."[72] Here, the "loop" is the "sense-decide-act"

operation the machine performs relative to a particular purpose. When humans are in the loop, the machine waits for human input after performing a task. When humans are on the loop, the machine can sense, decide, and act on its own, but a human monitors the system and can intervene to prevent it from acting in an undesirable manner. When humans are off the loop, the machine senses, decides, and acts on its own, without human supervision.[73] These are, of course, imprecise descriptions of how humans and machines interact. Even in a fully automated system, there is a human in the loop somewhere, whether in the design of the system, the software, or in decisions regarding employment.[74] These machines, after all, are human creations. However, for the purposes of this discussion, these categories of interaction are sufficient to illuminate a number of moral concerns.

Humans in the Loop

The Predator and Reaper unmanned aerial vehicles described above fit into the first category, where some elements of the system such as take-off, landing, and navigation may be automated, however, life-and-death decisions are made by humans. Locating moral responsibility for lethal decisions in such systems is not an ethical concern since it remains with humans. However, that responsibility may be mediated to the extent humans rely on AI-driven assessments, such as those Project Maven seeks to develop, in making those decisions.[75] Thus, human in the loop systems may not only lower the risk to combatants by distancing them from their targets, but they can also distance operators from their decisions to kill or otherwise act on AI-generated recommendation. This distancing gives rise to concerns regarding the psychological impact on operators as well as senior leaders. In the case of the former, it may desensitize them to the killing or other harms they do, making them easier to commit. In the case of the latter, such distancing lowers the physical and political risk associated with using force, making resorts to violence more likely.

Psychological Effects: Desensitization and Trauma

General William T. Sherman famously observed that "war is hell." Prefacing that comment he stated "[i]t is only those who have neither fired a shot nor heard the shrieks and groans of the wounded who cry aloud for blood, more vengeance, more desolation.[76]" With the advent of remote controlled "human-in-the-loop" systems war fighters may fire shots, but they no longer directly experience the associated shrieks and groans. In fact, much of the literature on LAWS cites studies by SLA Marshall's and David Grossman's observations that human soldiers come with a natural reluctance to kill. Marshall observed that only 15-20% of riflemen in World War II engaged the enemy and Grossman's observation that after the Battle of Gettysburg, 90% of the muskets recovered were still loaded or loaded with multiple rounds, suggesting soldiers pretended to fire.[77] These studies, and others like them, attribute—justifiably so—the reasons for this reluctance to shoot as fear of being killed and resistance to killing. Our fear of being killed makes us reluctant to take the risks necessary to engage the enemy, and even when we do so, an innate natural sympathy to our fellow humans can make us reluctant to pull the trigger. Distance plays a role in mitigating both these concerns.

In general, trends in military technology have been to distance soldiers from the killing they do. Crossbows allowed killing at greater distances than did swords, rifles farther than crossbows, cannons and artillery farther than rifles. What is different about autonomous weapons is that they do not just distance soldiers from killing, they can also distance soldiers from the decision to kill. While this is clearly true in fully autonomous systems, it can be true for semi-autonomous systems as well. As P.W. Singer notes in *Wired for War*, "By removing warriors completely from risk and fear, unmanned systems create the first complete break in the ancient connection that defines warriors and their soldierly values."[78]

As Singer goes on to observe, the traditional warrior identity arises from conquering profound existential fear, "not the absence of it."[79] The result is a fighting force that is not merely distanced from risk, but disconnected from it altogether. As one Air Force lieutenant reportedly said about conducting unmanned air

strikes in Iraq, "It's like a video game. The ability to kill. It's like … freaking cool."[80] In that case, a human was on the loop … imagine how desensitized we might become if we are off the loop altogether.

Of course, since there are few fully autonomous systems available, there is little information on how operators and commanders might respond to the killing such machines would do. Having said that, Singer offers us, perhaps unintentionally, a cautionary tale. After the fall of Saddam, U.S. forces embarked on a manhunt for the regime's former leaders, including Saddam's sons Uday and Qusay. When they were cornered by U.S. forces in a villa in Mosul where they were hiding, a large fire-fight ensued. Meanwhile, hundreds of miles away in Qatar, staff members at the command center gathered around the television screens showing the feed from an unmanned aerial vehicle circling the battle. Rather than being horrified at the carnage, they were entertained. As Singer notes, "it was like a Super Bowl party in there," adding that a number of participants brought snacks and would cheer when there was a "particularly big explosion."[81]

The point here is not to criticize what might be argued was a reasonable response to the demise of some particularly horrible people.[82] For all Singer, or the author, know, the soldiers on the ground experienced equal joy at the sound of large explosions and would have brought snacks had that been allowed.[83] Rather the point here is to observe that autonomous technologies *can* set conditions where operators take a casual approach to the killing they do. The key word here, of course, is "can." There are features of human-in-the loop operations that can actually increase sensitivity to killing and set conditions for post-traumatic stress syndrome or moral injury in its operators. In fact, in 2015 a large number of drone operators quit, some citing overwork, others citing the horrors they felt responsible for as reasons.[84]

As Samuel Issacharoff and Richard Pildes observe, the use of human-in-the-loop LAWS have increased the individuation of responsibility for killing and thus bring about a greater sense of responsibility for the killing they do.[85] One feature that increases sensitivity is the amount of time unmanned aerial vehicle pilots spend observing their targets and then watching the effects and after-effects of strikes they initiate. One unmanned aerial vehicle operator, Brandon Bryant, reported as a source of emotional stress the fact that after a strike he would not only sometimes have to watch his targets die, but also review the aftermath. Recounting one strike in Afghanistan, he not only observed the strike but also the bodies and body parts afterward. One particularly disturbing image was watching one of the individuals struck. As he recalls, "It took him a long time to die. I just watched him. I watched him become the same colour [sic] as the ground he was lying on."[86]

This kind of interaction is typically not a feature of conventional strikes. As one unmanned aerial vehicle pilot put it, "I doubted whether B-17 and B-20 [sic] pilots and bombardiers of World War II agonized much over dropping bombs over Dresden or Berlin as much as I did taking out one measly perp in a car."[87] The point here is not that increased use of semi-autonomous and autonomous weapons will bring about greater or less sensitivity to killing or greater or less trauma. The point here is that as the character of war changes different persons will respond differently. The ethical imperative is that leaders pay attention to those changes and take steps to mitigate their ill-effects.

Decreasing the Threshold to War

Lowering risk to soldiers also lowers risk for civilian leadership when it comes to decisions regarding when to use such weapons. Of course, this concern is not unique to autonomous systems. Any technology that distances soldiers from the violence they do or decreases harm to civilians will lower the political risks associated with using that technology. The ethical concern here is to ensure that decreased risk does not result in an increase in the number of unjust uses of these weapons.[88] Otherwise, the moral advantage gained from greater precision will be offset.

As Christian Enemark argues, "Political leaders, having less cause to contemplate the prospect of deaths, injuries and grieving families, might accordingly feel less anxious about using force to solve political

problems."[89] Like concerns regarding desensitization, concerns regarding lowering the threshold to war may be overstated. While arguably the use of unmanned aerial vehicle strikes has expanded over the last decade, instances of escalation into wider conflict have not. Even in areas such as Pakistan, Yemen, and Somalia, where the United States is not at war, the conflict in question preceded the use of unmanned systems, not the other way around. Thus the ethical question is whether, if human-in-the-loop technology were not available, would (and should) the United States do *something.* If the answer is yes, then to the extent the use of force is just and the use of LAWS makes the use of force more precise and humane, it is at least permissible. If the answer is no, then it is likely that *no* force would be permissible.

The difficulty in resolving this concern is that, much like the concern regarding desensitization, it pits a psychological claim regarding human motivations to employ violence against moral claims associated with the permissibility of violence. The answer to one question just is not an answer to the other. So while it may be true that lower risks make decisions about using force easier, it is irrelevant to whether such force is permissible. Having said that, to the extent the psychological concern is valid, it makes sense to ensure decisions to use risk-decreasing weapons are subject to strict oversight to ensure the conditions of justice are met as well as any other measures that might mitigate these effects. The absence of this oversight and transparency is, in fact, often cited in the literature as a genuine moral concern and has been a longstanding criticism of the U.S. unmanned aerial vehicle operations.[90] Given this concern, it makes sense to ensure such oversight and transparency is in place. In this way one can ensure the human reliance on the machine does not set humans up for moral failures they may otherwise not make.

Another concern is that even when LAWS are employed ethically in the service of legitimate U.S. interests, their use may drag the United States into local conflicts whose justice may be questionable. Enemark notes a debate within the Department of Defense regarding whether such strikes are permitted only against high value targets or also permitted against the larger number of low-level militants whose concerns are more local. He observes that the narrower set is more defensible as pre-emptive strikes to the extent these individuals are actively plotting against the United States and the lower-level militants who are motivated to fight for local concerns. As he states, "The narrow view is more easily defensible because individuals who are actively plotting to attack the United States more obviously attract (pre-emptive) defensive action than do individuals who merely happen to possess an antipathy towards the United States."[91]

Of course, this concern arises as much out of the fact that networks of terrorists threatening the United States draw on and cooperate with networks of oppositionists whose concerns are local, sometimes to the point where it is difficult to distinguish between the two. Thus, regardless of the means used, engaging the former risks expanding conflict to the United States with the latter, who would not otherwise be a threat. While this concern is real, it is more a feature of the character of the conflicts the U.S. finds itself in rather than the weapon system itself. In fact, it is conceivable that AI-assisted analysis could increase U.S. military's capability of differentiating between these local and transnational networks.[92] Having said that, the fact of this dynamic suggests the United States should adopt the narrower policy and employ a principle of conservatism when pressure to expand targeting to local targets increases. It may be permissible to do so; however, there should be a demonstrable relationship between the putative target and any threat to the United States.

Humans on the Loop: As noted previously, while human in the loop systems are the least morally risky, they can also reduce the benefit from employing AI significantly. To offset that disadvantage, one can put humans on the loop, where they monitor the activity of the machine and intervene only to prevent a violation or other malfunction. While still potentially limiting the full effectiveness of the machine, this relationship has the advantage of maintaining meaningful human control throughout a targeting or other decision cycle. However, in addition to the moral hazards associated with human-in-the-loop systems, human-on-the-loop systems come with a significant concern of their own.

When humans are on the loop, they are monitoring the machines behavior and, hopefully, evaluating it for appropriateness. To properly evaluate that behavior for appropriateness, however, humans have to be able to trust the information they receive. Sometimes that trust can be taken too far and humans may inappropriately subordinate their judgment to that of a less capable machine. Moreover, systems do not have to be that advanced for that inappropriate subordination to happen.

One of the most often cited examples of this phenomenon is the shoot-down of Iranian Air Flight 655 airliner by the USS Vincennes in 1988. The USS Vincennes was equipped with the AEGIS air defense system which is fully autonomous but has humans monitoring it as it goes through its targeting cycle. Humans can override the system at any point in this cycle and, in fact, the system was set to its lowest degree of autonomy. The jet's path and radio signature was consistent with civilian airliners; however, the system registered the aircraft as an Iranian F-14, and thus an enemy.[93]

As Singer retells the event,

> Even though the hard data were telling the human crew that the plane wasn't a fighter jet, they trusted what the computer was telling them more. Aegis was on semi-automatic mode, but not one of the eighteen sailors and officers on the command crew was willing to challenge the computer's wisdom. They authorized it to fire. … Only after the fact did the crew members realize that they had accidentally shot down an airliner, killing all 290 passengers and crew, including 66 children.[94]

The difficulty for humans in situations like this is that the complexity of machine "thinking" coupled with the pressure to act, especially in combat, disposes them to trust the machine, especially when doing so can absolve them of at least some of the responsibility of the action in question. Moreover, as at least one study has shown, that trust can emerge independent of the reliability of the machine. One study conducted by Korean researchers indicated that the most important factors in human assimilation of DSS systems were: institutional pressure, mature information technology infrastructure, and top management support. Quality of information, stated the report, had no significant impact on DSS assimilation.[95]

Thus the concern with human in the loop systems is that even though humans can prevent inappropriate machine behavior, often they will not. That counter-intuitive outcome arises from the fact that what the machine often presents to the human is a judgment, but the human takes it as fact. This certainly seemed to be case in shoot-down of the Iranian airliner. The *fact* was that there was an aircraft approaching the Vincennes, which the system *judged* was enemy. From the context, specifically the flight path and radio signature, the humans on board should have questioned the machine and aborted the attack.[96] As machine judgments become more complex, this concern is only going to get worse.

This point suggests that operators are going to need to develop sufficient expertise to know what sources of bad judgement are. It will also require operators to also adopt a "principle of conservatism," regarding when they should trust the machine without corroborating its output, and limit those times to only what is necessary to accomplishing the mission at hand. To facilitate that trust, as Scharre and Horowitz argue, designers will have to do their best to ensure the output of AI-systems are "explainable" to at least the operator, if not commander.[97]

The good news here, as Scharre points out, is that the most successful AI-systems will be those that rely on human-machine interaction, suggesting the most successful systems will either be human-in-the-loop or human-on-the-loop systems.[98] These systems, which he refers to as "Centaur Systems," are intentionally designed to maximize the speed and accuracy of a hybrid human-machine system in given situations. Examples include defensive systems such as the counter-rocket, artillery, and mortar systems that autonomously create "do not engage" sectors around friendly aircraft. These systems have a human

"fail-safe" to ensure engagements outside those sectors avoid fratricide or harm to civilian aircraft that might approach too closely.[99]

As noted above, such systems pose the least, though still real, concern regarding the responsibility gap and dehumanizing warfare. As long as those who employ those systems take care to address conditions that could lead to desensitization; rigorously enforce international law regarding the use of force, and conscientiously apply a principle of conservatism regarding whom to target as well as when and how to trust machine judgments, the employment of AI-systems would be morally permissible. However, there will still be times taking full advantage of a system will require the human to be off the loop. Regarding those situations, there will still remain concerns regarding whether such use is inherently evil. I will turn to that next.

Humans off the Loop: The difficulty for human-off-the-loop systems, of course, is that once they are launched, meaningful human control appears to be lost. However, as noted above, humans exert meaningful control in multiple ways, not just in deciding to pull the trigger to act on an AI-system's advice. Acquisition officials, designers, programmers, and manufacturer have opportunities to ensure the system meets the highest standards of operational and functional morality. Taking those opportunities, unfortunately, will not ensure ethical behavior by the machine any more than the most conscientious drill sergeant will ensure ethical behavior of the soldier. Even when all actors in the acquisition process are meeting their responsibilities, the complexity of the machine coupled with the complexity of the environment it operates in suggests that there will be violations for which it will be impossible to assign blame. It is easy to imagine that such situations will incentivize blaming the machine and obscuring the possibility of human error. That fact alone should give rise to moral concern.

It is worth noting that in addition to the fact the more successful systems will at least have the possibility of meaningful human control, there are few fully autonomous systems operational in combat environments. Slaughterbot swarms may be frightening and even technically possible, but so far there have been no actors who have employed anything like them. Currently, the Israeli Harpy, and the Chinese knock-off ASN-3-1, which is a copy of the Israeli system, are probably the best example of such systems. These systems are fully autonomous airborne loitering munitions that once launched orbit in a designated area search for radar signals associated with air defense systems. When they detect a signal, they fly into the source, destroying it.[100] Given the fact that their on-board explosive is relatively small, the chances of collateral harm are low. As long as the only radar operating in the area are enemy air defense radars, the Harpy and its variants are likely sufficiently discriminate to meet the standards of *jus in bello.*

However, as discussed above, the drive for autonomous systems suggests states will develop more complex systems for more complex environments. The task, then, is to determine what would count as meaningful human control in human-off-the-loop systems. At first blush, it may seem impossible to meet such a standard since, by definition, there is no human on the loop to offer meaningful control. However, as also discussed, humans impart control in a variety of ways throughout the acquisition and employment process. The question then is, how do we ensure all those ways give us the kind of control necessary.

The first difficulty that arises is that there is no set standard for meaningful human control that could apply. At one extreme, activist groups such as the International Committee for Robot Arms Control argue that meaningful human control must entail human operators must have full contextual and situational awareness of the target area. They also need sufficient time for deliberation on the nature of the target, the necessity and appropriateness of attack, and the likely collateral harms and effects. Finally, they must have the means to abort the attack if necessary to meet the other conditions.[101]

The difficulty with these criteria is that it essentially bans the use of off-the-loop systems. Moreover, it holds the systems that it does not ban to a higher standard than non-autonomous weapon systems already in use. Rarely in war do soldiers and their commanders have "full contextual and situational awareness of a target

area." Even when they do, soldiers who fire their rifles at an enemy have no ability to prevent the bullet from striking wherever they aimed it. It seems odd, then, to ban future weapons based on higher standards than the ones current weapons meet.[102] It makes less sense when one realizes that some of the capabilities that come along with autonomous weapons can set conditions for better moral decision-making.

As discussed previously, a greater reliance on autonomy can result not only in better contextual and situational awareness, but also greater deliberation when determining whether to engage a target than typically associated with humans. As Heyns, himself an opponent of LAWS, points out, "the increased depersonalization in the deployment of force brought about by [autonomous weapons systems] may thus lead to greater personalization in targeting outcomes and saving lives or preventing unwarranted injuries."[103] As previously discussed, there are a host of factors that can make deployment of AI-systems less lethal than leaving all the killing or operational decisions to humans.[104]

These facts, however, do not and cannot fully resolve concerns regarding the responsibility gap and the dehumanization of warfare. Moreover, it is not sufficient to say that the better outcomes these machines provide permits one to set aside the rights of those negatively affected. Doing so would reduce our ethical concerns to those of simple utilitarianism, which justifies any means given a certain ends.[105] That outcome is something to be avoided.

However, there is at least one good reason to set aside these concerns in favor of employing fully autonomous AI-systems. In general, to the extent some act makes someone better off and no one worse off, then that act is permissible. As Isaak Applbaum argues, "If a general principle sometimes is to a person's advantage and never is to that person's disadvantage, then actors who are guided by that principle can be understood to act for the sake of that person."[106] Thus, if Arkin and others are right about AI capabilities, then as long as their employment does not put any persons at more risk than if those systems were not employed, then arguably states are at least permitted, if not obligated, to use them. Because employing these systems under such conditions constitutes acting for the sake of those persons, it also counts as a demonstration of respect towards those persons, even if the interpersonal relationship Sparrow described is mediated, if not broken, by the machine.

One can further close the responsibility gap by paying attention to the circumstances in which AI-systems are employed. As previously discussed, command responsibility entails knowledge of what subordinates are doing and a responsibility for their actions. However, no commander has perfect knowledge of what his or her subordinates are doing, and, despite that fact, a commander can still be responsible for any violations those subordinates commit. Thus what matters to commanders is whether they trust those subordinates to function ethically and effectively in situations where the possibility of war convention violations arise. So, to the extent a commander can trust an AI-system to function at least as ethically and effectively as human soldiers, then employment of these systems should be morally permissible.

This bar is not as high as it sounds. Recall that the fact of harms to noncombatants by itself is not a violation of the war convention. Such harm only becomes a violation when it is the product of an indiscriminate or disproportionate use of force. Regarding DSS systems like Course of Action Display and Evaluation Tool, standards for humans should apply to the machine as well. Thus commanders have as much responsibility to be able to assess the output of DSS as they do for the output of human planning processes. These points suggest that while commanders may not be responsible for every mistake a machine makes, they will have to justify their trust in the machine and cease use of that machine when there are reasons to question that trust, just as they should for their human subordinates. Requiring commanders to account for this trust will help ensure AI-systems do not become an excuse for bad judgment and unaccountable wrong-doing. This requirement may not fully close the responsibility gap; however, it probably closes it enough to permit the use of human-off-the loop systems under the right conditions.

As previously discussed, determining what those conditions are requires an adequate understanding of how the machine works and more specifically how it will interact with its environment. Guarini and Bello argue that military activity occurs in theaters that can be described in terms of spectrums of activity with one end populated by combatants and the other populated by non-combatants. In space dominated by combatants, the task is to distinguish friend from foe. In space dominated by non-combatants, the task is not only to distinguish friend from foe, but also legitimate from illegitimate targets. To the extent friend and foe are relatively easily distinguished, autonomous systems will likely be able to perform it better than humans.

In non-combatant dominated space, the task is significantly more difficult, as previously discussed. In these "civilian theaters," where soldiers operate in the same space civilians conduct their daily lives, the tasks for AI-systems can be vastly more complicated. In such theaters, almost anything can become a threat, yet the preponderance of civilians entails that one cannot assume a threat. Someone carrying a pipe, for example, could be a plumber or a bomber.[107] To the extent an AI-system would have difficulty sorting through the complexities of making that determination, its use would not be permissible.

Thus, generally speaking, the greater combatant domination in a particular theater is, the more trust, all other things being equal, commanders may have in the machine. The greater the noncombatant domination, the less trust they should have. Determining where that point lies should not only be a matter of human judgment, but will also be a judgment for which we will hold humans to account. The uncertainties associated with these judgments are certainly complex, but this fact should encourage a conservative approach to employing human off-the-loop systems.

Conclusion

What this analysis has shown is that the arguments for considering military AI-systems, even fully autonomous ones, *mala en se* are on shakier ground than those that permit their use. It is possible to reduce, if not close, the responsibility gap and demonstrate respect for persons even in cases where the machine is making all the decisions. This point suggests that it is possible to align effective AI-systems development with our moral commitments and conform to the war convention.

Thus calls to eliminate or strictly reduce the employment of such weapons are off-base. If done right, the development and employment of such weapons can better deter war, or failing that, reduce the harms caused by war. If done wrong, however, these same weapons can encourage militaristic responses when other non-violent alternatives were available, resulting in atrocities for which no one is accountable, and desensitizing soldiers to the killing they do. To promote the former and avoid the latter, the United States should consider the following measures to ensure the ethical employment of these weapon systems. These measures include:

- *Work with AI-developing States to Update International Law:* As previously discussed, international law abhors a vacuum and makes the introduction of any system that mitigates or removes human responsibility problematic. On the other hand, many AI-developing states will take advantage of this vacuum to maximize the effectiveness of these systems, sometimes, if not often, without regard to the moral concerns discussed above. This point suggests the need to update the International Humanitarian Law and other applicable international law, to specify standards of responsibility for the employment of semi-autonomous and fully autonomous systems. These standards would include something like:
    - Humans in the loop would be fully responsible for the behavior of their systems. That responsibility can be mitigated to the extent the human operator relies on AI-driven DSS to make targeting decisions. In that case, responsibility for violations could fall to others involved in the procurement process. However, operators and commanders would have to be able to justify trust in any facts or judgments made by the machine. If that justification is inadequate, then they may be responsible for any violations.

- Humans on the loop would be responsible for the quality of information and assessment the machine provides and responsible for violations in the event inappropriate machine behavior is the result of poor quality information or assessment.
- Humans off the loop will be responsible for ensuring AI-systems are only employed in conditions for which they are designed to perform ethically. They are also responsible for monitoring the environment and the machine and ceasing operations when conditions changes in a way that sets conditions for violations.

- *Establish standards for diffusing responsibility.* States should establish standards for holding acquisition officials, programmers, designers and manufacturers responsible for machine violations. These standards will be especially important for fully autonomous systems where conditions for trust by commanders and operators are heavily dependent on the procurement side for ensuring that the machine meets standards associated with operational and functional morality. Meeting such standards would entail accounting for International Humanitarian Law when determining the features of the machine in the same way one might incorporate a safety device on a rifle.
- *Maintain a reasonably high threshold for use.* To ensure employment of AI-systems does not inappropriately lower the threshold to violence, states should agree to only employ these systems when the conditions of *jus ad vim* are met. These conditions permit an armed response for acts of aggression that fall short of war, but include the other standards of *jus ad bellum* as well as the requirement to take steps to avoid escalation. *Jus ad vim* also entails an obligation to ensure a high degree of probability that the use of force will achieve the desired objective.[108]
- *Specify conditions for employment.* Given the different human-machine relationships, states should specify conditions for use that ensure meaningful human control and the appropriate trust relationships are maintained. Update these standards as the technology evolves to avoid further gaps between effective use of AI-systems and moral commitments.
- *Regulate AI proliferation.* States that develop AI-systems for military use should establish proliferation standards similar to the ones the U.S. has established for the proliferation of unmanned aerial vehicles. At a minimum, these standards should include a commitment to only employ these systems in conflicts that meet the standards of *jus ad bellum*, and in a manner that meets the standards of *jus in bello*. Moreover, there should be a strong presumption of denial to recipients of the technology who have, in the past, been weak on their commitments to these standards.[109]
- *Preserve the Soldier Identity and address conditions that give rise to desensitization and other psychological trauma.* As the U.S. military becomes more reliant on AI-technology, soldiers will experience less risk, but not less trauma. Senior leaders should continue efforts to understand the nature of this trauma and take steps to mitigate. Moreover, disconnecting soldiers from the risk will also affect how society views and rewards military service. Senior leaders should take steps now to mitigate this potential moral hazard. One step could be to rotate AI-system operators in and out of assignments that expose them to risks commensurate with the conflict in question. Doing so will prevent the creation of a class of "riskless" soldiers and moderate the impact of this technology of civil-military relations.
- *Communicate the Principles regarding AI use*. Military leaders should develop a communications plan to explain the ethical framework for AI use to the public, media, and Congress.[110]

As Sharkey observes, the heavy manpower requirements with remotely controlled systems will place greater pressure to design and employ increasingly autonomous systems.[111] This point, coupled with the increased

effectiveness these systems afford, suggest the trend towards fully autonomous systems is inevitable. As this pressure mounts, commitments to keep humans in and on the loop will be increasingly difficult to keep. Fortunately, as the above analysis indicates, it is possible to manage the moral hazards associated with this technology to ensure moral commitments to human dignity, the rule of law, and a stable international order are met. Doing so may not assuage every Google employee; however, it will ensure that in acquiring these systems, the U.S. avoids evil.

End Notes

1 Scott Shane, Cade Metz, Daisuke Wakabayushi, "How a Pentagon Contract Became an Identity Crisis for Google," *New York Times*, May 30, 2018, available from https://www.nytimes.com/2018/05/30/technology/google-project-maven-pentagon.html, accessed June 28, 2018. It is worth noting that Google's commitment to avoiding evil is less than consistent, given their proposed cooperation with the Chinese government to provide a censored search engine. See Rob Schmitz, "Google Plans for a Censored Search Engine in China," *All Things Considered,* National Public Radio, August 2, 2018, available from https://www.npr.org/2018/08/02/635047694/google-plans-for-a-censored-search-engine-in-china, accessed August 6, 2018.

2 Future of Life Institute, *Lethal Autonomous Weapons Pledge, https://futureoflife.org/lethal-autonomous-weapons-pledge/?cn-reloaded=1,* accessed August 27, 2018.

3 Stanley Hauerwas, "Pacifism: Some Philosophical Considerations," in *War, Morality, and the Military Profession,* Malham M. Wakin, ed. (Boulder, CO: Westview Press, 1986), 282.

4 Paul Mozur, "Beijing Wants AI to be Made in China by 2030," in *The New York Times,* July 20, 2017. https://www.nytimes.com/2017/07/20/business/china-artificial-intelligence.html, accessed August 9, 2018.

5 Erwin Danneels, "Disruptive Technology Reconsidered: A Critique and Research Agenda," in *The Journal of Product Innovation Management,* 2004: 21, 249.

6 Paul Scharre, *Army of None: Autonomous Weapons and the Future of War,* New York: W.W. Norton and Company, 2018, 6.

7 Ronald G. Arkin, *Governing Lethal Behavior in Autonomous Robots,* Boca Raton: CRC Press, 2009, 7-8.

8 Larry Ground, Alexander Kott, and Ray Budd: "Coalition-based Planning of Military Operations: Adversarial Reasoning Algorithms in an Integrated Decision Aid," *Computing Research Repository,* 2015, 2-8. Quoted in James Boggess, "More than a Game: Third Offset and the Implications for Moral Injury," in *Closer than You Think: The Implications of the Third Offset Strategy,* Carlisle, PA: Strategic Studies Institute, 2017, 133. With the Course of Action Display and Evaluation Tool DSS, a human planner provides key operational goals and the program expands them into a detailed plan. In a 2002, the system competed against a team of field grade officers using the standard manual system. Course of Action Display and Evaluation Tool produced a plan in approximately two minutes. It took the field grade officers 16 hours to produce a comparable plan. See also Valerie Insinna, "U.S. Air Force looks to accelerate artificial intelligence contracts," *Defense News,* July 17, 2018, available from https://www.defensenews.com/digital-show-dailies/farnborough/2018/07/17/air-force-looks-to-accelerate-artificial-intelligence-contracts/?utm_source=Sailthru&utm_medium=email&utm_campaign=EBB-7-18&utm_term=Editorial%20-%20Early%20Bird%20Brief, accessed July 18, 2018. This article discusses efforts by the U.S. Air Force to acquire DSS to facilitate aircraft maintenance.

9 Scharre, *Army of None*, 16.

10 Lukas Mikelionis, "UC Berkeley's Professor Slaughterbots video goes viral," Fox News, November 21st, 2017, available from http://www.foxnews.com/tech/2017/11/21/uc-berkeley-professor-s-slaughterbots-video-on-killer-drones-goes-viral.html, accessed July 13, 2018. For other interesting, and less dystopian, depictions where the technology may heading, see also: "New Weapon Designed By Russian Inventor Demonstrating Of Destroying US, Israel and Russian Tanks," available from https://www.youtube.com/watch?v=gIJaG7X6918 or "New Weapons of the Russian Army 2018," https://www.youtube.com/watch?v=ihVszNrz6dQ, accessed July 13, 2018.

11 Scharre, *Army of None,* 11-13.

12 Paul Scharre, "Why You Shouldn't Fear Slaughterbots," *IEEE Spectrum,* December 22, 2017, https://

spectrum.ieee.org/automaton/robotics/military-robots/why-you-shouldnt-fear-slaughterbots, accessed July 13, 2018.

13 A common criticism of utility theory is that it can justify, depending on the balance of interests in question, unjust acts. See Louis P. Pojman, *Ethics: Discovering Right and Wrong,* Belmont, CA: Wadsworth Publishing Company, 1995, 122.

14 Michael Walzer, *Just and Unjust Wars,* 2nd ed, New York: Basic Books, 1992, 44-47.

15 Ibid., 131-137.

16 Ibid., 129.

17 Ibid., 144.

18 Arkin, 30.

19 Noel Sharkey, "Killing Made Easy," in Patrick Lin, Keith Abney, and George A. Berkey eds., *Robot Ethics,* Cambridge, MA: The MIT Press, 2014, 113.

20 Wendell Wallach and Colin Allen, *Moral Machines: Teaching Robots Right from Wrong*, Oxford: Oxford University Press, 2009, 4.

21 Paul Scharre and Michael Horowitz, *Artificial Intelligence: What Every Policy Maker Needs to Know,* Center for New American Security, June 2018, 11, available from https://www.cnas.org/publications/reports/artificial-intelligence-what-every-policymaker-needs-to-know, accessed August 7, 2018. As an example of "black box" thinking, Scharre and Horowitz note that an "AI image recognition system may be able to identify the image of a school bus, but not be able to explain what features of the image cause it to conclude that the picture is a bus."

22 Wendell Wallach and Colin Allen, "Framing robot arms control," *Ethics of Information Technology,* Vol. 15, 2013, 127.

23 Heather Roff, "Killing in War: Responsibility, liability, and lethal autonomous robots," in Fritz Allhoff, Nicholas G. Evans, and Adam Henschke, eds., *The Routledge Handbook of Ethics in War,* New York: Routledge, 2013, 355.

24 Wallach and Cohen, *Moral Machines*. 40.

25 Colnel James Boggess, Interview with author, January 4, 2017.

26 Hin Yan Liu, "Refining responsibility: differentiating two types of responsibility issues raised by autonomous weapons systems," in Nehal Bhuta, Susanne Beck, Robin Geiβ, Hin-Yan Liu, Claus Kreβ, eds., *Autonomous Weapon Systems: Law, Ethics, Policy,* Cambridge: Cambridge University Press, 2016, 341-344.

27 Geoffrey Brennan, Lina Eriksson, Robert Goodin, and Nicholas Southwood, *Explaining Norms,* Oxford: Oxford University Press, 2013, 35-39.

28 These principles are also stated in Army doctrine. See Headquarters, Department of the Army, Army Doctrine Reference Publication, No. 1, *The Army Profession*, Washington, D.C.: U.S. Government Printing Office, June 2015, 3-5. I owe this point to Michael Toler, Center for the Army Profession and Ethic.

29 Geoffrey S. Corn, "Risk, Transparency, and Legal Compliance," in Nehal Bhuta, Susanne Beck, Robin Geiβ, Hin-Yan Liu, Claus Kreβ, eds., *Autonomous Weapon Systems: Law, Ethics, Policy,* Cambridge: Cambridge University Press, 2016, 217-218.

30 Michael Walzer, *Arguing About War,* New Haven: Yale University Press, 2004, 23-32.

31 Scharre, *Army of None*, 287-288.

32 It is worth noting here that the Army Ethic acknowledges the importance of respect and includes as a principle, "In war and peace, we recognize the intrinsic dignity and worth of all people, treating them with respect." See ADRP-1, 2-7. I owe this point to Michael Toler, Center for the Army Profession and Ethic.

33 Robert Sparrow: "Robots and Respect: Assessing the Case Against Autonomous Weapon Systems," *Ethics & International Affairs*, Vol. 30, No. 1, 2016, 106.

34 Eliav Lieblich and Eyal Benvenisti, "The obligation to exercise discretion in warfare: why autonomous weapons systems are unlawful," in Nehal Bhuta, Susanne Beck, Robin Geiβ, Hin-Yan Liu, Claus Kreβ, eds., *Autonomous Weapon Systems: Law, Ethics, Policy,* Cambridge: Cambridge University Press, 2016, 266.

35 Lieblich and Benvenisti, 267.

36 The point here is not that the battlefield is a place to negotiate or that there has to be some kind of interaction independent of a targeting process to justify the decision to use lethal force. Rather, the point is that in any given situation where a human being may be harmed that the decision made to commit that harm is made by another human who can identify and consider the range of factors that would justify that harm. In this way, the person deciding to harm understands it is another person who he or she is harming and considers reasons not to do it. Autonomous systems may eventually be able to discern a number of relevant factors; however, that only entails they are considering reasons to harm, not reasons not to harm.

37 Sparrow, 101.

38 Ibid.,. 106-107.

39 Ibid., 108.

40 Lieblich and Benvenisti, 267.

41 Sparrow, 110-111.

42 Kenneth J. Arrow: "Uncertainty and the Welfare Economics of Medical Care," *The American Economic Review,* Vol. 53, No. 5, December 1963, 941, 961. See also Matthew McCaffrey, "Moral Hazard: Kenneth Arrow vs Frank Knight and the Austrians," MISES WIRE*,* March 14, 2017, available from https://mises.org/blog/moral-hazard-kennetharrow-vs-frank-knight-and-austrians, accessed August 27, 2018.

43 Wallach and Allen, *Moral Machines*, 16.

44 Ibid., 16.

45 Ibid.

46 Marcello Guarini and Paul Bello, "Robotic Warfare: Some Challenges in Moving from Noncivilian to Civilian Theaters," in Patrick Lin, Keith Abney, and George A. Berkey eds., *Robot Ethics,* Cambridge, MA: The MIT Press, 2014, 129-130.

47 Ibid., 131-132.

48 Zed Adams and Jacob Browning, "Introduction," in Zed Adams and Jacob Browning, eds., *Giving a Damn: Essays in Dialogue with John Haugeland,* Cambridge, MA: MIT Press, 2017, 4.

49 Kevin Warwick, "Robots with Biological Brains," in *,* Patrick Lin, Keith Abney, and George A. Bekey, eds, *Robot Ethics,* Cambridge, MA: MIT Press, 2014, 327-328.

50 Adams and Browning, 5.

51 Ibid., 13.

52 Harry Frankfurt: "Freedom of the Will and the Concept of the Person," *The Journal of Philosophy,* Vol. 68, No. 1, January 14, 1971, pp. 6-10. Persons, in Frankfurt's scheme, can have an infinite number of higher order desires; however, the actual number will be practically limited by "common sense" and "fatigue." His point here is that as long as there is a second order desire, the potential for free will exists. See Frankfurt, 16.

53 Ibid., 15.

54 Roff, 353.

55 Department of Defense, *Department of Defense Directive: Autonomy in Weapon Systems, Number 3000.09*, Washington, DC: U.S. Government Printing Office, November 21, 2012.

56 David Davenport, "The Two (Computational) Faces of AI," in Vincent C. Muller, ed., *Philosophy and Theory of Artificial Intelligence,* Heidleberg: Springer, 2013, 44-53.

57 Scharre and Horowitz, *Artificial Intelligence: What Every Policy Maker Needs to Know,* 10.

58 Wendell Wallach and Colin Allen, "Framing robot arms control," *Ethics of Information Technology,* Vol. 15, 2013, 126.

59 Christopher Heyns, "Autonomous Weapon Systems: living a dignified life and dying a dignified death," in Nehal Bhuta, Susanne Beck, Robin Geiβ, Hin-Yan Liu, Claus Kreβ, eds., *Autonomous Weapon Systems: Law, Ethics, Policy,* (Cambridge: Cambridge University Press, 2016), 10.

60 Wallach and Allen, *Moral Machines,* 25-26.

61 Ibid*.,* 25.

62 Yoram Dinstein, *War, Aggression, and Self-Defence,* 4th Ed, Cambridge, Cambridge University Press, 2005, 136.

63 Sanford Levinson, "Responsibility for War Crimes," in Marshall Cohen, Thomas Nagel, and Thomas Scanlon, eds., *War and Moral Responsibility,* Princeton, Princeton University Press, 1974, 117.

64 Liu, 329.

65 Ibid., 330-331.

66 Ibid., 340.

67 Ibid., 336.

68 Levinson, 118.

69 Ibid., 119.

70 Ibid., 119.

71 Walzer, *Just and Unjust Wars,* 316-320.

72 Jeffrey L. Caton, *Autonomous Weapon Systems: A Brief Survey of Developmental, Operational, Legal, and Ethical Issues*, Carlisle Barracks, PA: Strategic Studies Institute, December 2015, 2-3.

73 Scharre, *Army of None*, 28-29.

74 Heynes, 14.

75 Cheryl Pellerin, "Project Maven to Deploy Computer Algorithms to War Zones by Years End," *Defense*

One, July 21, 2017, available from https://www.defense.gov/News/Article/Article/1254719/project-maven-to-deploy-computer-algorithms-to-war-zone-by-years-end/, accessed August 13, 2018. Regarding AI-driven assessments: humans are better considered "on-the-loop" regarding the assessment itself, since they can monitor and intervene how the machine considers relevant input; however, to the extent they decide how to act on that assessment, they would be "in-the-loop." I owe this point to Colonel David Barnes, Department of English and Philosophy, United States Military Academy at West Point.

76 William T. Sherman, available from https://www.goodreads.com/author/quotes/177344.William_T_Sherman, accessed August 27, 2018.

77 Noel Sharkey, "Killing Made Easy: From Joy Sticks to Politics," in *Robot Ethics,* 111-112.

78 P.W. Singer, *Wired for War: The Robotics Revolution and Conflict in the 21st Century,* New York: Penguin Books, 2009, 332.

79 Ibid., 332.

80 Ibid., 395.

81 Ibid., 326-327.

82 Efram Karsh and Inari Rautsi, *Saddam Hussein,* New York: The Free Press, 1991. This book documents the human rights abuses of Saddam Hussein and family members, including Uday and Qusay.

83 Williamson Murray and Major General Robert H. Scales Jr., *The Iraq War: A Military History,* Cambridge, MA: The Belknap Press, 2003, 234-235.

84 Pratap Chatterjee, American Drone Operators are Quitting in Record Numbers, The Nation, March 5 2015, available from https://www.thenation.com/article/american-drone-operators-are-quitting-record-numbers, accessed August 2, 2018.

85 Samuel Issacharoff and Richard Pildes, "Drones and the Dilemmas of Modern Warfare," in Peter Bergen and Daniel Rothenberg, eds., *Drone Wars: Transforming Conflict, Law, and Policy,* Cambridge: Cambridge University Press, 2015, 399.

86 Phillip Sherwell, "U.S. drone pilot haunted by horrors of remote killings; Operator sickened by death count" London Daily Telegraph, October 25, 2013, available from https://www.telegraph.co.uk/news/worldnews/northamerica/usa/10403313/Confessions-of-a-US-drone-operator-I-watched-him-die.-It-took-a-long-time.html, accessed August 10, 2018.

87 Matt J. Martin and Charles Sasser, *Predator: The remote-control air war over Iraq and Afghanistan: A pilot's story,* Minneapolis, MN: Zenith Press, 2010, quoted in Isaccharoff and Pildes, 416.

88 Christian Enemark, *Armed Drones and the Ethics of War,* London: Routledge, 2014, 22.

89 Ibid., 23.

90 Sharkey, 115. As Sharkey states, "it is now unclear what type and level of evidence is being used to sentence nonstate actors to death by Hellfire attack without right to appeal or right to surrender." The point here is not that U.S. targeting procedures are unjust, but to the extent they are not transparent—at least to the extent possible without compromising the process—the perception of injustice will persist, generating the kind of response evidence by the Google employees.

91 Enemark, 25.

92 I owe this point to Colonel David Barnes, Department of English and Philosophy, U.S. Military Academy at West Point.

93 Singer, 125.

94 Ibid., 125.

95 Hyun-Ku Lee and Hangjung Zo, "Assimilation of Military Group Decision Support Systems in Korea: The Mediating Role of Structural Appropriation" *Information Development* 33, Vol 21, No. 1, 2017, quoted in James Boggess, "More than a Game: Third Offset and the Implications for Moral Injury," in *Closer than You Think: The Implications of the Third Offset Strategy,* Carlisle, PA: Strategic Studies Institute, 2017, 133.

96 David Evans, "Vincennes: A Case Study," *Proceedings,* Vol 119/8/1086, August, 1993.

97 Scharre and Horowitz, *What Policy Makers Need to Know,* 11.

98 Scharre, *Army of None,* 321.

99 Ibid., *Army of None,* 323-324.

100 Ami Rojkes Dombe, "China Unveils a Harpy-type Loitering Munition," *IsraelDefense,* March 1, 2017, http://www.israeldefense.co.il/en/node/28716, accessed July 16, 2018.

101 Noah Sharkey, "Guidelines for the human control of Weapon Systems," *International Committee for Robot Arms Control*, April 2018, https://www.icrac.net/wp-content/uploads/2018/04/Sharkey_Guideline-for-the-human-control-of-weapons-systems_ICRAC-WP3_GGE-April-2018.pdf, accessed August 10, 2018.

102 Scharre and Horowitz, *Autonomy in Weapons Systems,* Washington, D.C.: Center for a New American Security, 2015, 16.

103 Heyns, 7.

104 Arkin, 30.

105 James Rachels, *The Elements of Moral Philosophy,* 3rd Ed., Boston: McGraw Hill, 1999, 107-121.

106 Arthur Isaak Applbaum, *Ethics for Adversaries: The Morality of Roles in Public and Professional Life,* Princeton, NJ: Princeton University Press, 1999, pp. 162-166. Applbaum refers to situations where someone is better off and no one is worse off as "avoiding Pareto-inferior outcomes." Avoiding such outcomes can count as "fair" and warrant overriding consent.

107 Guarini and Bello, 132.

108 Daniel Brunstetter and Megan Braun, "From Jus ad Bellum to Jus ad Vim: Recalibrating our Understanding of the Moral Use of Force," *Ethics & International Affairs*, 27, no. 1 (2013).

109 Fact Sheet, *U.S. Policy on the Export of Unmanned Aerial Systems,* U.S. State Department, April 19, 2018. https://www.state.gov/r/pa/prs/ps/2018/04/280619.htm. This policy is a little more permissive than the one implemented by the last administration. The new policy permits the sale of armed unmanned systems through Direct Commercial Sales and removes language regarding a "strong presumption of denial" for systems that cross the Missile Technology Control Regime thresholds, which in this case are systems with a range greater than 300 kilometers and are capable of carrying 500 kilogram payloads or more. Both sets of standards require upholding international law. See Fact Sheet, *U.S. Policy on the Export of Unmanned Aerial Systems,* U.S. State Department, February 17, 2015. https://2009-2017.state.gov/r/pa/prs/ps/2015/02/237541.htm. Accessed August 10, 2018.

110 I owe this point to Dr. Steve Metz, Strategic Studies Institute, U.S. Army War College.

111 Sharkey, 115.